frontiers | Research Topics

Reviews in astronomy and space sciences

Edited by
Christopher H. K. Chen, Alberto Fairén, Jonathan Eastwood, Olga V. Khabarova, Georgios Balasis, Rudolf A. Treumann, Philip J. Erickson, Nuno C. Santos, Steve B. Howell, Matthew Parry, Daniel Hestroffer, Yanxia Zhang, Majdi Hochlaf, Ankan Das, Ryan C. Fortenberry, David Leisawitz, Lucia Kleint, Mauro D'Onofrio, Milan S. Dimitrijevic, Joseph M. Trigo-Rodiguez, Carrie Anderson, Zhonghua Yao, Chunshan Lin, Nadja S. Magalhaes and Guillem Domenech

Published in
Frontiers in Astronomy and Space Sciences

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-4936-0
DOI 10.3389/978-2-8325-4936-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Reviews in astronomy and space sciences

Topic editors

Christopher H. K. Chen — Queen Mary University of London, United Kingdom
Alberto Fairén — Center for Astrobiology (CAB), CSIC-INTA, Spain
Jonathan Eastwood — Imperial College London, United Kingdom
Olga V. Khabarova — Tel Aviv University, Israel
Georgios Balasis — National Observatory of Athens, Greece
Rudolf A. Treumann — Ludwig Maximilian University of Munich, Germany
Philip J. Erickson — Massachusetts Institute of Technology, United States
Nuno C. Santos — Instituto de Astrofisica e Ciências do Espaço (IA), Portugal
Steve B. Howell — National Aeronautics and Space Administration, United States
Matthew Parry — University of Otago, New Zealand
Daniel Hestroffer — Université Paris Sciences & Lettres PSL, France
Yanxia Zhang — Chinese Academy of Sciences (CAS), China
Majdi Hochlaf — Université Paris Est Marne la Vallée, France
Ankan Das — Institute of Astronomy Space and Earth Science, India
Ryan C. Fortenberry — University of Mississippi, United States
David Leisawitz — National Aeronautics and Space Administration, United States
Lucia Kleint — University of Geneva, Switzerland
Mauro D'Onofrio — University of Padua, Italy
Milan S. Dimitrijevic — Astronomical Observatory, Serbia
Joseph M. Trigo-Rodiguez — Spanish National Research Council (CSIC), Spain
Carrie Anderson — National Aeronautics and Space Administration, United States
Zhonghua Yao — Chinese Academy of Sciences (CAS), China
Chunshan Lin — Jagiellonian University, Poland
Nadja S. Magalhaes — Federal University of São Paulo, Brazil
Guillem Domenech — National Institute of Nuclear Physics of Padova, Italy

Citation

Chen, C. H. K., Fairén, A., Eastwood, J., Khabarova, O. V., Balasis, G., Treumann, R. A., Erickson, P. J., Santos, N. C., Howell, S. B., Parry, M., Hestroffer, D., Zhang, Y., Hochlaf, M., Das, A., Fortenberry, R. C., Leisawitz, D., Kleint, L., D'Onofrio, M., Dimitrijevic, M. S., Trigo-Rodiguez, J. M., Anderson, C., Yao, Z., Lin, C., Magalhaes, N. S., Domenech, G., eds. (2024). *Reviews in astronomy and space sciences*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4936-0

Table of contents

06 Editorial: Reviews in space physics
Olga Khabarova, Georgios Balasis, Radoslav Bučík, Jonathan P. Eastwood, Philip J. Erickson and Rudolf A. Treumann

09 Editorial: Reviews in astrobiology
Alberto G. Fairén

11 Fifty Years of ^{3}He-Rich Events
Donald V. Reames

20 The grit crust: A poly-extremotolerant microbial community from the Atacama Desert as a model for astrobiology
Patrick Jung, Lukas W. Lehnert, Jörg Bendix, Guillaume Lentendu, Martin Grube, Fernando D. Alfaro, Camilo del Rio, José Luis Gutiérrez Alvarado, Liesbeth van den Brink and Michael Lakatos

29 Kinematics of the diffuse intragroup and intracluster light in groups and clusters of galaxies in the local universe within 100 Mpc distance
Magda Arnaboldi and Ortwin Gerhard

53 Space Weather: From solar origins to risks and hazards evolving in time
Natalia Buzulukova and Bruce Tsurutani

75 Setting the geological scene for the origin of life and continuing open questions about its emergence
Frances Westall, André Brack, Alberto G. Fairén and Mitchell D. Schulte

107 Detection of organic matter on Mars, results from various Mars missions, challenges, and future strategy: A review
Arif H. Ansari

121 Dark blue-green: Cave-inhabiting cyanobacteria as a model for astrobiology
Patrick Jung, Felix Harion, Shujie Wu, Dennis J. Nürnberg, Francesco Bellamoli, Antonio Guillen, Manuel Leira and Michael Lakatos

130 On the phenomenology of magnetosheath jets with insight from theory, modelling, numerical simulations and observations by Cluster spacecraft
Marius Echim, Mirela Voiculescu, Costel Munteanu, Eliza Teodorescu, Gabriel Voitcu, Cătălin Negrea, Simona Condurache-Bota and Emilian Bujor Dănilă

149 Role of the variable solar wind in the dynamics of small-scale magnetosheath structures
Liudmila Rakhmanova, Maria Riazantseva, Georgy Zastenker and Yuri Yermolaev

157 Numerical modeling of solar wind and coronal mass ejection in the inner heliosphere: A review
Man Zhang, Xueshang Feng, Huichao Li, Ming Xiong, Fang Shen, Liping Yang, Xinhua Zhao, Yufen Zhou and Xiaojing Liu

177 Cosmic ray propagation in turbulent magnetic fields
Alex Lazarian, Siyao Xu and Yue Hu

189 SPGCam: A specifically tailored camera for solar observations
David Orozco Suárez, Daniel Álvarez García, Antonio C. López Jiménez, María Balaguer Jiménez, David Hernández Expósito, Pierre Labrousse, Francisco J. Bailén, Isabel Bustamante Díaz, Eduardo Bailón Martínez, Beatriz Aparicio del Moral, José M. Morales Fernández, Antonio Sánchez Gómez, Ángel Tobaruela Abarca, Antonio J. Moreno Mantas, José L. Ramos Más, Isabel Pérez Grande, Javier Piqueras Carreño, Yukio Katsukawa, Masahito Kubo, Yusuke Kawabata, Takayoshi Oba, Manuel Rodríguez Valido, Eduardo Magdaleno Castelló and Jose Carlos Del Toro Iniesta

208 Warm giant exoplanet characterisation: current state, challenges and outlook
Simon Müller and Ravit Helled

217 Solar wind with Hydrogen Ion charge Exchange and Large-Scale Dynamics (SHIELD) DRIVE Science Center
Merav Opher, John Richardson, Gary Zank, Vladimir Florinski, Joe Giacalone, Justyna M. Sokół, Gabor Toth, Sanlyn Buxner, Marc Kornbleuth, Matina Gkioulidou, Romina Nikoukar, Bart Van der Holst, Drew Turner, Nicholas Gross, James Drake, Marc Swisdak, Kostas Dialynas, Maher Dayeh, Yuxi Chen, Bertalan Zieger, Erick Powell, Chika Onubogu, Xiaohan Ma, Ethan Bair, Heather Elliott, Andre Galli, Lingling Zhao, Laxman Adhikari, Masaru Nakanotani, Matthew E. Hill, Parisa Mostafavi, Senbei Du, Fan Guo, Daniel Reisenfeld, Stephen Fuselier, Vladislav Izmodenov, Igor Baliukin, Alan Cummings, Jesse Miller, Bingbing Wang, Keyvan Ghanbari, Jozsef Kota, Abraham Loeb, Juditra Burgess, Sarah Chobot Hokanson, Cherilyn Morrow, Adam Hong and Andrea Boldon

243 The importance of Phobos simulants: a review on our current knowledge
Catarina Miranda, Manish Patel, Mário Nuno Berberan-Santos, Tiago Hormigo, Alexandre C. M. Correia, Bruno Pedras and Zita Martins

254 Constraining cosmological parameters from N-body simulations with variational Bayesian neural networks
Héctor J. Hortúa, Luz Ángela García and Leonardo Castañeda C.

264 Measurement methods for gamma-ray bursts redshifts
Mengci Li, Zhe Kang, Chao Wu, Chengzhi Liu, Jirong Mao, Zhenwei Li, Shiyu Deng, Bingli Niu and Ping Jiang

280 Interplanetary scintillation observation and space weather modelling
Ming Xiong, Xueshang Feng, Bo Li, Jiansen He, Wei Wang, Yanchen Gao, Man Zhang, Liping Yang, Zhenghua Huang, Jun Cheng, Cang Su, Yihua Yan and Kairan Ying

299 A review of studies on mass-movements on the Moon
Kaixuan Li, Jiao Wang and Yijian Shi

317 Hard rock dark biosphere and habitability
Cristina Escudero and Ricardo Amils

328 From the laboratory to the interstellar medium: a strategy to search for exotic molecules in space
Cristina Puzzarini, Silvia Alessandrini, Luca Bizzocchi, Mattia Melosso and Víctor M. Rivilla

347 Simulating asteroid impacts and meteor events by high-power lasers: from the laboratory to spaceborne missions
Martin Ferus, Antonín Knížek, Giuseppe Cassone, Paul B. Rimmer, Hitesh Changela, Elias Chatzitheodoridis, Inna Uwarova, Ján Žabka, Petr Kabáth, Franz Saija, Homa Saeidfirozeh, Libor Lenža, Miroslav Krůs, Lukáš Petera, Lukáš Nejdl, Petr Kubelík, Anna Křivková, David Černý, Martin Divoký, Michael Pisařík, Tomáš Kohout, Lakshika Palamakumbure, Barbora Drtinová, Klára Hlouchová, Nikola Schmidt, Zita Martins, Jorge Yáñez, Svatopoluk Civiš, Pavel Pořízka, Tomáš Mocek, Jona Petri and Sabine Klinkner

359 Point spread function modelling for astronomical telescopes: a review focused on weak gravitational lensing studies
Tobías I. Liaudat, Jean-Luc Starck and Martin Kilbinger

394 Exploring the high abundance discrepancy in the planetary nebula IC 4663
M. Mohery, A. Ali, Aya A. Khames, S. Snaid and A. Mindil

406 Observations key to understanding solar cycles: a review
Sara F. Martin

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Editorial
PUBLISHED 19 July 2023
DOI 10.3389/fspas.2023.1254235

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY:
Luca Sorriso-Valvo,
National Research Council (CNR), Italy

*CORRESPONDENCE
Olga Khabarova,
olgakhabar@tauex.tau.ac.il

RECEIVED 06 July 2023
ACCEPTED 07 July 2023
PUBLISHED 19 July 2023

CITATION
Khabarova O, Balasis G, Bučík R, Eastwood JP, Erickson PJ and Treumann RA (2023), Editorial: Reviews in space physics.
Front. Astron. Space Sci. 10:1254235.
doi: 10.3389/fspas.2023.1254235

COPYRIGHT
© 2023 Khabarova, Balasis, Bučík, Eastwood, Erickson and Treumann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Reviews in space physics

Olga Khabarova[1]*, Georgios Balasis[2], Radoslav Bučík[3], Jonathan P. Eastwood[4], Philip J. Erickson[5] and Rudolf A. Treumann[6]

[1]Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel, [2]Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Athens, Greece, [3]Southwest Research Institute, San Antonio, TX, United States, [4]Department of Physics, Imperial College, London, United Kingdom, [5]Haystack Observatory, Massachusetts Institute of Technology, Westford, MA, United States, [6]International Space Science Institute, Bern, Switzerland

KEYWORDS

space physics, heliosphere, space weather, magnetosphere, cosmic rays, energetic particles, solar wind, turbulence

Editorial on the Research Topic
Reviews in space physics

The present Research Topic opens the series of "*Reviews in space physics*" with a Research Topic comprising a limited number of submissions discussing the current state of research in the wide domain of Space Physics. Within the framework of this Research Topic, Frontiers will publish reviews, covering diverse aspects of heliospheric, solar, magnetospheric, and ionospheric physics. Reviews included in this Research Topic are aimed at covering processes from the upper terrestrial atmosphere to distant astronomical space, critically discussing past and future developments in theory, observation, analysis, and instrumentation.

This first 2023 Research Topic of Reviews in Space Physics contains six articles written by 60 authors. Its Research Topic focus on 1) space weather, its geo-effective mechanisms and possible threats to civilization, including radiation hazards; 2) studies of dynamical structures in the small- and large-scale solar wind and the consequences of their interaction with the terrestrial magnetosphere, with special focus on processes in Earth's magnetosheath (MSH); 3) observations and modelling of MSH plasma jets, and 4) acceleration and propagation of solar energetic particles (SEPs) and cosmic rays (CRs) in the heliosphere and beyond.

Space weather and its effects on human activities have become an important Research Topic, especially as recent global efforts inexorably embed space technology and services in everyday life. Buzulukova and Tsurutani cover the basic features of space weather and its technological effects overviewing the chain of physical processes responsible for space weather hazards, tracing from its solar origins to effects in interplanetary space and impacts at Earth. All kinds of solar phenomena causing space weather variability, from coronal mass ejections (CMEs) to coronal holes, are considered. The authors show that extreme space weather poses a risk to the technosphere and may be associated with spacecraft/satellite failures, communication/navigation issues, power blackouts and human health negative changes caused by geomagnetically-induced currents and complex ionospheric-atmospheric-thermospheric effects. A detailed overview of geomagnetic storms, substorms,

and Auroral Electrojet activity intervals provides information on Sun-Earth-system-coupled responses. Future concerted community efforts required to fill the knowledge gap are discussed.

Rakhmanova et al. assign a significant geoeffective role to the coupling between MSH structures and the variable solar wind at kinetic scales. The turbulent MSH is sensitive to ~10^{-3}–10^{-2} Hz solar wind and foreshock variations of amplitudes 5%–10% above background representing so-called jets, rather than waves. Transient dynamic pressure enhancements forming jets in 3D are physically plasmoids or flux ropes or, in 2D, magnetic islands, holes, and Alfvén vortices contributing to MSH turbulence and intermittency. Aligned with the Sun–Earth line, they become geoeffective. The article advocates future diversified missions active over the entire solar cycle to investigate their role in overall Sun-Earth physics.

Echim et al. focus on the same MSH structures as Rakhmanova et al. calling them "multiscale dynamical irregularities of the background MSH plasma state". The authors review their properties in a multi-pronged effort referring to theory, numerical simulations and observations. Using ~1000 jet observations from Cluster 3, in combination with theory and simulations they observe a dawn-dusk asymmetry with stronger density perturbations along the dusk flank. At dawn, denser and colder jets dominate. Approaching the Earth, jet speeds decrease while perpendicular jet ion temperatures increase, which indicates adiabatic breaking potentially impacting ionospheric dynamics.

The presence of dangerous SEPs and CRs in the inner heliosphere is a permanent threat for human space exploration. Reames provides an in-depth review of impulsive (or ^{3}He-rich) SEP events, describing the evolution of views on the nature of SEPs with an unusual elemental composition discovered in the early 1970s and showing ratios ^{3}He/^{4}He, Fe/O enhanced above the coronal abundances later detected in high-Z elements. Impulsive SEP events with large proton excess (see Figure 1) and rare ^{4}He-pure events are discussed, and various acceleration theories are presented, from resonant plasma wave acceleration to acceleration on magnetic islands collapsing in reconnection. The significance of radio and imaging observations in discerning acceleration processes in SEPs is addressed.

Magnetic fields play a significant role in space physics, particularly in the propagation of CRs in heliospheric plasmas, preventing their escape and facilitating diffusion. Spatial properties and temporal variations of the magnetic fields, modulated by turbulence and intermittency, determine the path and the intensity of CRs in the heliosphere, the interstellar and intracluster media. The heliosphere shields ~70% of galactic CRs, and knowing the way they propagate to the Earth is very important for radiation monitoring and protection. The current issue contains two articles concerning this Research Topic.

Lazarian et al. provide an overview of recent theoretical findings on CR propagation in turbulent magnetic fields in astrophysical plasmas and summarize the key implications and mechanisms involved. They discuss details of how pre-existing magnetohydrodynamic turbulence impacts parallel and perpendicular CR diffusion, as well as perpendicular super diffusion of CRs. Parallel diffusion prevails only in ideal plasma. In realistic astrophysical plasmas perpendicular diffusion cannot be ignored

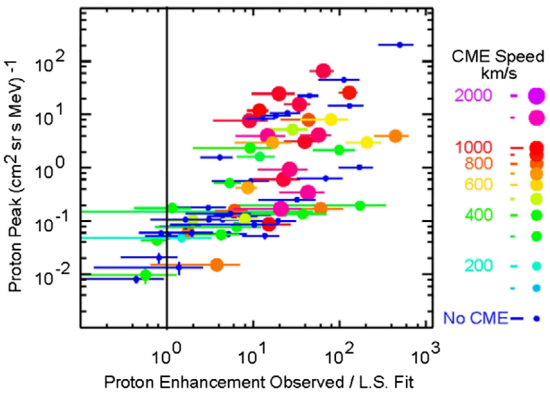


FIGURE 1
Scatter plot of proton peak intensity at 2–2.4 MeV vs. relative proton excess for impulsive SEPs associated and not associated with CMEs (from Reames, distributed under the terms of the Creative Commons Attribution License (CC BY)). CME speed is given in colored dots, and non-CME events - in blue. Reames distinguishes 4 SEP-types: Small impulsive SEP1 (protons fit power law of heavier ions), SEP2 (impulsive with enhanced proton excess; accelerated by CME-driven shock), SEP3 (gradual; re-accelerated SEP1, ion dominated), and SEP4 (gradual; shock accelerates primarily ambient population). Most CMEs with the speed >500 km/s accelerate SEPs, with non-CME SEP sources contributing to the pool of events with high excess as well. Non-CME SEPs are probably shock-enhanced SEP2s seeded by SEP1s and ambient coronal seed particles. The proton excess plot allows distinguishing pure SEP events from those in which SEPs are re-accelerated.

because magnetic fluctuations in trans-Alfvénic turbulence are comparable to the mean magnetic field, being even larger in super-Alfvénic turbulence. In strong turbulence, the magnetic field becomes not frozen into the plasma, which leads to super diffusion at small scales. Perpendicular diffusion of CRs is impacted by turbulent magnetic fields regulated by Alfvén modes, while parallel diffusion depends on resonance broadening effects and transit-time damping interactions.

Opher et al. focus on global processes in the heliosphere related to CR transport. Rather than a more traditional review, the article presents key questions and problems through a comprehensive description of the proposal for the now NASA-funded DRIVE science center SHIELD, studying the structure and dynamics of the heliosphere through integrated research. The authors indicate knowledge gaps that can be filled with building self-consistent global models of the heliosphere and its interaction with the interstellar medium as well as creating CR transport models. The article outlines the SHIELD research plan and discusses the ways the center will contribute in predicting variations of the radiation environment of the Earth and studying the evolution of pickup ions and their impact on heliospheric processes.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

OK's research was carried out with the assistance of the Center for Absorption in Science, Ministry of Immigration and Absorption, State of Israel. RB was supported by NASA grants 80NSSC22K0757, 80NSSC21K1316 and NASA contract NNN06AA01C. JE was supported by UKRI/STFC grant ST/W001071/1.

Acknowledgments

The editors are grateful to the authors of the reviews for their work on collecting the material and its analysis. They thank Joseph E. Borovsky and Christopher H. K. Chen for their help with editorial duties. The authors thank the reviewers of their articles, namely, Laxman Adhikari, Joseph Borovsky, Jungyeon Cho, Jingnan Guo, Olga Gutynska, George Ho, Jeffrey Linsky, Siming Liu, Terry Zixu Liu, Scott William McIntosh, Luis Preisser, and Ilan Roth.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Editorial
PUBLISHED 21 September 2023
DOI 10.3389/fspas.2023.1292594

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Julio Navarro,
University of Victoria, Canada

*CORRESPONDENCE
Alberto G. Fairén,
agfairen@cab.inta-csic.es

RECEIVED 11 September 2023
ACCEPTED 14 September 2023
PUBLISHED 21 September 2023

CITATION
Fairén AG (2023), Editorial: Reviews in astrobiology.
Front. Astron. Space Sci. 10:1292594.
doi: 10.3389/fspas.2023.1292594

COPYRIGHT
© 2023 Fairén. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Reviews in astrobiology

Alberto G. Fairén[1,2]*

[1]Centro de Astrobiología (CAB), CSIC-INTA, Madrid, Spain, [2]Department of Astronomy, Cornell University, Ithaca, NY, United States

KEYWORDS

astrobiology, origins of life, deep biosphere, Mars, organics, extremophiles

Editorial on the Research Topic
Reviews in astrobiology

Astrobiology was defined in 1995 by the now defunct NASA Astrobiology Institute (NAI) as "the study of the living Universe", therefore explicitly including investigations on the origin and evolution of life on Earth. It is true that Earth's biology constitutes the only known subject of study for Astrobiology so far, and that's the reason why Astrobiology is vigorously working to increase the sample size searching for additional occurrences of living entities outside our planet. This Research Topic "*Reviews in Astrobiology*" is focused on revisiting the most recent advances in this search, as progress in Astrobiology over the past 2 decades has been very rapid.

This Research Topic of 5 review articles touches on several of the key challenges for future lines of research in Astrobiology, combining the diverse perspectives gained by a breadth of experts. The Research Topic addresses a variety of advances related to the origin of life on Earth, the habitability of the deep subsurface on Earth and exoplanets, the search for organics on Mars, and two comprehensive overviews on two different models of microbial communities as models for astrobiological exploration.

In their review article "*Setting the geological scene for the origin of life and continuing open questions about its emergence*," Westall et al. discuss much of what we know about the dynamic and overlapping geological environments on the early Earth, and how certain molecular building blocks could have been provided both on and to the early Earth, exploring endogenous and exogenous production. The latter section of the paper revisits open questions in origins research.

Addressing the prokaryotic diversity in the Earth's continental hard rock deep subsurface, Escudero and Amils provide a comprehensive overview of planetary habitability in their review article "*Hard rock dark biosphere and habitability*". Understanding the importance of the dark biosphere on continental systems on Earth, how we have acquired information about it, and its relevance to assess the habitability of exoplanets, is of paramount relevance for the future of Astrobiology.

The hurdles of finding organic molecules on Mars are discussed in "*Detection of organic matter on Mars, results from various Mars missions, challenges, and future strategy: A review*" by Ansari. This article reviews the historical detections of organics made from the Vikings to the Perseverance rover missions, with the aim of understanding the elusive carbon cycling of Mars, identifying the sites with better preservation potential for hydrocarbons, and describing how developing methodologies are being refined for improved extraction of indigenous organic molecules with minimum contamination.

Two studies in this Research Topic provided elegant analog microbial models for future astrobiological exploration. In "*Dark blue-green: Cave-inhabiting cyanobacteria as a model for astrobiology*", Jung et al. introduce cave-inhabiting cyanobacteria as a model system for astrobiology, presenting an overview of literature that describe these fascinating phototrophs in the context of their ecology, pigment composition, ability to produce bioplastic and to create living biomaterials. This mini-review summarizes known information about cyanobacteria and places it in the framework of astrobiology.

And in "*The grit crust: A poly-extremotolerant microbial community from the Atacama Desert as a model for astrobiology*", Jung et al. provide a short review in which the research on a recently described biocrust in the Coastal Range of the Atacama region of Chile is summarized. The grit crust's extremophiles are described as suitable for mass cultivation in photobioreactors to provide food and oxygen for crewed missions, as screenings for Chlorophyll f therefore allowing photosynthesis of cyanobacteria in low light environments such as on icy moons or exoplanets, and as a model for symbiotic interactions useful to understand ancient life forms.

In summary, the articles in this Research Topic provide an overview of issues related to the recent advances of Astrobiology on the origin and distribution of life on Earth and on developing strategies in the search for life beyond our planet. I hope that this Research Topic inspires new ideas and helps to foster deeper collaboration between biologists, geologists, physicists and chemists to help solving the numerous and significant challenges of Astrobiology, a true 21st-century science.

As a concluding remark, I would like to take this opportunity to thank all contributing authors, reviewers, editors, and research funding agencies for their help in completing this Research Topic of reviews on the latest consensus in Astrobiology.

Author contributions

AF: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work is a contribution from the Project "MarsFirstWater", funded by the European Research Council, Consolidator Grant no 818602.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Astronomy and Space Sciences

REVIEW
published: 27 September 2021
doi: 10.3389/fspas.2021.760261

Fifty Years of ^{3}He-Rich Events

*Donald V. Reames**

Institute for Physical Science and Technology, University of Maryland, College Park, MD, United States

The early 1970s saw a new and surprising feature in the composition of solar energetic particles (SEPs), resonant enhancements up to 10,000-fold in the ratio ^{3}He/^{4}He that could even make ^{3}He dominant over H in rare events. It was soon learned that these events also had enhancements in the abundances of heavier elements, such as a factor of ~10 enhancements in Fe/O, which was later seen to be part of a smooth increase in enhancements vs. mass-to-charge ratio *A/Q* from H to Pb, rising by a factor of ~1000. These events were also associated with streaming 10–100 keV electrons that produce type III radio bursts. In recent years we have found these "impulsive" SEP events to be accelerated in islands of magnetic reconnection from plasma temperatures of 2–3 MK on open field lines in solar jets. Similar reconnection on closed loops traps the energy of the particles to produce hot (>10 MK), bright flares. Sometimes impulsive SEP intensities are boosted by shock waves when the jets launch fast coronal mass ejections. No single theory yet explains both the sharp resonance in ^{3}He and the smooth increase up to heavier elements; two processes seem to occur. Sometimes the efficient acceleration even exhausts the rare ^{3}He in the source region, limiting its fluence.

Keywords: solar energetic particles, solar jets, shock waves, solar system abundances, magnetic reconnection

OPEN ACCESS

Edited by:
Radoslav Bucik,
Southwest Research Institute (SwRI),
United States

Reviewed by:
George Ho,
Johns Hopkins University,
United States
Siming Liu,
Purple Mountain Observatory (CAS),
China
Ilan Roth,
University of California, Berkeley,
United States

***Correspondence:**
Donald V. Reames
dvreames@gmail.com

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *17 August 2021*
Accepted: *13 September 2021*
Published: *27 September 2021*

Citation:
Reames DV (2021) Fifty Years of ^{3}He-Rich Events. Front. Astron. Space Sci. 8:760261. doi: 10.3389/fspas.2021.760261

INTRODUCTION

The solar energetic particle (SEP) events observed first (Forbush 1946) were the largest and most energetic examples we know, where GeV protons produce a shower of particles cascading through the atmosphere to ground level in excess of that from the galactic cosmic rays (GCRs). However, these ground-level events (GLEs) provided little information on the composition of the incoming beam.

Measurement of multiple element abundances in SEPs began when Fichtel and Guss (1961) used nuclear emulsion detectors on a 4-min sounding-rocket flight from Ft. Churchill, Manitoba to observe a sampling of elements with $6 \le Z \le 16$. The principle elements up to the Fe abundance peak were observed in the next solar cycle by Bertsch et al. (1969) using the same technique. Resolution of He isotopes and the continuous time coverage needed to observe smaller events would only begin when detector telescopes were flown on satellites.

^{3}HE

The first clearly-enhanced abundance ratio of ^{3}He/^{4}He = $(2.1 \pm 0.4) \times 10^{-2}$ was reported by Hsieh and Simpson (1970), perhaps 50 times the value seen in the corona or solar wind, although differences in the spectra of the He isotopes were noted here.

Nearly every scientist involved in the early study of SEPs had previous experience with GCRs. After acceleration by shock waves at supernovae, GCRs spend ~10^7 years colliding with interstellar H to produce secondary ^{2}H, ^{3}He, and isotopes of Li, Be, and B. Thus the observation of ^{3}He/^{4}He of 2%

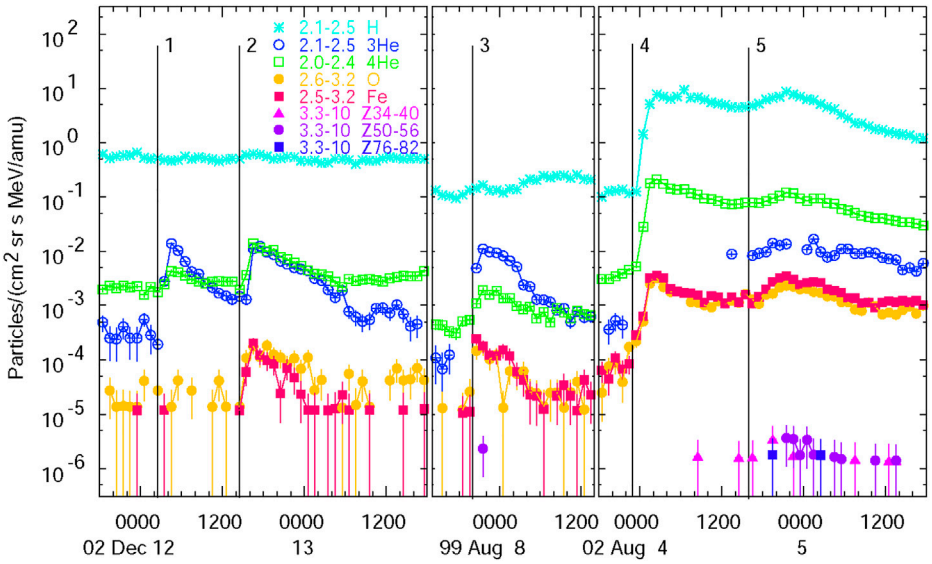


FIGURE 1 | Ion intensities at the indicated MeV amu^{-1} are shown in a sample of ^{3}He-rich or "impulsive" SEP events of increasing intensity. ^{3}He exceeds ^{4}He in Events 1 and 3. Fe and O are below background in Event 1, but Fe/O ≈ 1 in the remaining events. H is below background in Events 1–3. ^{3}He/^{4}He < 0.1 in Event 4 and is omitted because of possible resolution errors; it is ≈0.1 in Event 5 and high-*Z* heavy elements are beginning to appear (Reames and Ng, 2004). Other properties of the last four of these events are listed in Reames et al. (2014a).

by Hsieh and Simpson (1970) was immediately misinterpreted as evidence that the SEPs had traversed enough material to fragment some ^{4}He into ^{3}He, just like the GCRs. However, it was soon found that there were many events like that seen by Serlemitsos and Balasubrahmanyan (1975) with ^{3}He/^{4}He = 1.52 ± 0.10 but ^{3}He/^{2}H > 300. How could there be more ^{3}He than ^{4}He and yet no ^{2}H? Such abundances were definitely not compatible with fragmentation. Subsequently, there were limits established on Be/O and B/O in SEP events that were found to be $<2 \times 10^{-4}$ (e.g., McGuire et al., 1979; Cook et al., 1984). Thus these ^{3}He-rich events were not just an accident of fragmentation; they must involve a completely new resonance phenomenon.

Figure 1 shows a sample of ^{3}He-rich events as we see them above 2 MeV amu^{-1}. As event intensities increase it becomes possible to see rarer ion species. Event 1 has only pre-event background levels of H, O, and Fe and ^{4}He barely appears. ^{3}He/^{4}He varies greatly in events and vs energy, often peaking in the region 1–10 MeV amu^{-1} (Mason, 2007), while Fe/O is much more stable once events are large enough to provide a measurable sample, so Fe/O is often used to define "impulsive" events (e.g., Reames et al., 2014a). As discussed below, very high-*Z* elements begin to appear in the larger impulsive events and ^{3}He fluence may be limited by depletion of ^{3}He ions in the source volume.

Electrons

Some of the earliest information on SEPs in space was provided by the radio emission produced by ~10–100 keV electrons streaming out from the Sun. The radio emission frequency varies as the square root of the local electron density which decreases sharply with distance from the Sun. In an early review article on solar radio bursts, Wild et al. (1963) distinguished type III radio bursts as produced by electrons streaming out rapidly along the magnetic field from a source at the Sun, and type II bursts which moved out at the slower ~1,000 km s^{-1} speed of an interplanetary shock wave. Thus they saw two types of events: "electron events," which produced the type III bursts, and "proton events," accelerated at shock waves where associated electrons produce the type II burst as they are accelerated then overtaken, often at the flanks of the shock.

Early instruments flown on satellites measured 40 keV electrons associated with X-ray bursts at the Sun and type III radio bursts in space (Lin, 1970, 1974). These events were clearly different from the large proton events and some seemed to be "pure" electron events, i.e., lacking measurable ion intensities. The electrons events conformed to the picture presented by Wild et al. (1963).

It was not until 1985 that "pure" electron events turned out to be ^{3}He-rich events (Reames et al., 1985). ^{3}He-rich events were strongly associated with type III bursts, both the metric radio events, near the Sun (Reames et al., 1985), and the kilometric events below 2 MHz produced as the electrons continued out beyond ~6 solar radii (Reames and Stone, 1986).

Abundances of Elements

Measurements of element abundances soon began to show periods when Fe/O ≥ 1 (Mogro-Compero and Simpson, 1972; Gloeckler et al., 1975), an enhancement by a factor of ~10, relative to "coronal" abundances determined by the average of gradual SEP events (e.g., Reames, 1995a, 2014; Reames, 2021a), and such enhancements were shown to correspond with ^{3}He-rich events (e.g., Mason et al., 1986), as seen in **Figure 1**. Reames (1988) looked at daily averages of SEPs to measure the overall distribution of abundances. He found a bimodal pattern with two branches of Fe/O. The branch near Fe/O ≈ 1 was ^{3}He rich,

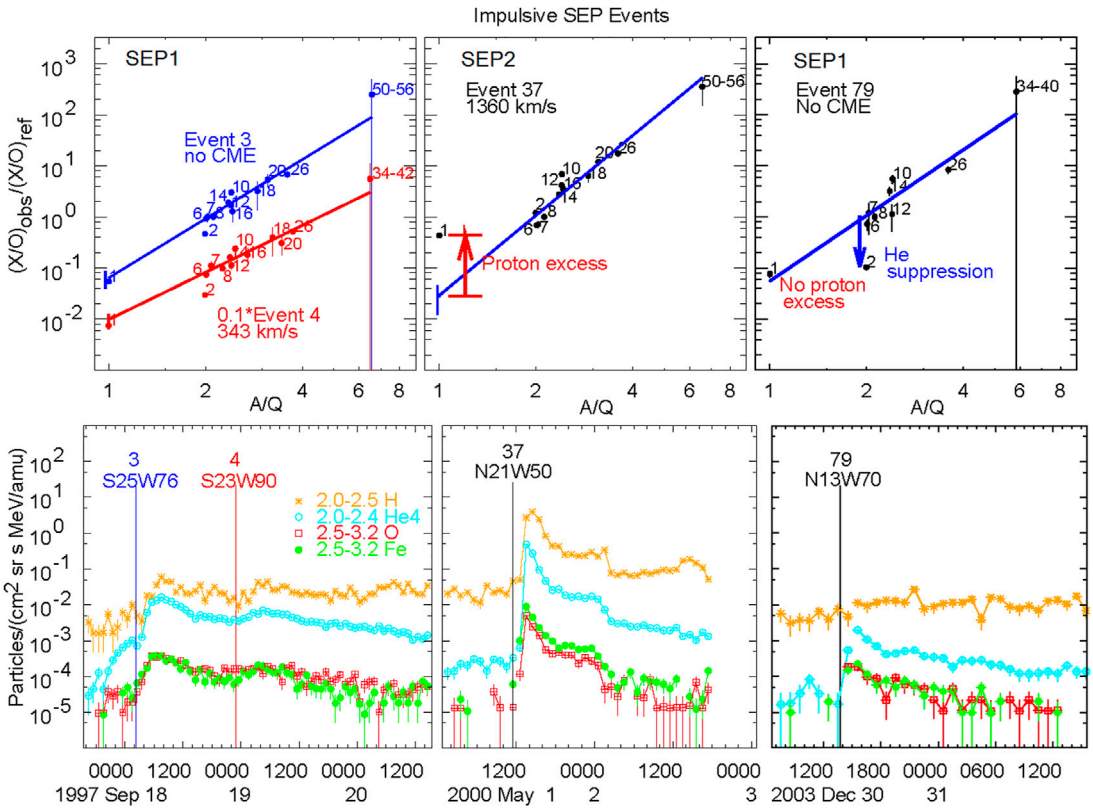


FIGURE 2 | Lower panels show time histories of H, ^{4}He, O and Fe, at the indicated MeV amu^{-1}, for four impulsive SEP events, two small SEP1 events on the **left** and a larger SEP2 event in the **center**, and a ^{4}He-poor SEP1 event on the **right**. Event numbers shown above source coordinates refer to the event list of Reames et al. (2014a). Power-law fits to the abundance enhancements, noted by *Z*, in each event, are shown in the **upper** panels and CME speeds are listed when CMEs are seen. The **center** Event 37 is an SEP2 class event because of its large proton excess noted. For the SEP1 events on the **left** and **right**, the protons abundances lie on the extrapolated power-law fits (Reames, 2019b, 2020).

electron rich, and proton (H/^{4}He) poor relative to the branch near Fe/O ≈ 0.1. These bimodal abundances were helpful in distinguishing the physics of impulsive and gradual SEP events.

Reames et al. (1994) found that on average in ^{3}He-rich events, the elements He, C, N, and O were unenhanced relative to coronal abundances, Ne, Mg, and Si were enhanced a factor of ~2.5 and Fe was enhanced a factor of ~7. This pattern would occur if He–O were fully ionized and Ne, Mg, and Si were in a stable state with two orbital electrons, which occurs in the temperature range 3–5 MK. This suggested that patterns of element abundance enhancements could be used to determine source plasma temperatures, since the pattern of *Q* values and *A*/*Q*, thus element enhancements, was dependent upon temperature (Reames et al., 2014b; Reames, 2018). The temperature variations among impulsive events turned out to be small, so the technique was more useful for gradual SEP events (Reames, 2016) with larger variations in temperature.

The extensive enhancement of very heavy elements was suggested early when Shirk and Price (1974) studied etch pits in a glass window of the *Apollo 16* lunar command module and found $(Z > 44)/\mathrm{Fe} = 120^{+120}_{-60}$ at $0.6 \le E \le 2.0$ MeV amu^{-1} from a small SEP event in April 1972. Routine measurement that resolved elements above Fe with $\delta Z/Z \sim 2\%$ began with the launch of the *Wind* spacecraft in November 1994. Reames (2000) found significant enhancements up to $(70 \le Z \le 82)$, at $3.3 \le E \le 10$ MeV amu^{-1}, but only in impulsive SEP events. These measurements improved statistically with time (e.g., Reames and Ng, 2004) until Reames et al. (2014a) found the dependence of enhancements rising at the 3.64 ± 0.15 power of *A*/*Q* (at ≈3 MK) from He to Pb, with the $(76 \le Z \le 82)$/O interval enhanced by a factor of ≈900. Below 1 MeV amu^{-1}, Mason et al. (2004) found enhancements varying as a power of *A*/*Q* of 3.26 and the interval $180 \le A \le 200$ was enhanced a factor of ≈200. With the exception of ^{3}He, enhancements are not strongly energy dependent.

When the power-law fits to abundance enhancements vs *A*/*Q* for elements $Z \ge 6$ in an event were extrapolated down to H at *A*/*Q* = 1 (Reames, 2019b), there were small impulsive SEP events that seemed to fit the protons extremely well (called SEP1 events; Reames 2020) and some larger events with a large proton excess (called SEP2 events) as shown by the examples in **Figure 2**.

We will see below that the event in the central panels of **Figure 2** was one of the first impulsive events to be associated with a narrow CME produced by a solar jet (Kahler et al., 2001).

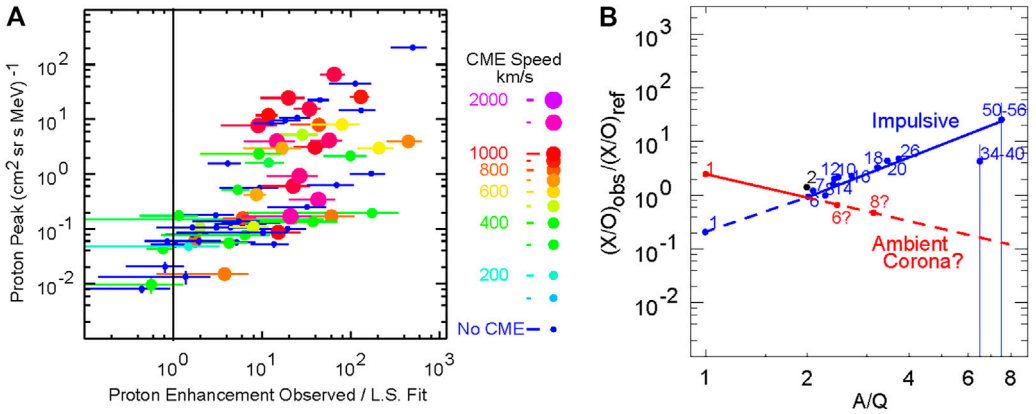


FIGURE 3 | Panel **(A)** shows the peak proton intensity at 2–2.5 MeV vs the proton excess relative to the $Z > 2$ power-law fit for impulsive SEP events with the symbol size and color determined by CME speed as shown. Events with fast CMEs have proton excesses. Panel **(B)** suggests two possible contributions to the plot of enhancement vs A/Q for the shock-enhanced SEP2 impulsive events: reaccelerated impulsive SEP1 seed particles (**blue**) and accelerated ambient coronal seed particles (**red**). Note that a shock wave can accelerate outside the volume that contains the SEP1 seed particles. If the shock is very strong, the ambient corona could dominate and the event could become a shock-dominated gradual event.

These SEP2 events are generally more intense and were often associated with CMEs fast enough (1,360 km s^{-1} in this case) to drive shock waves that could reaccelerate the SEP1 impulsive suprathermal ions from the earlier magnetic reconnection as well as ions, at least H ions and occasionally He, from the ambient plasma. The correlation of SEP proton intensity with proton excess is shown in **Figure 3** along with the suggested explanation of the excess.

At a temperature of 2.5–3 MK the elements ^{4}He and C are both fully ionized with $A/Z = 2$, so ^{4}He/C should represent the underlying coronal abundance, yet ~6% of impulsive SEP events are extremely ^{4}He poor (Reames 2019a) with ^{4}He/C ≈ 15 vs an average 137 ± 8. One such event is shown as Event 79 in the right panel of **Figure 2** where H, near background level, lies near the power-law fit of high-Z elements, but He lies well below it; another is shown as Event 3 in the central panel of **Figure 1**, where ^{4}He is obviously much closer to O and Fe than it is in other events. It has been suggested that the high first ionization potential (FIP) of He, of 24.6 eV can delay its ionization and elevation into the corona, but Ne with FIP = 21.6 eV is unaffected. Any possible Z^2/A effects from matter traversal would seem to suppress C more than ^{4}He. What causes this occasional ^{4}He poverty?

Note that we have studied variations of H/O and ^{4}He/O separately in the above and *not* H/^{4}He. Thus in **Figure 2** we identify separate processes where H variation from the extrapolated power-law fit of $Z > 2$ depends upon the presence and activity of fast shock waves, while ^{4}He (presumably) depends upon the especially high FIP of He. Studying the ratio H/^{4}He would have blurred together the effects of these extremely different physical processes. This important feature was not initially obvious.

Isotope resolution has been extended as high as Fe (e.g., Leske et al., 1999, 2007). These measurements show variations in A/Q that supplement similar measurements that have been shown with power-law fits using multiple elements.

Ionization States

Luhn et al. (1984, 1987) provided important early measurements of ionic charge states, Q, up to Fe for energies 0.34–1.8 MeV amu^{-1}. For large gradual events they found average values of Q_{Si} = 11.0 ± 0.3 and Q_{Fe} = 14.1 ± 0.2, appropriate for a temperatures of ≈2 MK. However, for ^{3}He-rich events they found Q_{Si} ≈ 14 and Q_{Fe} = 20.5 ± 1.2, which meant either a source temperature of ≈10 MK, or stripping of the ions after acceleration. How could the average abundances of Ne, Mg, and Si get enhanced if they were all fully ionized with $A/Q = 2$ during acceleration, just like He, C, and O? This dilemma was finally resolved when DiFabio et al. (2008) found that the ionization states in the impulsive SEP events increased with ion speed, suggesting that the ions had come to ionization equilibrium during traversal of a small amount of material *after* acceleration. DiFabio et al. (2008) concluded that acceleration must have occurred near 1.5 solar radii.

Acceleration Theory

Most of the early attempts to explain the enhancement of ^{3}He involved the preferential absorption of some form of wave energy, in resonance with the gyrofrequency of ^{3}He, to produce selective preheating that would enhance the tail of thermal distribution of ^{3}He so as to inject more ions into some unspecified acceleration mechanism. Ibragimov and Kocharov (1977) and Kocharov and Kocharov (1978, 1984) were first, considering ion-sound wave heating, but Weatherall (1984) pointed out that this could not account for all abundances. Fisk (1978) and Varvoglis and Papadopoulis (1983) suggested selective heating by absorption of electrostatic ion cyclotron waves. Winglee (1989) considered the ion-ion streaming instability to enhance heavy ions, and Riyopoulos (1991) considered electrostatic two-ion (H–^{4}He) hybrid waves.

Temerin and Roth (1992) considered the ubiquitous associated streaming electrons that produced electromagnetic ion cyclotron (EMIC) waves that were adequate to actually

resonantly accelerate the ^{3}He that absorbed the waves while mirroring in the magnetic field, in analogy with "ion conics" seen in the Earth's aurorae. Roth and Temerin (1997) suggested that heavier ions were enhanced through their second harmonic, but there was no smooth power law in *A/Q*. Miller et al. (1993a, b) considered the effects of other electron-beam-generated wave modes and Steinacker et al. (1997) considered warm-plasma broadening of spectral lines to produce a "He valley" of wave absorption.

Litvinenko (1996) considered effects of EMIC waves and Coulomb energy losses on the ^{3}He spectrum and Liu et al. (2004, 2006) were able to fit the complex spectra of ^{3}He and ^{4}He with a model of stochastic acceleration by a power-law spectrum of plasma-wave turbulence. Roughly speaking, as the ions are accelerated to higher energy, the ^{3}He begins to form a distinct energy peak as the ^{3}He in the source volume becomes depleted, while the ^{4}He does not.

Separately, explanation of the power-law dependence of enhancements on *A/Q* was first found by Drake et al. (2009) in particle-in-cell simulations of collapsing islands of magnetic reconnection. Ions are Fermi accelerated as they scatter back and forth from the ends of the collapsing islands. A similar process is found to accelerate electrons (Arnold et al., 2021) and its efficiency depends upon the strength of the out-of-plane guide field. However, ^{3}He has not been discussed at all in this process, although there would seem to be abundant opportunities for mirroring ions to absorb waves.

There is evidence that the contribution of ^{3}He is often saturated, i.e., all the ^{3}He in the accelerating volume is actually accelerated. Reames (1999) estimated the total number of energetic ^{3}He ions integrated over energy, space, and time and found it could be comparable with the number that exist in a typical flare volume. Subsequently, (Ho et al., 2005; Ho et al., 2019; Petrosian et al., 2009) found that this effect limited the maximum possible fluence of ^{3}He. It has been suggested (Kahler, private communication) that comparing the calculated number of SEP ^{3}He with the calculated number of ^{3}He in the observed volume of a jet or reconnection region, on an event-by-event basis, might help determine where and what fraction of the ^{3}He is accelerated. Is there enough ^{3}He in the reconnection volume or is it accelerated throughout the jet as the Temerin and Roth (1992) model may suggest? Can anything be said about possible second-harmonic acceleration of some heavy ions like Fe, or the rare, low-energy resonant peaks of Si or S (Mason et al., 2016), for which the low-energy spectra roll over like that of ^{3}He? Is SEP output bounded by the reconnection volume? Accelerating volumes are limited in impulsive SEPs while, in contrast, in gradual SEP events, source shock waves of huge area sweep out extensive volumes.

FLARES, CMES, SHOCKS, AND JETS

Reconnection of the solar magnetic field drives nearly all solar activity, and flares provide a ubiquitous visible marker of magnetically trapped heating from that reconnection. Perhaps it is not surprising that the earliest evidence of GLEs was associated with obvious, bright flares, but this association was taken much too literally.

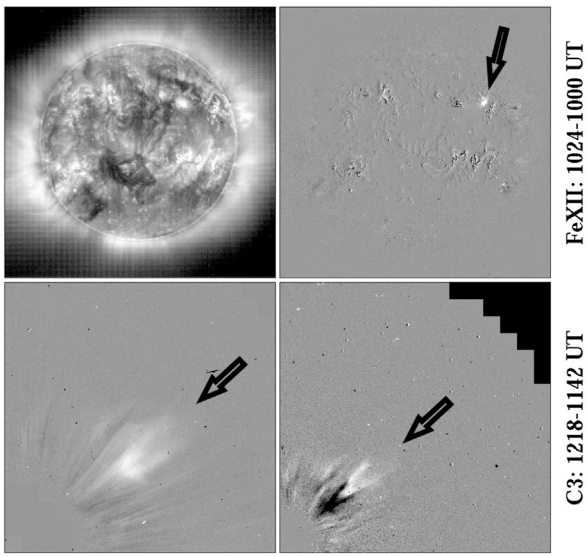


FIGURE 4 | Images of the impulsive SEP event of May 1, 2000 (see also **Figure 2**) are: **upper left** a full disk SOHO/EIT image at 195 Å *before* the flare, **upper right** subtracted EIT image at 195 Å with arrow showing new small white source in the NW quadrant, **lower left** subtracted image of NW quadrant of SOHO/LASCO/C2 coronagraph 2–6 R_S with arrow showing narrow CME, **lower right** subtracted NW quadrant of C3 image 4–30 R_S with arrow showing narrow CME (Kahler et al., 2001). Tracking and magnetic configuration for this event are shown by Nitta et al. (2006) and by Wang et al. (2006).

Radio data were the first to distinguish two acceleration mechanisms for SEPs in space (Wild et al., 1963), type III bursts that associate with electrons streaming from open magnetic reconnection in impulsive SEP events, and type II bursts that associate with shock acceleration and large gradual SEP events, but, unfortunately, the radio evidence was largely ignored in early SEP history.

After CMEs were identified and their observation became common, Kahler et al. (1984) found a 96% correlation between large SEP events and fast, wide CMEs. Shock waves driven by the CMEs could explain the extremely broad longitude span of the events. Mason et al. (1984) found that the minimal rigidity dependences of abundance variations across longitude were inconsistent with a point source origin and they discussed large-scale shock acceleration that they labelled LSSA. The importance of CMEs and shocks was not taken seriously by some flare enthusiasts until Gosling (1993, 1994) review article entitled "The solar flare myth" was published. This article was then found to "wage an assault on the last 30 years of solar-flare research" (Zirin, 1994) even though most published "solar-flare research" ignored SEPs entirely. However, the importance of CMEs and the existence of two mechanisms of SEP acceleration began to be recognized (Reames, 1988, 1995b, 1999, 2013, 2015, 2020, 2021a, 2021b; Zank et al., 2000, 2007; Kahler, 2001; Cliver et al., 2004; Lee, 2005; Gopalswamy et al.,

2012; Lee et al., 2012; Mewaldt et al., 2012; Desai and Giacalone, 2016; Kouloumvakos et al., 2019).

For a brief interval, distinguishing impulsive and gradual SEP events seemed simple. Impulsive events were ^{3}He-rich and gradual events were not. Then Mason et al. (1999) found a small but significant enhancement of ^{3}He in a large SEP event that would be called gradual in all other respects. It became clear that shock waves in gradual SEP events could reaccelerate residual suprathermal ions left over from previous impulsive SEP events. In fact these pre-accelerated ions might be preferred in some cases, as in quasi-perpendicular shocks when ions needed to overtake the shock from downstream (Tylka et al., 2001, 2005; Tylka and Lee, 2006). In fact, it became evident that large pools of ^{3}He-rich, Fe-rich suprathermal ions were extremely common, and available for shocks to traverse (Richardson et al., 1990; Desai et al., 2003; Wiedenbeck et al., 2008; Bučík et al., 2014, 2015; Chen et al., 2015). Whenever there was no large SEP event in progress the default suprathermal ion abundances below ~1 MeV amu^{-1} seemed to be ^{3}He-rich and Fe-rich, suggesting a large number of small unresolved jets (nanojets?) could generate SEPs faster than the solar wind could sweep them away. Thus, for 24% of gradual events, the $Z \geq 2$ elements are dominated by reaccelerated ions from impulsive events, called SEP3 events (Reames, 2020), and in 69% of gradual events the shock predominantly accelerates ions, even $Z > 2$ ions, from the ambient coronal plasma (SEP4 events; Reames, 2020a; Reames, 2021a; Reames, 2021b).

Impulsive SEP events were small and difficult to associate with coronal features, but Kahler et al. (2001) were able to associate several of the larger impulsive SEP events with narrow CMEs. **Figure 4** shows a narrow (54°), fast (1,360 km s^{-1}) CME from the impulsive event of May 1, 2000 SEP, seen by the Solar and Heliospheric Observatory (SOHO; https://sohowww.nascom.nasa.gov/), for which SEPs were shown in the **central** panel of **Figure 2**. Such CMEs had been associated with type III bursts and solar jets (Shimojo and Shibata, 2000). Unlike flares, jets involve magnetic reconnection on open field lines so the SEPs (and the CMEs) easily escape. Tracking of impulsive SEPs back to the Sun led to jets, often on the boundary between active regions and coronal holes (Wang et al., 2006; Nitta et al., 2006, 2015). However, onset times are poorly defined in small SEP events, making associations with type III bursts, etc. more difficult. Reconnection is often triggered by large-scale waves moving across the corona (Bučík et al., 2016). It is now possible to consider the nature of the associated jets directly for many ^{3}He-rich events (Bučík et al., 2018a, 2018b, 2021; see review; Bučík, 2020). It has even been possible to compare the temperatures derived from the extreme ultraviolet (EUV) images of coronal source regions of 24 solar jets associated with ^{3}He-rich events (Bučík et al., 2021) with the abundance-derived temperatures from SEPs (Reames et al., 2014b).

However, impulsive SEP events are not necessarily derived purely from magnetic reconnection; the CMEs from these jets are often fast enough to drive shock waves which can reaccelerate SEPs from the reconnection as is certainly the case for the 1,360 km s^{-1} CME in the May 1, 2000 event shown in **Figure 4**. It may be possible to distinguish the pure (SEP1) events from those reaccelerated by a shock (SEP2) from the proton excess produced in the latter events as shown in **Figure 3**.

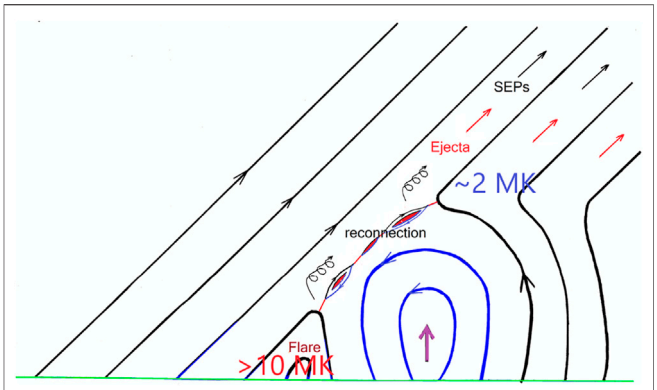


FIGURE 5 | A jet is produced when newly emerging magnetic flux (*blue*) reconnects with oppositely directed field **(*black*)** in the ***red*** region. This reconnection region is not a uniform surface but forms multiple islands of reconnection. Energetic particles and plasma can escape toward the ***upper right*** and a newly-enclosed flaring region labeled "Flare" forms at the ***lower left*** that is heated by trapping SEPs. Real jets can be much more complex, involving twisted fields, etc.

A sketch showing a jet formed by newly emerging magnetic field is shown in **Figure 5**. When the emerging field has opposite polarity from that of the preexisting field, reconnection takes place, not at a single point but in a series of "islands." In realistic jets, the opposing fields rarely exactly cancel, but rather leave an out-of-plane, residual "guide field." Particle acceleration occurs as particles, mainly electrons (Arnold et al., 2021), are Fermi-accelerated as they pitch-angle scatter in the evolving fields. As the SEPs and CME plasma are ejected on open field lines at the upper right in **Figure 5**, newly closing loops capture some SEPs in the lower left region labeled "flare" where they deposit their energy as heat. While much of the jet, including the SEPs, retain temperatures of 2–3 MK and emit EUV (Reames et al., 2014b; Bučík et al., 2021), the "flare" region with its captured SEP energy heats to 10–20 MK and emits X-rays. These X-rays often provide the source location. It is important to realize that the presence of the X-rays does not define the nature of the SEPs in space; both environments coexist since the reconnection that opens some field lines closes others. Some day someone will be able to provide us with an overlay of X-ray and EUV images that will map details of these source regions. As reconnection events become larger and more complex, both X-ray and EUV regions increase, contributing to a "big flare syndrome" (Kahler 1982), i.e., a misleading correlation between X-rays and the SEPs in space.

What about solar flares? There is evidence from measurements of Doppler-broadened γ-ray lines that the ions accelerated in large solar flares are ^{3}He-rich (Mandzhavidze et al., 1999; Murphy et al., 2016) and Fe-rich (Murphy et al., 1991), just like the impulsive SEPs from jets that we see in space. The SEPs accelerated on closed loops, dominated by electrons, soon scatter into the loss cone and plunge into the denser corona below, scattering against ions to produce X-ray bremstrahlung, and heating the plasma which expands back up into the loops creating a hot (>10 MK), bright flare. SEPs from jets are not hot (~2–3 MK; Reames et al., 2014b; Bučík et al., 2021) because the

SEPs and the CME plasma escape. Flares exist precisely because their energy is magnetically trapped and can only escape as heat, light, or neutral particles. Reconnection of closed fields with other closed fields cannot produce open fields, except when those fields are eventually carried outward by a CME.

Shock waves, driven by fast, wide CMEs are the basis of gradual SEP events. However, the narrow (<60°) CMEs emitted from solar jets are also fast enough to drive significant shocks, such as the 1,360 km s^{-1}, 54° width CME in the May 1, 2000 events discussed above. Yet, often, these fast, narrow CMEs from jets lack type II radio bursts that can signify shock acceleration, and the CME's contribution to SEP acceleration may be questioned (e.g., Bronarska et al., 2018; Kahler et al., 2019). At quasi-parallel shock waves, ions are accelerated by scattering back and forth across the shock against Alfvén waves (often self-generated), but non-relativistic electrons cannot resonate with Alfvén waves, so they can only be accelerated in the $\mathbf{V_S}$ x $\mathbf{B}$ electric field in more quasi-perpendicular regions of the shock. A big wide hemispherical shock surely has some regions where type II electrons get accelerated, while a narrow CME may produce only quasi-parallel regions—accelerating few electrons but plenty of ions. The shock from the May 1, 2000 event may not accelerate enough electrons, but it surely accelerates more ions than any of the shocks we have been able to observe directly *in situ* with shock speeds as low as 300 km s^{-1} (Reames 2012). The one feature that seems to show the presence or absence of shock acceleration in impulsive SEP events is the proton excess shown in **Figures 2**, **3** above (Reames 2019b), but the response of electrons and protons can be completely different.

Some people continue to defined "flare" to include every energetic phenomenon on the Sun, including flares, jets, CMEs, shocks, and even the SEPs at 1 AU. That, of course, makes flare research the most important discipline of all, studying the cause of everything solar, by definition. This deliberately blurs the physics, elevating the importance of flares and diminishing that of CMEs; this was exactly Gosling (1993) objection to "The Solar Flare Myth." Applying "flare" to everything is not only meaningless but helps no one understand any physics. Beginning with Carrington (1860), many of us still think of "the flare" as that localized sudden bright flash of white light, Hα, or X-ray emission driven because reconnection energy is trapped in closed magnetic loops on the Sun—an event recorded by the Solar Flare Patrol and documented by its timing, latitude, longitude, and C-M-X-scale soft X-ray intensity. These flares are limited in spatial extent. In contrast, the shock wave that accelerates SEPs in a large gradual event is a nominally hemispherical structure initially active from 2 to over 3 R_S (Reames 2009a, b; Cliver et al., 2004), thus enclosing an accelerating volume perhaps ~10 times the volume of the Sun itself—this is *not* a flare. Thus the source of SEPs has a major effect on their spatial distributions, among many other things, and lumping all possible sources together, to inflate the egos of a few flare researchers, is no help. please distinguish CMEs from flares in your publications.

This field began with a frequent assumption that all SEPs were somehow actually accelerated in flares. First we found that SEPs in gradual events were accelerated by CME-driven shocks instead. Then we found that the remaining impulsive SEPs in space came from jets. Now we find that it may be more correct to say that flares are caused by SEPs trapped on loops than the converse.

AUTHOR CONTRIBUTIONS

All work on this article was performed by DR.

ACKNOWLEDGMENTS

The author thanks Steve Kahler for helpful comments on this article.

REFERENCES

Arnold, H., Drake, J., Swisdak, M., Guo, F., Dahlin, J., Chen, B., et al. (2021). Electron Acceleration during Macroscale Magnetic Reconnection. *Phys. Rev. Lett.* 126, 13. doi:10.1103/PhysRevLett.126.135101

Bertsch, D. L., Fichtel, C. E., and Reames, D. V. (1969). Relative Abundance of Iron-Group Nuclei in Solar Cosmic Rays. *ApJ* 157, L53. doi:10.1086/180383

Bronarska, K., Wheatland, M. S., Gopalswamy, N., and Michalek, G. (2018). Very Narrow Coronal Mass Ejections Producing Solar Energetic Particles. *A&A* 619, A34. doi:10.1051/0004-6361/201833237

Bučík, R., Innes, D. E., Mall, U., Korth, A., Mason, G. M., and Gómez-Herrero, R. (2014). Multi-Spacecraft Observations of Recurrent ^{3}He-Rich Solar Energetic Particles. *ApJ* 786, 71. doi:10.1088/0004-637X/786/1/71

Bučík, R., Innes, D. E., Chen, N. H., Mason, G. M., Gómez-Herrero, R., and Wiedenbeck, M. E. (2015). Long-lived Energetic Particle Source Regions on the Sun. *J. Phys. Conf. Ser.* 642, 012002. doi:10.1088/1742-6596/642/1/012002

Bučík, R., Innes, D. E., Mason, G. M., and Wiedenbeck, M. E. (2016). Association of ^{3}He-Rich Solar Energetic Particles With Large-Scale Coronal Waves. *ApJ* 833, 63. doi:10.3847/1538-4357/833/1/63

Bučík, R., Innes, D. E., Mason, G. M., Wiedenbeck, M. E., Gómez-Herrero, R., and Nitta, N. V. (2018a). ^{3}He-rich Solar Energetic Particles in Helical Jets on the Sun. *ApJ* 852, 76. doi:10.3847/1538-4357/aa9d8f

Bučík, R., Wiedenbeck, M. E., Mason, G. M., Gómez-Herrero, R., Nitta, N. V., and Wang, L. (2018b). ^{3}He Rich Solar Energetic Particles from Sunspot Jets. *ApJ* 869, L21. doi:10.3847/2041-8213/aaf37f

Bučík, R. (2020). ^{3}He-Rich Solar Energetic Particles: Solar Sources. *Space Sci. Rev.* 216, 24. doi:10.1007/s11214-020-00650-5

Bučík, R., Mulay, S. M., Mason, G. M., Nitta, N. V., Desai, M. I., and Dayeh, M. A. (2021). Temperature in Solar Sources of ^{3}He-Rich Solar Energetic Particles and Relation to Ion Abundances. *ApJ* 908, 243. doi:10.3847/1538-4357/abd62d

Carrington, R. C. (1859). Description of a Singular Appearance Seen in the Sun on September 1, 1859. *Monthly Notices R. Astronomical Soc.* 20, 13–15. doi:10.1093/mnras/20.1.13

Chen, N.-h., Bučík, R., Innes, D. E., and Mason, G. M. (2015). Case Studies of Multi-day ^{3}He-Rich Solar Energetic Particle Periods. *A&A* 580, A16. doi:10.1051/0004-6361/201525618

Cliver, E. W., Kahler, S. W., and Reames, D. V. (2004). Coronal Shocks and Solar Energetic Proton Events. *ApJ* 605, 902–910. doi:10.1086/382651

Cook, W. R., Stone, E. C., and Vogt, R. E. (1984). Elemental Composition of Solar Energetic Particles. *ApJ* 279, 827. doi:10.1086/161953

Desai, M. I., Mason, G. M., Dwyer, J. R., Mazur, J. E., Gold, R. E., Krimigis, S. M., et al. (2003). Evidence for a Suprathermal Seed Population of Heavy Ions Accelerated by Interplanetary Shocks Near 1 AU. *ApJ* 588, 1149–1162. doi:10.1086/374310

Desai, M., and Giacalone, J. (2016). Large Gradual Solar Energetic Particle Events. *Living Rev. Sol. Phys.* 13. doi:10.1007/s41116-016-0002-5

DiFabio, R., Guo, Z., Möbius, E., Klecker, B., Kucharek, H., Mason, G. M., et al. (2008). Energy-dependent Charge States and Their Connection with Ion Abundances in Impulsive Solar Energetic Particle Events. *Astrophysical J.* 687, 623–634. doi:10.1086/591833

Drake, J. F., Cassak, P. A., Shay, M. A., Swisdak, M., and Quataert, E. (2009). A Magnetic Reconnection Mechanism for Ion Acceleration and Abundance Enhancements in Impulsive Flares. *ApJ* 700, L16–L20. doi:10.1088/0004-637X/700/1/L16

Fichtel, C. E., and Guss, D. E. (1961). Heavy Nuclei in Solar Cosmic Rays. *Phys. Rev. Lett.* 6, 495–497. doi:10.1103/PhysRevLett.6.495

Fisk, L. A. (1978). ^{3}He-Rich Flares - A Possible Explanation. *ApJ* 224, 1048. doi:10.1086/156456

Forbush, S. E. (1946). Three Unusual Cosmic-Ray Increases Possibly Due to Charged Particles from the Sun. *Phys. Rev.* 70, 771–772. doi:10.1103/PhysRev.70.771

Gloeckler, G., Hovestadt, D., Vollmer, O., and Fan, C. Y. (1975). Unusual Emission of Iron Nuclei from the Sun. *ApJ* 200, L45. doi:10.1086/181893

Gopalswamy, N., Xie, H., Yashiro, S., Akiyama, S., Mäkelä, P., and Usoskin, I. G. (2012). Properties of Ground Level Enhancement Events and the Associated Solar Eruptions during Solar Cycle 23. *Space Sci. Rev.* 171, 23–60. doi:10.1007/s11214-012-9890-4

Gosling, J. T. (1993). The Solar Flare Myth. *J. Geophys. Res.* 98, 18937–18949. doi:10.1029/93JA01896

Gosling, J. T. (1994). Correction to "The Solar Flare Myth". *J. Geophys. Res.* 99, 4259. doi:10.1029/94JA00015

Ho, G. C., Roelof, E. C., and Mason, G. M. (2005). The Upper Limit on ^{3}He Fluence in Solar Energetic Particle Events. *ApJ* 621, L141–L144. doi:10.1086/429251

Ho, G. C., Mason, G. M., and Allen, R. C. (2019). ^{3}He-Rich Solar Energetic Particle Events with No Measurable ^{4}He Intensity Increases. *Sol. Phys.* 294, 33. doi:10.1007/s11207-019-1420-z

Hsieh, K. C., and Simpson, J. A. (1970). Galactic^{3}He above 10 MeV Per Nucleon and the Solar Contributions of Hydrogen and Helium. *ApJ* 162, L197. doi:10.1086/180653

Ibragimov, I. A., and Kocharov, G. E. (1977). Possible Mechanism for Enrichment of Solar Cosmic Rays by Helium-Three and Heavy Nuclei. *15th Int. Conf. Cosmic Rays (Plovdiv: Bulgarian Acad. Sciences)* 11, 340.

Kahler, S. W. (1982). The Role of the Big Flare Syndrome in Correlations of Solar Energetic Proton Fluxes and Associated Microwave Burst Parameters. *J. Geophys. Res.* 87, 3439. doi:10.1029/JA087iA05p03439

Kahler, S. W., Sheeley, N. R., Jr., Howard, R. A., Koomen, M. J., Michels, D. J., McGuire, R. E., et al. (1984). Associations between Coronal Mass Ejections and Solar Energetic Proton Events. *J. Geophys. Res.* 89, 9683. doi:10.1029/JA089iA11p09683

Kahler, S. W. (2001). The Correlation between Solar Energetic Particle Peak Intensities and Speeds of Coronal Mass Ejections: Effects of Ambient Particle Intensities and Energy Spectra. *J. Geophys. Res.* 106, 20947–20955. doi:10.1029/2000JA002231

Kahler, S. W., Reames, D. V., and Sheeley, Jr., N. R., Jr. (2001). Coronal Mass Ejections Associated with Impulsive Solar Energetic Particle Events. *ApJ* 562, 558–565. doi:10.1086/323847

Kahler, S. W., Ling, A. G., and Gopalswamy, N. (2019). Are Solar Energetic Particle Events and Type II Bursts Associated with Fast and Narrow Coronal Mass Ejections?. *Sol. Phys.* 294, 134. doi:10.1007/s11207-019-1518-3

Kocharov, L. G., and Kocharov, G. E. (1984). ^{3}He-rich Solar Flares. *Space Sci. Rev.* 38, 89. doi:10.1007/BF00180337

Kocharov, G. E., and Kocharov, L. G. (1978). "Present State of Experimental and Theoretical Investigations of Solar Events Enriched by Helium-3" in Proc. 10th Leningrad Sympos. on Cosmic Rays, Leningrad. Leningrad: A. F. Yoffe Phys.-Tech. Inst., 37.

Kouloumvakos, A., Rouillard, A. P., Wu, Y., Vainio, R., Vourlidas, A., Plotnikov, I., et al. (2019). Connecting the Properties of Coronal Shock Waves with Those of Solar Energetic Particles. *ApJ* 876, 80. doi:10.3847/1538-4357/ab15d7

Lee, M. A. (2005). Coupled Hydromagnetic Wave Excitation and Ion Acceleration at an Evolving Coronal/interplanetary Shock. *Astrophys J. Suppl. S* 158, 38–67. doi:10.1086/428753

Lee, M. A., Mewaldt, R. A., and Giacalone, J. (2012). Shock Acceleration of Ions in the Heliosphere. *Space Sci. Rev.* 173, 247–281. doi:10.1007/s11214-012-9932-y

Leske, R. A., Mewaldt, R. A., Cohen, C. M. S., Cummings, A. C., Stone, E. C., Wiedenbeck, M. E., et al. (1999). Event-to-event Variations in the Isotopic Composition of Neon in Solar Energetic Particle Events. *Geophys. Res. Lett.* 26, 2693–2696. doi:10.1029/1999GL900561

Leske, R. A., Mewaldt, R. A., Cohen, C. M. S., Cummings, A. C., Stone, E. C., Wiedenbeck, M. E., et al. (2007). Solar Isotopic Composition as Determined Using Solar Energetic Particles. *Space Sci. Rev.* 130, 195–205. doi:10.1007/s11214-007-9185-3

Lin, R. P. (1970). The Emission and Propagation of ~40 KeV Solar Flare Electrons. *Sol. Phys.* 12, 266. doi:10.1007/BF00227122

Lin, R. P. (1974). Non-relativistic Solar Electrons. *Space Sci. Rev.* 16, 189. doi:10.1007/BF00240886

Litvinenko, Y. E. (1996). "On the Formation of the Helium-3 Spectrum in Impulsive Solar Flares," in *High Energy Solar Physics, AIP Conf. Proc. 374*. Editors R. Ramaty, N. Mandzhavidze, and X.-M. Hua (Woodbury, NY: AIP Press), 498. doi:10.1063/1.50985

Liu, S., Petrosian, V., and Mason, G. M. (2004). Stochastic Acceleration of ^{3}He and ^{4}He by Parallel Propagating Plasma Waves. *ApJ* 613, L81–L84. doi:10.1086/425070

Liu, S., Petrosian, V., and Mason, G. M. (2006). Stochastic Acceleration of ^{3}He and4He in Solar Flares by Parallel-propagating Plasma Waves: General Results. *ApJ* 636, 462–474. doi:10.1086/497883

Luhn, A., Klecker, B., Hovestadt, D., Gloeckler, G., Ipavich, F. M., Scholer, M., et al. (1984). Ionic Charge States of N, Ne, Mg, Si and S in Solar Energetic Particle Events. *Adv. Space Res.* 4, 161–164. doi:10.1016/0273-1177(84)90307-7

Luhn, A., Klecker, B., Hovestadt, D., and Moebius, E. (1987). The Mean Ionic Charge of Silicon in He-3-Rich Solar Flares. *ApJ* 317, 951. doi:10.1086/165343

Mandzhavidze, N., Ramaty, R., and Kozlovsky, B. (1999). Determination of the Abundances of Subcoronal ^{4}He and of Solar Flare-Accelerated ^{3}He and ^{4}He from Gamma-ray Spectroscopy. *Astrophys. J.* 518, 918. doi:10.1086/307321

Mason, G. M., Gloeckler, G., and Hovestadt, D. (1984). Temporal Variations of Nucleonic Abundances in Solar Flare Energetic Particle Events. II - Evidence for Large-Scale Shock Acceleration. *ApJ* 280, 902. doi:10.1086/162066

Mason, G. M., Reames, D. V., von Rosenvinge, T. T., Klecker, B., and Hovestadt, D. (1986). The Heavy-Ion Compositional Signature in ^{3}He-Rich Solar Particle Events. *ApJ* 303, 849. doi:10.1086/164133

Mason, G. M., Mazur, J. E., and Dwyer, J. R. (1999). ^{3}He Enhancements in Large Solar Energetic Particle Events. *Astrophys. J. Lett.* 525, L133–L136. doi:10.1086/312349

Mason, G. M., Mazur, J. E., Dwyer, J. R., Jokipii, J. R., Gold, R. E., and Krimigis, S. M. (2004). Abundances of Heavy and Ultraheavy Ions in ^{3}He-rich Solar Flares. *ApJ* 606, 555–564. doi:10.1086/382864

Mason, G. M. (2007). ^{3}He-Rich Solar Energetic Particle Events. *Space Sci. Rev.* 130, 231–242. doi:10.1007/s11214-007-9156-8

Mason, G. M., Nitta, N. V., Wiedenbeck, M. E., and Innes, D. E. (2016). Evidence for a Common Acceleration Mechanism for Enrichments of ^{3}He And Heavy Ions in Impulsive Sep Events. *ApJ* 823, 138. doi:10.3847/0004-637X/823/2/138

McGuire, R. E., von Rosenvinge, T. T., and McDonald, F. B. (1979). A Survey of Solar Cosmic ray Composition. *Proc. 16th Int. Cosmic Ray Conf*, 5, 61..

Mewaldt, R. A., Looper, M. D., Cohen, C. M. S., Haggerty, D. K., Labrador, A. W., Leske, R. A., et al. (2012). Energy Spectra, Composition, and Other Properties of Ground-Level Events during Solar Cycle 23. *Space Sci. Rev.* 171, 97–120. doi:10.1007/s11214-012-9884-2

Miller, J., Viñas, A., and Reames, D. V. (1993a). Selective ^{3}He and Fe Acceleration in Impulsive Solar Flares. *23rd Intl. Cosmic-ray Conf. (Calgary)* 3, 13.

Miller, J., Viñas, A., and Reames, D. V. (1993b). Heavy Ion Acceleration and Abundance Enhancements in Impulsive Solar Flares. *23rd Intl. Cosmic-ray Conf. (Calgary)* 3, 17.

Mogro-Campero, A., and Simpson, J. A. (1972). Enrichment of Very Heavy Nuclei in the Composition of Solar Accelerated Particles. *ApJ* 171, L5. doi:10.1086/180856

Murphy, R. J., Ramaty, R., Reames, D. V., and Kozlovsky, B. (1991). Solar Abundances from Gamma-ray Spectroscopy - Comparisons with Energetic Particle, Photospheric, and Coronal Abundances. *ApJ* 371, 793. doi:10.1086/169944

Murphy, R. J., Kozlovsky, B., and Share, G. H. (2016). Evidence for Enhanced^{3}He in Flare-Accelerated Particles Based on New Calculations of the Gamma-ray Line Spectrum. *ApJ* 833, 196. doi:10.3847/1538-4357/833/2/196

Nitta, N. V., Reames, D. V., DeRosa, M. L., Liu, Y., Yashiro, S., and Gopalswamy, N. (2006). Solar Sources of Impulsive Solar Energetic

Particle Events and Their Magnetic Field Connection to the Earth. *ApJ* 650, 438–450. doi:10.1086/507442

Nitta, N. V., Mason, G. M., Wang, L., Cohen, C. M. S., and Wiedenbeck, M. E. (2015). Solar Sources of ^{3}He-Rich Solar Energetic Particle Events in Solar Cycle 24. *ApJ* 806, 235. doi:10.1088/0004-637X/806/2/235

Petrosian, V., Jiang, Y. W., Liu, S., Ho, G. C., and Mason, G. M. (2009). Relative Distributions of Fluences of ^{3} He and ^{4}He In Solar Energetic ParticleS. *ApJ* 701, 1–7. doi:10.1088/0004-637X/701/1/1

Reames, D. V., von Rosenvinge, T. T., and Lin, R. P. (1985). Solar ^{3}He-Rich Events and Nonrelativistic Electron Events - A New Association. *ApJ* 292, 716. doi:10.1086/163203

Reames, D. V., and Stone, R. G. (1986). The Identification of Solar He-3-Rich Events and the Study of Particle Acceleration at the Sun. *ApJ* 308, 902. doi:10.1086/164560

Reames, D. V. (1988). Bimodal Abundances in the Energetic Particles of Solar and Interplanetary Origin. *ApJ* 330, L71. doi:10.1086/185207

Reames, D. V., Meyer, J. P., and von Rosenvinge, T. T. (1994). Energetic-particle Abundances in Impulsive Solar Flare Events. *ApJS* 90, 649. doi:10.1086/191887

Reames, D. V. (1995a). Coronal Abundances Determined from Energetic Particles. *Adv. Space Res.* 15 (7), 41–51. doi:10.1016/0273-1177(94)00018-v

Reames, D. V. (1995b). Solar Energetic Particles: A Paradigm Shift. *Rev. Geophys.* 33, 585. doi:10.1029/95RG00188

Reames, D. V. (1999). Particle Acceleration at the Sun and in the Heliosphere. *Space Sci. Rev.* 90, 413–491. doi:10.1023/A:1005105831781

Reames, D. V. (2000). Abundances of Trans-iron Elements in Solar Energetic Particle Events. *Astrophys. J. Lett.* 540, L111–L114. doi:10.1086/312886

Reames, D. V., and Ng, C. K. (2004). Heavy-Element Abundances in Solar Energetic Particle Events. *ApJ* 610, 510–522. doi:10.1088/0004-637X/723/2/128610.1086/421518

Reames, D. V. (2009a). Solar Release Times of Energetic Particles in Ground-Level Events. *ApJ* 693, 812–821. doi:10.1088/0004-637X/693/1/812

Reames, D. V. (2009b). Solar Energetic-Particle Release Times in Historic Ground-Level Events. *ApJ* 706, 844–850. doi:10.1088/0004-637X/706/1/844

Reames, D. V. (2012). Particle Energy Spectra at Traveling Interplanetary Shock Waves. *ApJ* 757, 93. doi:10.1088/0004-637X/757/1/93

Reames, D. V. (2013). The Two Sources of Solar Energetic Particles. *Space Sci. Rev.* 175, 53–92. doi:10.1007/s11214-013-9958-9

Reames, D. V. (2014). Element Abundances in Solar Energetic Particles and the Solar corona. *Sol. Phys.* 289, 977–993. doi:10.1007/s11207-013-0350-4

Reames, D. V., Cliver, E. W., and Kahler, S. W. (2014a). Abundance Enhancements in Impulsive Solar Energetic-Particle Events with Associated Coronal Mass Ejections. *Sol. Phys.* 289, 3817–3841. doi:10.1007/s11207-014-0547-1

Reames, D. V., Cliver, E. W., and Kahler, S. W. (2014b). Variations in Abundance Enhancements in Impulsive Solar Energetic-Particle Events and Related CMEs and Flares. *Sol. Phys.* 289, 4675–4689. doi:10.1007/s11207-014-0589-4

Reames, D. V. (2015). What Are the Sources of Solar Energetic Particles? Element Abundances and Source Plasma Temperatures. *Space Sci. Rev.* 194, 303–327. doi:10.1007/s11214-015-0210-7

Reames, D. V. (2016). Temperature of the Source Plasma in Gradual Solar Energetic Particle Events. *Solar Phys.* 291. doi:10.1007/s11207-016-0854-9

Reames, D. V. (2018). Abundances, Ionization States, Temperatures, and FIP in Solar Energetic Particles. *Space Sci. Rev.* 214, 61. doi:10.1007/s11214-018-0495-4

Reames, D. V. (2019a). Helium Suppression in Impulsive Solar Energetic-Particle Events. *Sol. Phys.* 294, 32. doi:10.1007/s11207-019-1422-x

Reames, D. V. (2019b). Hydrogen and the Abundances of Elements in Impulsive Solar Energetic-Particle Events. *Sol. Phys.* 294, 37. doi:10.1007/s11207-019-1427-5

Reames, D. V. (2020). Four Distinct Pathways to the Element Abundances in Solar Energetic Particles. *Space Sci. Rev.* 216, 20. doi:10.1007/s11214-020-0643-5

Reames, D. V. (2021a). Solar Energetic Particles. in *Lec. Notes Phys 978*. Second Edition (Cham, Switzerland: Springer Nature). doi:10.1007/978-3-030-66402-2

Reames, D. V. (2021b). Sixty Years of Element Abundance Measurements in Solar Energetic Particles. *Space Sci. Rev.* 217, 72. doi:10.1007/s11214-021-00845-4

Richardson, I. G., Reames, D. V., Wenzel, K.-P., and Rodriguez-Pacheco, J. (1990). Quiet-time Properties of Low-Energy (Less Than 10 MeV Per Nucleon) Interplanetary Ions during Solar Maximum and Solar Minimum. *ApJ* 363, L9. doi:10.1086/185853

Riyopoulos, S. (1991). Subthreshold Stochastic Diffusion with Application to Selective Acceleration of ^{3}He in Solar Flares. *ApJ* 381, 578. doi:10.1086/170682

Roth, I., and Temerin, M. (1997). Enrichment of ^{3}He and Heavy Ions in Impulsive Solar Flares. *ApJ* 477, 940–957. doi:10.1086/303731

Serlemitsos, A. T., and Balasubrahmanyan, V. K. (1975). Solar Particle Events with Anomalously Large Relative Abundance of ^{3}He. *ApJ* 198, 195. doi:10.1086/153592

Shimojo, M., and Shibata, K. (2000). Physical Parameters of Solar X-Ray Jets. *ApJ* 542, 1100–1108. doi:10.1086/317024

Shirk, E. K., and Price, P. B. (1974). Observation of Transiron Solar-Flare Nuclei in an Apollo 16 Command Module Window. *ApJ* 190, 695. doi:10.1086/152927

Steinacker, J., Meyer, J. P., Steinacker, A., and Reames, D. V. (1997). The Helium Valley: Comparison of Impulsive Solar Flare Ion Abundances and Gyroresonant Acceleration with Oblique Turbulence in a Hot Multi-Ion Plasma. *ApJ* 476, 403–427. doi:10.1086/303589

Temerin, M., and Roth, I. (1992). The Production of ^{3}He and Heavy Ion Enrichment in ^{3}He-Rich Flares by Electromagnetic Hydrogen Cyclotron Waves. *ApJ* 391, L105. doi:10.1086/186408

Tylka, A. J., Cohen, C. M. S., Dietrich, W. F., Maclennan, C. G., McGuire, R. E., Ng, C. K., et al. (2001). Evidence for Remnant Flare Suprathermals in the Source Population of Solar Energetic Particles in the 2000 Bastille Day Event. *Astrophys. J. Lett.* 558, L59–L63. doi:10.1086/323344

Tylka, A. J., Cohen, C. M. S., Dietrich, W. F., Lee, M. A., Maclennan, C. G., Mewaldt, R. A., et al. (2005). Shock Geometry, Seed Populations, and the Origin of Variable Elemental Composition at High Energies in Large Gradual Solar Particle Events. *ApJ* 625, 474–495. doi:10.1086/429384

Tylka, A. J., and Lee, M. A. (2006). A Model for Spectral and Compositional Variability at High Energies in Large, Gradual Solar Particle Events. *ApJ* 646, 1319–1334. doi:10.1086/505106

Varvoglis, H., and Papadopoulos, K. (1983). Selective Nonresonant Acceleration of He-3(2+) and Heavy Ions by H(+) Cyclotron Waves. *ApJ* 270, L95. doi:10.1086/184077

Wang, Y. M., Pick, M., and Mason, G. M. (2006). Coronal Holes, Jets, and the Origin of ^{3}He-rich Particle Events. *ApJ* 639, 495–509. doi:10.1086/499355

Weatherall, J. (1984). Turbulent Heating in Solar Cosmic-ray Theory. *ApJ* 281, 468. doi:10.1086/162119

Wiedenbeck, M. E., Cohen, C. M. S., Cummings, A. C., de Nolfo, G. A., Leske, R. A., Mewaldt, R. A., et al. (2008). Persistent Energetic ^{3}He in the Inner Heliosphere. *Proc. 30th Int. Cosmic Ray Conf. (Mérida)* 1, 91.

Wild, J. P., Smerd, S. F., and Weiss, A. A. (1963). Solar Bursts. *Annu. Rev. Astron. Astrophys.* 1, 291–366. doi:10.1146/annurev.aa.01.090163.001451

Winglee, R. M. (1989). Heating and Acceleration of Heavy Ions during Solar Flares. *ApJ* 343, 511. doi:10.1086/167726

Zank, G. P., Rice, W. K. M., and Wu, C. C. (2000). Particle Acceleration and Coronal Mass Ejection Driven Shocks: A Theoretical Model. *J. Geophys. Res.* 105, 25079–25095. doi:10.1029/1999JA000455

Zank, G. P., Li, G., and Verkhoglyadova, O. (2007). Particle Acceleration at Interplanetary Shocks. *Space Sci. Rev.* 130, 255–272. doi:10.1007/s11214-007-9214-2

Zirin, H. (1994). Solar Storminess, Sky and Telescope. *Nov* 9, 9.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Reames. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Mini Review
PUBLISHED 09 November 2022
DOI 10.3389/fspas.2022.1052278

Check for updates

OPEN ACCESS

EDITED BY
Alberto Fairén,
Center for Astrobiology (CSIC), Spain

REVIEWED BY
Roberto Barbieri,
University of Bologna, Italy
Jean-Pierre Paul de Vera,
German Aerospace Center (DLR), Germany

*CORRESPONDENCE
Patrick Jung,
Patrick_jung90@web.de

SPECIALTY SECTION
This article was submitted to Astrobiology, a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 23 September 2022
ACCEPTED 26 October 2022
PUBLISHED 09 November 2022

CITATION
Jung P, Lehnert LW, Bendix J, Lentendu G, Grube M, Alfaro FD, Rio Cd, Gutiérrez Alvarado JL, van den Brink L and Lakatos M (2022), The grit crust: A poly-extremotolerant microbial community from the Atacama Desert as a model for astrobiology. *Front. Astron. Space Sci.* 9:1052278. doi: 10.3389/fspas.2022.1052278

COPYRIGHT
© 2022 Jung, Lehnert, Bendix, Lentendu, Grube, Alfaro, Rio, Gutiérrez Alvarado, van den Brink and Lakatos. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The grit crust: A poly-extremotolerant microbial community from the Atacama Desert as a model for astrobiology

Patrick Jung[1]*, Lukas W. Lehnert[2], Jörg Bendix[3], Guillaume Lentendu[4], Martin Grube[5], Fernando D. Alfaro[6], Camilo del Rio[7], José Luis Gutiérrez Alvarado[8], Liesbeth van den Brink[9] and Michael Lakatos[1]

[1]Integrative Biotechnology, University of Applied Sciences, Kaiserslautern, Germany, [2]Department of Geography, Ludwig-Maximilians-University, Munich, Germany, [3]Faculty of Geography, Laboratory for Climatology and Remote Sensing, Philipps-University of Marburg, Marburg, Germany, [4]Institute of Biology, Laboratory of Soil Biodiversity, University of Neuchâtel, Neuchâtel, Switzerland, [5]Institute of Biology, University of Graz, Graz, Austria, [6]GEMA Center for Genomics, Ecology & Environment, Universidad Mayor, Santiago, Chile, [7]Instituto de Geografía & Centro UC Desierto de Atacama, Pontificia Universidad Católica de Chile, Santiago, Chile, [8]Corporación Nacional Forestal (CONAF), Santiago, Chile, [9]Plant Ecology Group University of Tübingen, Tübingen, Germany

The grit crust is a recently discovered, novel type of biocrust made of prokaryotic cyanobacteria, eukaryotic green algae, fungi, lichens and other microbes that grow around and within granitoid stone pebbles of about 6 mm diameter in the Coastal Range of the Atacama Desert, Chile. The microbial community is very well adapted towards the extreme conditions of the Atacama Desert, such as the highest irradiation of the planet, strong temperature amplitudes and steep wet-dry cycles. It also has several other striking features making this biocrust unique compared to biocrusts known from other arid biomes on Earth. It has already been shown that the grit crust mediates various bio-weathering activities in its natural habitat. These activities prime soil for higher organisms in a way that can be envisioned as a proxy for general processes shaping even extra-terrestrial landscapes. This mini-review highlights the potential of the grit crust as a model for astrobiology in terms of extra-terrestrial microbial colonization and biotechnological applications that support human colonization of planets.

KEYWORDS

grit crust, astrobiology, Atacama, Mars, cyanobacteria, lichens, green algae

Introduction

Terrestrial extremophile microorganisms have ever since inspired hypotheses about life strategies in extra-terrestrial environments, e.g., on Mars, and about their applications in biotechnology-driven space activities. Such extremophiles usually include several hetero- and chemoautotrophic bacteria, archaea and fungi as well as photoautotrophic cyanobacteria and eukaryotic algae (Liu et al., 2008; Beblo-Vranesevic et al., 2018; Sundarasami et al., 2019; Coleine and Delgado-Baquerizo, 2022). A specific aspect of extreme environments has frequently been overlooked in this context: the symbiotic lifestyle, which is perfectly exemplified by biological soil crusts (hereafter: biocrusts). Moreover, the close coexistence and interacting of the biocenosis in a confined space also supports the response to sudden abiotic environmental fluctuations or changes and gives the community overall better resilience through different physiological use and response to change. These symbiotic associations of microorganisms are mainly found in hot and cold deserts across the globe, where they are often the main biological component colonizing the soil's surface (Weber et al., 2022). Soil particles are aggregated through the presence and activity of this often extremo-tolerant biota that desiccates regularly. The resulting living crust mantels the ground's surface as a coherent layer. It was estimated that biocrusts cover about 12% of the Earth's terrestrial surface and about 30% of all dryland soils (Rodriguez-Caballero et al., 2018). In these extreme environments, the microbial consortium regularly faces long periods of desiccation, intense UV radiation, and severe nutrient and resource limitations (Belnap et al., 2008; Munzi et al., 2019).

To overcome environmental limitations biocrust organisms have developed several strategies such as extracellular polymeric substances (EPS) that help them to retain water (Mugnai et al., 2018), UV-protecting pigments such as scytonemin (Miralles et al., 2017) or specific compounds including sucrose and trehalose (Xin et al., 2015; Baubin et al., 2021) that minimize metabolic desiccation stress. Most of these strategies have not yet been fully explored, and a plethora of novel genes and metabolites is expected from biocrust-forming microorganisms from extreme environments which can be used to obtain a wide variety of products using biotechnological processes (Lakatos and Strieht, 2017). Cyanobacteria and green algae are already optimal candidates for present biotechnological application on Earth due to low needs for maintenance (Schweiger et al., 2022). Therefore, this micro-algal-based technology could also be part of regenerative life support systems in space.

Beside such applications, biocrust-forming microorganisms as early terrestrial microbial associations on Earth can also help to address ancient humanity's questions, such as whether we are alone or how and where did life appear and evolve? This again has to do with their extremo-tolerant character, and the early time of their origins or their ability to interact with each other. Cyanobacteria, for example, which evolved during the late Archaean (Sánchez-Baracaldo and Cardona, 2020), can be found in virtually any environment on Earth and in symbiosis with fungi (Rikkinen, 2015) and plants (Chang et al., 2019). These microorganisms evolved at a time where the atmospheric composition of our planet and radiation income of the Sun were different. The Archaean Earth was devoid of oxygen and enriched by greenhouse gases (Hessler et al., 2004) including an increased exposition to higher fluxes of UV-B (280–320 nm) and UV-C (200–280 nm) radiation (Cockell, 2002). However, cyanobacteria could not only cope with these conditions (Shih, 2015), they also supposedly led to the Great Oxygenation Event during which they changed Earth's atmosphere and became ancestral to all other eukaryotic photosynthetic organisms such as algae and plants (Sánchez-Baracaldo and Cardona, 2020). This implies that cyanobacteria (and other microorganisms) resemble concepts of simple life forms that have a high adaptation potential to extreme conditions. Thus, we imagine that relatable life forms might be expected beyond Earth even if the abiotic conditions differ. Whether they are alive or extinct, they have and they still left their footprints on Earth, and we can learn to interpret the pattern they left as biosignatures to support the search for life on other planets at least in the Solar System and particular on planets such as Mars. Cyanobacteria, for example, are well known for their ability to form stromatolites on shores and lakes ranging in height from millimeters to tens of meters and extending laterally from small lenses to aggregates of hundreds of kilometers wide as mineralized domes (Neilan et al., 2002). If cyanobacteria alone leave such landscape-dominating traces behind in aquatic environments over millennia, how pronounced must then their traces be when they are an integral part of biocrusts together with lichens, green algae and other microorganisms that cover 12% of the Earth's terrestrial surface?

Recently, a novel type of biocrust was discovered in the Atacama Desert, one of the world's oldest and driest deserts, covering most of its floor: the grit crust (Jung et al., 2020a). The grit crust's microbial community is unique amongst biocrusts found in all other arid parts of the world because cyanobacteria, eukaryotic green algae, lichens and specific fungi grow on, in, and around small granitoid quartz pebbles called grits (locally called "maicillo") of about 6 mm in diameter. These structures can be seen with naked eyes and cause a blackish, landscape filling pattern, on the otherwise bare ground. The Atacama Desert is extremely interesting in that sense as it resembles an environment close to what is expected to be found on other planetary bodies, such as Mars, in terms of bio-geochemistry, nutrient composition or topological similarities - or at least mimics particular characters (Navarro-González, et al., 2003; Schmidt, et al., 2018). This can be explained by the unique combination of its environmental extremes such as its long lasting history of hyper-aridity, the highest UV radiation levels on Earth, oxidizing soils and other factors that can be

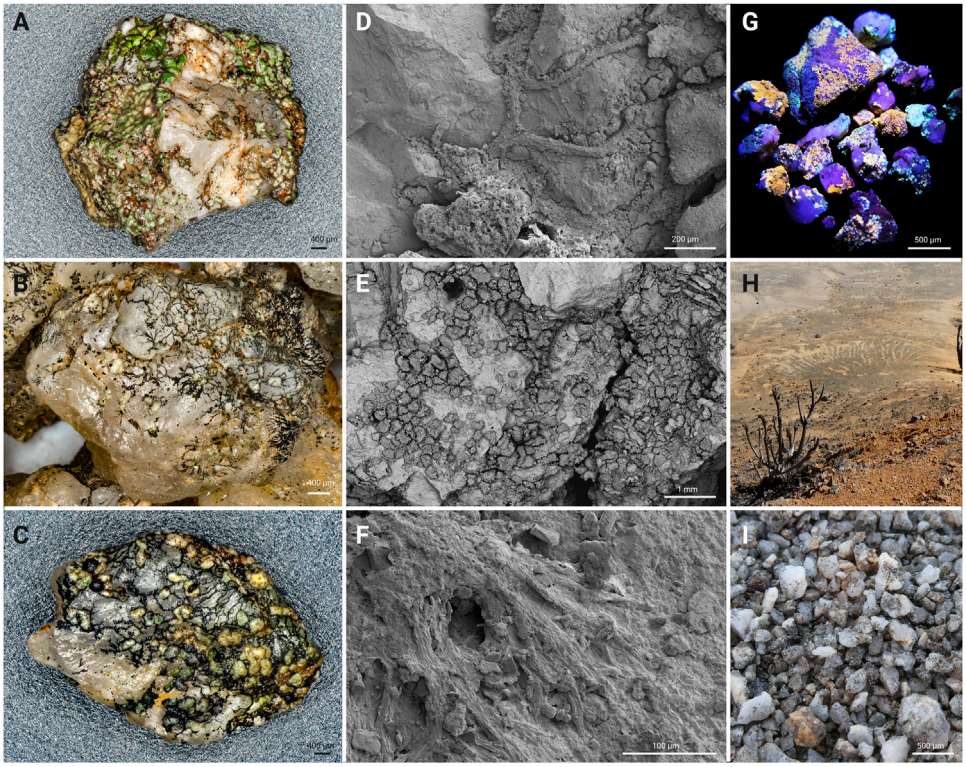


FIGURE 1
Potential fossilization traces of the grit crust depicted by various techniques. **(A–C)** micrographs of single grit stones in their hydrated state with lichens and fungi creating a vein-like network of micro-structures on the grit's surface. **(D–F)** scanning electron microscopy showing fungal trench lines in **(D)**, lichen squamules covered by mineral particles on grit stones in **(E)** and lichen pro-thallus in **(F)**. **(G–I)** photographs of the grit crust with glowing lichens in G caused by reflectance of secondary lichen substances under UV light, erosion rifts in **(H)** and a grit crust overview image with grits overgrown by various concatenating microorganisms in **(I)**.

compared with the conditions present on Mars (Azua-Bustos et al., 2022). For these reasons, an increasing number of experiments spanning from testing technologies over to experiments in the life sciences have been conducted at the Atacama Desert (Sobron et al., 2012; Abarca et al., 2013; Gunes-Lasnet et al., 2014; Schulze-Makuch et al., 2018). Thus, studying the survival and microbial diversity of the grit crust of the Atacama Desert during future studies will provide valuable insights into understanding the evolution of life, the habitability beyond Earth, and biotechnological applications of these microorganisms that could support space exploration missions.

Currently, the research project "Grit Life" funded by the German Research Foundation, aims to untangle the microbiota of the grit crust based on metabarcoding data applied to a recovery experiment over several years. This will not only give detailed insights into the diversity of bacteria, eukaryotes, and fungi but also tracks their re-colonization potential. The DNA-based approach is accompanied by monitoring climate and micro-climate data of the grit crust environment so that the extremophile character of the microbiota can be determined. In addition, fungi, eukaryotic green algae, and cyanobacteria of the grit crust will successively be isolated and curated in the context of a microbial archive in order to guarantee future investigations, e.g., on the microbe's biotechnological potential. As such, this project is the first of its kind that will generate a microbial archive of biocrusts where the extremophile properties of the microorganisms are characterized, allowing for transfer into biotechnological approaches. Exploiting this unique biocenosis will be discussed in the following in terms of the search for life beyond Earth based on novel potential fossilization structures known from the grit crust (Figure 1), and the utilization of the biocenosis' microorganisms for biotechnological approaches in space (Figure 2).

Suitability of the grit crust's ecology for the colonization of other planets

The grit crust covers a roughly 40 km broad strip, 2.5 km off the Coastal Range of the southern Atacama Desert (~26°S), spanning over 1.500 km to the northern Atacama Desert. In these areas, it forms blackish patterns on the desert's ground

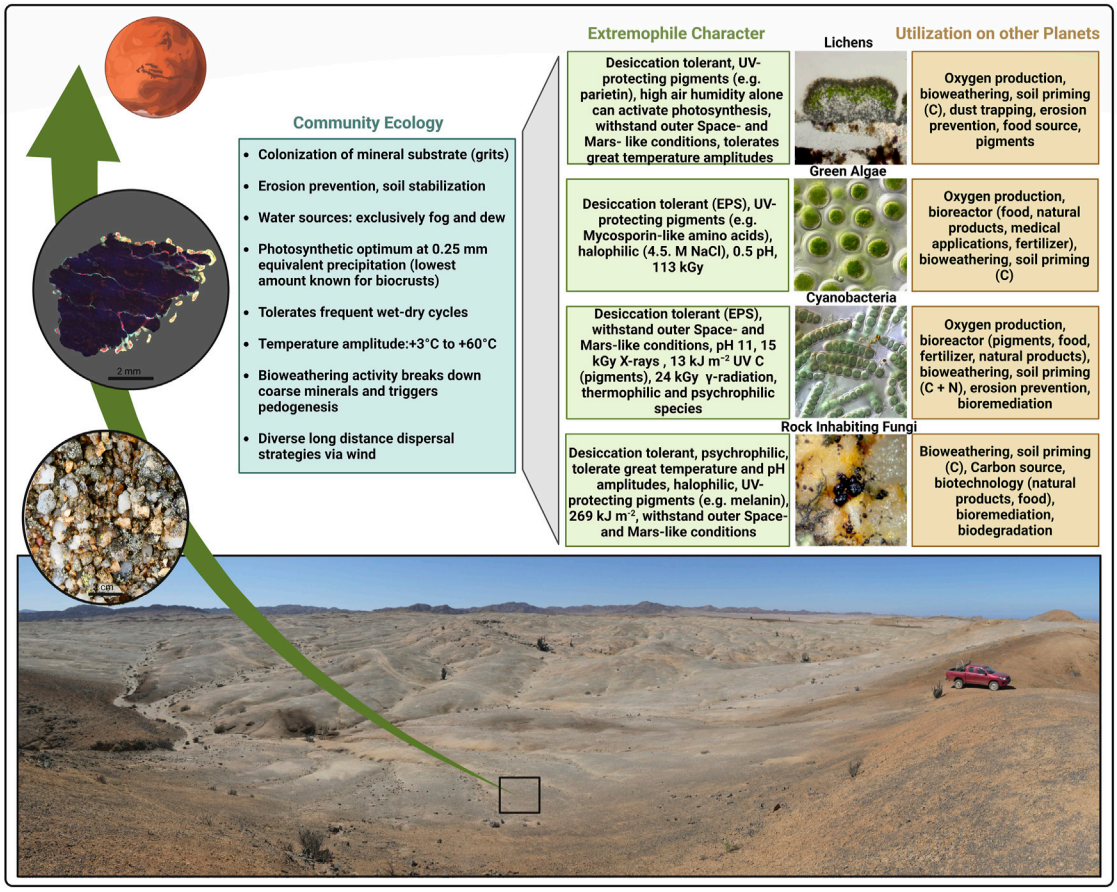


FIGURE 2
Ecology, extremophile character and possible applications of the grit crust in the context of astrobiology. The bottom picture shows the landscape of the coastal Atacama Desert with blackish patterns on the ground formed by a high microbial colonization degree. The bottom circle on the left shows a close-up photography of the ground with the grits colonized by diverse microbes such as lichens. The top circle shows a thin section of a single grit under the microscope using autofluorescence where fungal mycelium (lichens) appear whitish while algae appear red, which also colonize inner structures of the stone.

caused by a high degree of microbial colonization of the otherwise whitish granitoid quartz pebbles of approximately 6 mm diameter. These blackish patterns are so rich in organic material that they are well visible by the naked eye in the landscape. The reason for this is the cyclic appearance of fog and dew water input that leads to biomass growth equivalent to about 140 mg chlorophyll$_{a+b}$ m^{-2} and more (Jung et al., 2020a). These high biomass stocks are reached in a xeric environment where soil temperature can exceed 60 °C during summer days and falls below 3 °C during winter times. Average water input from dew is only 0.025–0.088 mm per day, water input from fog is 0.38–1.25 mm per day, and intense irradiation, with maxima of more than 1,500 µmol photons m^2 s^{-1} is frequent (Jung et al., 2020a). In general, the Atacama Desert is considered as a radiation hotspot where the highest irradiation on the planet was reported (Rondanelli et al., 2015). Constituents of the microbial community are diverse micro-lichens (*Buellia* spp., *Pleopsidium chlorophanum*, *Tetramelas pulverulentus*, *Aspicilia* spp., *Diploschistes* spp.), with green algal photobionts (*Trebouxia* spp., *Myrmecia* spp.), free living eukaryotic green algae (*Stichococcus* spp., *Pseudostichococcus* spp., *Chlorella* spp., *Klebsormidium* spp.) and cyanobacteria (*Nostoc* spp., *Microcoleus vaginatus*, *Kastovskya adunca*, *Myxacorys chilensis*, *Pleurocapsa* spp.), as well as melanized micro-colonial fungi (*Constantinomyces* spp., *Lichenothelia* spp.) (Jung et al., 2020a). These harsh conditions under which they thrive may indicate that the microbial community building the grit crust - or at least individual strains contained therein - might be suitable candidates for testing (their) vitality in outer Space or under Martian conditions (Coleine and Delgado-Baquerizo, 2022). Here, it is also worth to mention that unicellular life forms evolved early in the Earth's history. On the other hand, the climate of the Atacama Desert is considered to be stable for at least 150 million years (Hartley at al., 2005), which leads to the

idea that the microbial life forms that jointly build the grit crust of the Atacama Desert are well adapted to extreme environments. Although other microbial consortia are known from hyper-arid regions of the Atacama Desert (e.g. Schulze-Makuch et al., 2018), the grit crust is exposed to low amounts of fog in areas devoid of any higher vegetation (Jung et al., 2020a) and it is the only microbial assemblage which fills entire landscapes of hundreds of square kilometers (Jung et al., 2020a). This demonstrates the ecophysiological adaptation potential of the grit crust on the large scale environment which might be beneficial for microbial colonization of other planets with relatable conditions.

Surviving prolonged desiccation is a common feature of all organisms mentioned here, but in addition, it has been shown that the photosynthetic activity of the grit crust is strongly inhibited when more than the equivalent of 0.25 mm water is present in the organisms (Jung et al., 2020a). This amount also marks the range in which the net photosynthesis of the microbial community is at its optimum. At the same time, this is the lowest amount of water for optimum photosynthesis detected for biocrusts on Earth (Jung et al., 2020a). Amongst biocrusts, the grit crust from the Atacama Desert is also unique in its adaptation to wet-dry cycles caused by several fog events during a single day because these cyclic conditions have been shown to be harmful for other biocrusts (Belnap et al., 2004; Reed et al., 2012). This implies that the grit crust's microbial community could be ideal for biotechnological applications outside its natural environment, such as humanly induced microbial colonization of other planets, where cyclic artificial wetting will for sure be essential, but in the case of the grit crust - compared to other biocrusts - not lethal due to their specialized desiccation tolerance.

One of the biggest challenges during human colonization of other planets has been the formation of soil as a weathering product of mainly unweathered rocks found on other planets. Access to nutrient rich soil would subsequently allow the growth of microorganisms and/or plants in order to support human life. The grit crust is the first biocrust detected on a pure mineral, coarse substrate, while most other biocrust types are established on soil, a product of a later stage in the weathering of minerals (Weber et al., 2022). This resembles another benefit of the grit crust compared with other biocrusts because coarse minerals are the predominant substrate found on, e.g., Mars and the Earth's moon. Mars' surface consists mostly of coarse regolith of largely weathered basalt (McSween et al., 2009), which contains the basic macro elements (C, H, O, N, P, S, K, Mg, Na, and Ca) and minor elements (Mn, Cr, Ni, Mo, Cu, Fe, and Zn) essential for life (Greenwood et al., 2007; Stern et al., 2015; Verseux et al., 2016). However, due to the lack of organic carbon, the basaltic regolith substrate has a poor water-holding capacity and limited nutritional bioavailability (Wamelink et al., 2014).

Despite colonization of the quartz pebbles' surface, the grit crust microorganisms also inhabit the undersurface and naturally occurring cracks where they are protected from harsh UV radiation but can still perform photosynthesis due to the transparent to translucent character of the stones. It has also been shown that these microorganisms actively mediate bio-weathering activities regarding nutrient acquisition, which can lead to the formation of a terrestrial protopedon as initial pedogenesis (Jung et al., 2020b). This terrestrial protopedon is a fine substrate enriched in nitrogen from cyanobacterial nitrogen fixation and carbon from photosynthesis that primes the substrate for the establishment of other organisms and leaches into deeper layers (Dojani et al., 2007; Young et al., 2022). Triggering this succession is a well-known feature of biocrusts worldwide (Belnap and Weber, 2013; Barger et al., 2016; Bao et al., 2019), but the grit crust is the only biocrust type on Earth where minerals are the starting material and not (the more weathered) sand or soil. At the same time, the leaked or extracellular polymeric substances (EPS) of cyanobacteria, for example, together with the filamentous nature of some algae, the lichens and fungi concatenate the first millimeters of the substrate that protects it from erosion by, e.g., wind (Chamizo et al., 2017). This leads to the idea that the grit crust might have the potential to enrich mineral substrates of other planets over the long term and protects it from strong wind erosion, which is crucial on some Solar System bodies, e.g. Mars, where wind speeds can be immense (White, 1979; Viúdez-Moreiras et al., 2019). However, such strong winds could also be beneficial for colonizing great landscapes on other planets by the grit crust organisms, as these winds produce a lot of dust that can be transported over wide distances (e.g., Ferri et al., 2003), on which microbes can hitchhike. The Martian atmosphere is full of dust, with loads that greatly fluctuate with the year's season (Toigo and Richardson, 2000), so aeolian transport, including microbes, is a likely scenario that could support the microbial colonization of other planets that support human life. In the Atacama Desert, for example, it has been shown that certain soil-borne microorganisms were transported more than 100 km off the coast towards the hyper-arid core of the Atacama (Azua-Bustos et al., 2019).

Microalgae, lichens and fungi of the grit crust as models for astrobiology

On Earth, phototrophic extremophiles have been found in hostile environments, such as hot and cold deserts, hot, acidic, or alkaline springs, caves, or Antarctic sub-glacial lakes considered as Earth analogs of Mars and the icy moons of Jupiter and Saturn (Martins et al., 2017; Coleine and Delgado-Baquerizo, 2022). One of the most interesting microalgae occupying various ecological niches of such extreme habitats is the prokaryotic cyanobacterial order Chroococcidiopsidales, which has been a focal point for astrobiology research. For specific strains of the genus Chroococcidiopsis *sensu lato*, for example, it has been shown that they tolerate at least 4 years of air-drying (Billi, 2009;

Fagliarone et al., 2017), up to 13 kJ m^{-2} of UV C radiation (Baqué et al., 2013), 15 kGy of X-rays (Billi et al., 2000) and 12 kGy of γ-radiation (Verseux et al., 2016; Verseux et al., 2017). They even have an enhanced radiation resistance, coping with 30 kJ m^{-2} of a simulated Martian UV flux (Cockell et al., 2005), 24 kGy of γ-radiation and 2 kGy of Fe ions (Verseux et al., 2016). In their dried state, these strains also survived 2.5 years of exposure to Mars and Space-like conditions (Billi, 2019). Also a comparative analysis of whole-genome sequences showed no increased variant numbers in the space-derivate compared to triplicates of the reference strain maintained on the ground (Napoli et al., 2022). Closely related species that also belong to the order Chroococcidiopsidales, such as *Aliterella chasmolithica* (Jung et al., 2020c) and *Gloeocapsopsis diffluens* (Jung et al., 2021) were found in the grit crust environment, and a relative of the latter has been shown to produce great amounts of trehalose (145 µg per mg dry weight) and sucrose (64 µg per mg dry weight) as desiccation protection along with copious amounts of EPS (Azua-Bustos et al., 2019). This implies that these organisms have strong capabilities to protect themselves from damage during prolonged desiccation periods, a great characteristic for exposure to Space-like conditions during travel or terraformation approaches. Although these results were achieved for several strains of the order Chroococcidiopsidales, they likely also apply to other cyanobacteria—especially if they were isolated from desert environments such as the grit crust, where most of these extreme conditions come together. In addition to Chroococcidiopsidales, other cyanobacteria such as *Pleurocapsa* spp., the endemic *Kastovskya adunca*, *Myxacorys chilensis* and several *Nostoc* species have been found to be part of the biocenosis of the grit crust (Jung et al., 2019) and are expected to possess similar desiccation resistance functions.

In contrast to the great insights on the survivability of prokaryotic cyanobacteria exposed to Space- or Mars-like conditions, little is known about these capabilities for eukaryotic green algae. One of the most notorious green algae is the spherical and highly abundant genus *Chlorella* which can adapt to a range of extreme and harsh terrestrial conditions (Krienitz et al., 2015). Members of this genus isolated from desert sand crusts can resist intense UV radiation and γ-rays (Treves et al., 2013). In contrast to cyanobacteria, experiments, where *Chlorella* from desert environments was exposed to a Mars-like near-space environment showed that they survived but got significantly damaged on a sub-cellular and physiological level (Wang et al., 2021). We already detected several *Chlorella* and *Chlorella*-like types as frequent constituents of the grit crust, including *Pseudochlorella* spp. or *Pseudostichococcus* spp. which share similar characteristics. However, *Chlorella* is already used in large-scale bioreactors on Earth for commercial interests due to its ability to quickly adapt to cultivation conditions, because of the significant amounts of biomass generated in a short time, its CO_2 capturing ability and role as a food source (Coronado-Reyes et al., 2020). For these reasons, there are various ideas and concepts that suggest such green algae to be an integral part of lunar bases (Detrel, 2021) or on space stations (Helisch et al., 2016; Niederwieser et al., 2018), where a photobioreactor volume of 482 L has been shown to be suitable to ensure a daily CO_2 reprocessing for one crew member (Helisch et al., 2020).

Another component of the grit crust from the Coastal Range of the Atacama Desert are lichens, which -in general- represent a micro-ecosystem built by a fungus and an algal photobiont partner, including a vast microbiome (Hawksworth and Grube, 2020). Several lichen species have already been detected in the grit crust belonging to the genera *Aspicilia*, *Buellia* and *Pleopsidium* (Jung et al., 2020a), including some micro-colonial fungi of the genera *Lichenothelia* and *Constantinomyces* that are poly-extremotolerant (Gostinčar et al., 2012; Grube et al., 2013) and can facultatively be in symbiotic relationship with free-living eukaryotic green algae (Muggia et al., 2013). The lichen *Pleopsidium chlorophanum*, also present in the grit crust, was exposed uninterruptedly to simulated conditions of protected Martian surface niche conditions (269 kJ m^{-2}) for 34 days. The lichen not only survived and remained photosynthetically active but even adapted physiologically by increasing its photosynthetic activity over 34 days (de Vera et al., 2014). Similar results have been achieved for numerous other lichens (de Vera, 2012; Meeßen et al., 2013; Brandt et al., 2015; De la Torre Noetzel et al., 2017). In addition, black micro-colonial fungi (MCF), also referred to as rock-inhabiting fungi (RIF) (Sterflinger, 1998; Gorbushina, 2003), can cope with extremes of temperatures, acidity, osmotic stress and salinity, dehydration, solar and UV radiation and even radioactivity (Sterflinger 1998; Gorbushina, 2003; Onofri et al., 2007), with some of them being able to survive even real space exposure and simulated Martian conditions (Onofri et al., 2007). This indicates that lichens, their photobionts and RIFs as extremophile organisms, in general, can be great indicators to determine and prime the habitability of a planet and for the search of possible life-supporting surroundings on planets like Mars (Armstrong, 2019; Joseph et al., 2020).

Not only the singular microbes of the grit crust represent promising candidates for astrobiology but also their ecology as a community can give new impulses: the grit crust creates unique micro- and macroscopic patterns that can potentially get fossilized over time and that have until now not been included in the search for life on other planets. Paying attention to these structures is crucial and can complement investigations of fossilized microbial structures on other bodies of the Solar System or exoplanets. The structures of the grit crust microbes such as lichen thalli and fungal- or algal filaments, for example, can trap mineral particles rich in calcium carbonate that can fossilize as a venous network

(Figures 1A–C). A similar pattern might be created by trench lines generated by fungi that etch the quartz surface (Figure 1D). The later is a consequence of bio-weathering activity which has already been shown for the grit crust (Jung et al., 2020b). In addition, water erosion based on occasional heavy rain events play a crucial role in the Atacama Desert (Alcayaga et al., 2022). As the organisms concatenate the mineral substrate, the grits can create small gutters where excess water runs off, forming a slightly wavy micro-topography (Figure 1H). These "waves" are quite stable due to high contents of calcium carbonate that concatenates the waves which can later be fossilized over time. These and other structures can be visualized by various techniques such as simple photography using the full light spectrum or UV-radiation, scanning electron microscopy or stereoscopic micrographs (Figure 1), all of which are common for space missions (Joseph et al., 2020).

Concluding remarks

The Coastal Range of the Atacama Desert has recently been identified as the home to an outstanding biocenosis called grit crust, with a tremendous potential for various aspects of astrobiology. The ecology of the grit crust, including its extremophile microbial constituents, are unique amongst biocrusts on Earth that can be beneficial for human colonization of other planets or rock bodies such as the Earth's Moon or Mars. The new research project Grit Life aims to untangle the full microbial biodiversity of this grit environment, including isolating prominent microorganisms that will support future research in the context of astrobiology and space sciences. During these studies, the grit crust's extremophiles can be used to test their suitability during mass cultivation in photobioreactors (food-, oxygen source for crewed missions), screenings for Chlorophyll f (allows photosynthesis of cyanobacteria in low light environments such as on other planets) and symbiotic interactions (expands the knowledge on ancient life forms). For these reasons, the grit crust as a consortium of various microorganisms on a mineral substrate opens up a new opportunity to test hypotheses and ideas in the context of astrobiology that surpasses other biocrusts known on Earth.

Author contributions

PJ prepared the first draft of the manuscript and the figures. All authors revised the manuscript and approved the final.

Funding

The project "Grit Life" is funded by the German Research Foundation (DFG) with the grant number JU 3228/1-1. ML was funded by the Ministry of Science and Health Rhineland-Palatinate (PhytoBioTech, 724-0116#2021/004-1501 15405) and by the Federal Ministry of Education and Research (W2V-Strategy2Value, 03WIR4502A). FA and CR were supported by FONDECYT-1220358 and FONDECYT-11200789 respectively, Agencia Nacional de Investigación y Desarrollo (ANID), Chile.

Acknowledgments

The authors want to thank CONAF (Corporación Nacional Forestal) for their help in the context of this research project and the kind contributions of the rangers. Images were prepared with the great help of Biorender.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abarca, M., Diaz, A., de las Casas, H., Lazarte, A., Trujillo, S., and Onchi, E. (2013). *Establishment of a Mars analogue research station in pampas de la Joya (Atacama Desert in southern Peru)*. Lakewood, Colorado, U.S: The Mars Society.

Alcayaga, H., Soto-Alvarez, M., Laronne, J. B., Caamaño, D., Mao, L., and Urrutia, R. (2022). Runoff volume and sediment yield from an endorheic watershed generated by rare rainfall events in the Atacama Desert. *Geomorphology* 400, 108107. doi:10.1016/j.geomorph.2021.108107

Armstrong, R. A. (2019). The lichen symbiosis: Lichen "extremophiles" and survival on Mars. *J. Astrobiol. Space Sci. Rev.* 1, 378–397.

Azua-Bustos, A., González-Silva, C., and Fairén, A. G. (2022). The Atacama Desert in northern Chile as an analog model of Mars. *Front. Astron. Space Sci.* 8, 810426. doi:10.3389/fspas.2021.810426

Azua-Bustos, A., González-Silva, C., Fernández-Martínez, M. Á., Arenas-Fajardo, C., Fonseca, R., Martín-Torres, F. J., et al. (2019). Aeolian transport of viable microbial life across the Atacama Desert, Chile: Implications for Mars. *Sci. Rep.* 9 (1), 1–11. doi:10.1038/s41598-019-47394-z

Bao, T., Zhao, Y., Gao, L., Yang, Q., and Yang, K. (2019). Moss-dominated biocrusts improve the structural diversity of underlying soil microbial communities

by increasing soil stability and fertility in the Loess Plateau region of China. *Eur. J. Soil Biol.* 95, 103120. doi:10.1016/j.ejsobi.2019.103120

Baqué, M., Viaggiu, E., Scalzi, G., and Billi, D. (2013). Endurance of the endolithic desert cyanobacterium Chroococcidiopsis under UVC radiation. *Extremophiles* 17, 161–169. doi:10.1007/s00792-012-0505-5

Barger, N. N., Weber, B., Garcia-Pichel, F., Zaady, E., and Belnap, J. (2016). "Patterns and controls on nitrogen cycling of biological soil crusts," in *Biological soil crusts: An organizing principle in drylands* (Cham: Springer), 257–285.

Baubin, C., Ran, N., Siebner, H., and Gillor, O. (2021). The response of desert biocrust bacterial communities to hydration-desiccation cycles. *SOIL Discuss.*, 1–48.

Beblo-Vranesevic, K., Bohmeier, M., Perras, A. K., Schwendner, P., Rabbow, E., Moissl-Eichinger, C., et al. (2018). Lack of correlation of desiccation and radiation tolerance in microorganisms from diverse extreme environments tested under anoxic conditions. *FEMS Microbiol. Lett.* 365 (6), fny044. doi:10.1093/femsle/fny044

Belnap, J., Phillips, S. L., Flint, S., Money, J., and Caldwell, M. (2008). Global change and biological soil crusts: Effects of ultraviolet augmentation under altered precipitation regimes and nitrogen additions. *Glob. Change Biol.* 14 (3), 670–686. doi:10.1111/j.1365-2486.2007.01509.x

Belnap, J., Phillips, S. L., and Miller, M. E. (2004). Response of desert biological soil crusts to alterations in precipitation frequency. *Oecologia* 141 (2), 306–316. doi:10.1007/s00442-003-1438-6

Belnap, J., and Weber, B. (2013). Biological soil crusts as an integral component of desert environments. *Ecol. Process.* 2 (1), 11–12. doi:10.1186/2192-1709-2-11

Billi, D., Friedmann, E. I., Hofer, K. G., Grilli Caiola, M., and Ocampo-Friedmann, R. (2000). Ionizing-radiation resistance in the desiccationtolerant cyanobacterium Chroococcidiopsis. *Appl. Environ. Microbiol.* 66, 1489–1492. doi:10.1128/aem.66.4.1489-1492.2000

Billi, D. (2009). Subcellular integrities in Chroococcidiopsis sp. CCMEE 029 survivors after prolonged desiccation revealed by molecular probes and genome stability assays. *Extremophiles* 13, 49–57. doi:10.1007/s00792-008-0196-0

Billi, D. (2019). Desert cyanobacteria under space and planetary simulations: A tool for searching for life beyond Earth and supporting human space exploration. *Int. J. Astrobiol.* 18 (5), 483–489. doi:10.1017/s147355041800037x

Brandt, A., de Vera, J. P., Onofri, S., and Ott, S. (2015). Viability of the lichen Xanthoria elegans and its symbionts after 18 months of space exposure and simulated Mars conditions on the ISS. *Int. J. Astrobiol.* 14 (3), 411–425. doi:10.1017/s1473550414000214

Chamizo, S., Rodríguez-Caballero, E., Román, J. R., and Cantón, Y. (2017). Effects of biocrust on soil erosion and organic carbon losses under natural rainfall. *Catena* 148, 117–125. doi:10.1016/j.catena.2016.06.017

Chang, A. C. G., Chen, T., Li, N., and Duan, J. (2019). Perspectives on endosymbiosis in coralloid roots: Association of cycads and cyanobacteria. *Front. Microbiol.* 10, 1888. doi:10.3389/fmicb.2019.01888

Cockell, C. S., Schuerger, A. C., Billi, D., Friedmann, E. I., and Panitz, C. (2005). Effects of a simulated martian UV flux on the cyanobacterium, Chroococcidiopsis sp. 029. *Astrobiology* 5, 127–140. doi:10.1089/ast.2005.5.127

Cockell, C. S. (2002). Photobiological uncertainties in the Archaean and post-Archaean world. *Int. J. Astrobiol.* 1 (1), 31–38. doi:10.1017/s1473550402001003

Coleine, C., and Delgado-Baquerizo, M. (2022). Unearthing terrestrial extreme microbiomes for searching terrestrial-like life in the Solar System. *Trends Microbiol.* 30, 1101–1115. doi:10.1016/j.tim.2022.04.002

Coronado-Reyes, J. A., Salazar-Torres, J. A., Juarez-Campos, B., and Gonzalez-Hernandez, J. C. (2020). *Chlorella* vulgaris, a microalgae important to be used in biotechnology: A review. *Food Sci. Technol.*

de la Torre Noetzel, R., Miller, B., Cubero, A. Z., Sancho, L. G., Jordão, L., Rabbow, E., and de Vera, J. P. (2017). *Survival of lichens on the ISS-II: Ultrastructural and morphological changes of circinaria gyrosa after space and mars-like conditions*.

de Vera, J. P. (2012). Lichens as survivors in space and on Mars. *Fungal Ecol.* 5 (4), 472–479. doi:10.1016/j.funeco.2012.01.008

de Vera, J. P., Schulze-Makuch, D., Khan, A., Lorek, A., Koncz, A., Möhlmann, D., et al. (2014). Adaptation of an Antarctic lichen to Martian niche conditions can occur within 34 days. *Planet. Space Sci.* 98, 182–190. doi:10.1016/j.pss.2013.07.014

Detrell, G. (2021). *Chlorella* vulgaris photobioreactor for oxygen and food production on a moon base—potential and challenges. *Front. Astron. Space Sci.* 8, 700579. doi:10.3389/fspas.2021.700579

Dojani, S., Lakatos, M., Rascher, U., Wanek, W., Lüttge, U., and Büdel, B. (2007). Nitrogen input by cyanobacterial biofilms of an inselberg into a tropical rainforest in French Guiana. *Flora - Morphol. Distribution Funct. Ecol. Plants* 202 (7), 521–529. doi:10.1016/j.flora.2006.12.001

Fagliarone, C., Mosca, C., Ubaldi, I., Verseux, C., Baqué, M., Wilmotte, A., et al. (2017). Avoidance of protein oxidation correlates with the desiccation and radiation resistance of hot and cold desert strains of the cyanobacterium Chroococcidiopsis. *Extremophiles* 21, 981–991. doi:10.1007/s00792-017-0957-8

Ferri, F., Smith, P. H., Lemmon, M., and Rennó, N. O. (2003). Dust devils as observed by Mars pathfinder. *J. -Geophys. -Res.* 108 (E12). doi:10.1029/2000je001421

Gorbushina, A. (2003). Microcolonial fungi: Survival potential of terrestrial vegetative structures. *Astrobiology* 3 (3), 543–554. doi:10.1089/153110703322610636

Gostinčar, C., Muggia, L., and Grube, M. (2012). Polyextremotolerant black fungi: Oligotrophism, adaptive potential, and a link to lichen symbioses. *Front. Microbiol.* 3, 390. doi:10.3389/fmicb.2012.00390

Greenwood, J. P., Blake, R. E., Barron, V., and Torrent, J. (2007). in *Phosphorous geochemistry of Mars: Evidence for an early acidic hydrosphere* (Pasadena, California: Seventh International Conference on Mars).

Grube, M., Muggia, L., and Gostinčar, C. (2013). "Niches and adaptations of polyextremotolerant black fungi," in *Polyextremophiles* (Dordrecht: Springer), 551–566.

Gunes-Lasnet, S., Kisidi, A., van Winnendael, M., Josset, J. L., Ciarletti, V., Barnes, D., and Chong Diaz, G. (2014). *Safer: The promising results of the Mars mission simulation campaign in Atacama, Chile.*

Hartley, A. J., Chong, G., Houston, J., and Mather, A. E. (2005). 150 million years of climatic stability: Evidence from the Atacama Desert, northern Chile. *J. Geol. Soc. Lond.* 162, 421–424. 10.1144/0016-764904-071.

Hawksworth, D. L., and Grube, M. (2020). Lichens redefined as complex ecosystems. *New Phytol.* 227 (5), 1281–1283. doi:10.1111/nph.16630

Helisch, H., Keppler, J., Bretschneider, J., Belz, S., Henn, N., Fasoulas, S., et al. (2016). Preparatory ground-based experiments on cultivation of *Chlorella* vulgaris for the ISS experiment PBR@ LSR. 46th. *Int. Conf. Environ. Syst.*

Helisch, H., Keppler, J., Detrell, G., Belz, S., Ewald, R., Fasoulas, S., et al. (2020). High density long-term cultivation of *Chlorella* vulgaris SAG 211-12 in a novel microgravity-capable membrane raceway photobioreactor for future bioregenerative life support in SPACE. *Life Sci. Space Res.* 24, 91–107. doi:10.1016/j.lssr.2019.08.001

Hessler, A. M., Lowe, D. R., Jones, R. L., and Bird, D. K. (2004). A lower limit for atmospheric carbon dioxide levels 3.2 billion years ago. *Nature* 428, 736–738. doi:10.1038/nature02471

Joseph, R. J., Graham, L., Büdel, B., Jung, P., Kidron, G. J., Latif, K., and Schild, R. (2020). Mars: Algae, lichens, fossils, minerals, microbial mats, and stromatolites in gale crater. *J. Astrobiol. Space Sci. Rev.* 3 (1), 40–111.

Jung, P., Azua-Bustos, A., Gonzalez-Silva, C., Mikhailyuk, T., Zabicki, D., Holzinger, A., et al. (2021). Emendation of the coccoid cyanobacterial genus Gloeocapsopsis and description of the new species Gloeocapsopsis diffluens sp. nov. And Gloeocapsopsis dulcis sp. nov. Isolated from the coastal range of the Atacama Desert (Chile). *Front. Microbiol.* 12, 671742. doi:10.3389/fmicb.2021.671742

Jung, P., Baumann, K., Emrich, D., Springer, A., Felde, V. J., Dultz, S., et al. (2020b). Lichens bite the dust–A bioweathering scenario in the Atacama Desert. *Iscience* 23 (11), 101647. doi:10.1016/j.isci.2020.101647

Jung, P., Baumann, K., Lehnert, L. W., Samolov, E., Achilles, S., Schermer, M., et al. (2020a). Desert breath—how fog promotes a novel type of soil biocenosis, forming the coastal Atacama Desert's living skin. *Geobiology* 18 (1), 113–124. doi:10.1111/gbi.12368

Jung, P., Mikhailyuk, T., Emrich, D., Baumann, K., Dultz, S., and Büdel, B. (2020c). Shifting boundaries: Ecological and geographical range extension based on three new species in the cyanobacterial genera Cyanocohniella, Oculatella, and, Aliterella. *J. Phycol.* 56 (5), 1216–1231. doi:10.1111/jpy.13025

Jung, P., Schermer, M., Briegel-Williams, L., Baumann, K., Leinweber, P., Karsten, U., et al. (2019). Water availability shapes edaphic and lithic cyanobacterial communities in the Atacama Desert. *J. Phycol.* 55 (6), 1306–1318. doi:10.1111/jpy.12908

Krienitz, L., Huss, V. A., and Bock, C. (2015). *Chlorella*: 125 years of the green survivalist. *Trends plant Sci.* 20 (2), 67–69. doi:10.1016/j.tplants.2014.11.005

Lakatos, M., and Strieht, D. (2017). Terrestrial microalgae—novel concepts for biotechnology and applications. *Progress in Botany* 79, 269–312.

Liu, Y., Cockell, C. S., Wang, G., Hu, C., Chen, L., and De Philippis, R. (2008). Control of lunar and martian dust—Experimental insights from artificial and natural cyanobacterial and algal crusts in the desert of inner Mongolia, China. *Astrobiology* 8 (1), 75–86. doi:10.1089/ast.2007.0122

Martins, Z., Cottin, H., Kotler, J. M., Carrasco, N., Cockell, C. S., de la Torre Noetzel, R., et al. (2017). Earth as a tool for astrobiology – A European perspective. *Space Sci. Rev.* 209, 43–81. doi:10.1007/s11214-017-0369-1

McSween, H. Y., Taylor, G. J., and Wyatt, M. B. (2009). Elemental composition of the Martian crust. *Science* 324 (5928), 736–739. doi:10.1126/science.1165871

Meeßen, J., Sánchez, F. J., Brandt, A., Balzer, E. M., de la Torre, R., Sancho, L. G., et al. (2013). Extremotolerance and resistance of lichens: Comparative studies on five species used in astrobiological research I. Morphological and anatomical characteristics. *Orig. Life Evol. Biosph.* 43 (3), 283–303. doi:10.1007/s11084-013-9337-2

Miralles, I., Jorge-Villar, S. E., van Wesemael, B., and Lázaro, R. (2017). Raman spectroscopy detection of biomolecules in biocrusts from differing environmental conditions. *Spectrochimica Acta Part A Mol. Biomol. Spectrosc.* 171, 40–51. doi:10.1016/j.saa.2016.07.035

Muggia, L., Gueidan, C., Knudsen, K., Perlmutter, G., and Grube, M. (2013). The lichen connections of black fungi. *Mycopathologia* 175 (5), 523–535. doi:10.1007/s11046-012-9598-8

Mugnai, G., Rossi, F., Felde, V. J. M. N. L., Colesie, C., Büdel, B., Peth, S., et al. (2018). Development of the polysaccharidic matrix in biocrusts induced by a cyanobacterium inoculated in sand microcosms. *Biol. Fertil. Soils* 54 (1), 27–40. doi:10.1007/s00374-017-1234-9

Munzi, S., Varela, Z., and Paoli, L. (2019). Is the length of the drying period critical for photosynthesis reactivation in lichen and moss components of biological soil crusts? *J. Arid Environ.* 166, 86–90. doi:10.1016/j.jaridenv.2019.04.019

Napoli, A., Micheletti, D., Pindo, M., Larger, S., Cestaro, A., de Vera, J. P., et al. (2022). Absence of increased genomic variants in the cyanobacterium Chroococcidiopsis exposed to Mars-like conditions outside the space station. *Sci. Rep.* 12 (1), 8437–8439. doi:10.1038/s41598-022-12631-5

Navarro-González, R., Rainey, F. A., Molina, P., Bagaley, D. R., Hollen, B. J., de la Rosa, J., et al. (2003). Mars-like soils in the Atacama Desert, Chile, and the dry limit of microbial life. *Science* 302 (5647), 1018–1021. doi:10.1126/science.1089143

Neilan, B. A., Burns, B. P., Relman, D. A., and Lowe, D. R. (2002). Molecular identification of cyanobacteria associated with stromatolites from distinct geographical locations. *Astrobiology* 2 (3), 271–280. doi:10.1089/153110702762027853

Niederwieser, T., Kociolek, P., and Klaus, D. (2018). Spacecraft cabin environment effects on the growth and behavior of Chlorella vulgaris for life support applications. *Life Sci. space Res.* 16, 8–17. doi:10.1016/j.lssr.2017.10.002

Onofri, S., Selbmann, L., De Hoog, G. S., Grube, M., Barreca, D., Ruisi, S., et al. (2007). Evolution and adaptation of fungi at boundaries of life. *Adv. Space Res.* 40 (11), 1657–1664. doi:10.1016/j.asr.2007.06.004

Reed, S. C., Coe, K. K., Sparks, J. P., Housman, D. C., Zelikova, T. J., and Belnap, J. (2012). Changes to dryland rainfall result in rapid moss mortality and altered soil fertility. *Nat. Clim. Chang.* 2 (10), 752–755. doi:10.1038/nclimate1596

Rikkinen, J. (2015). *Cyanolichens. Biodivers. Conservation* 24 (4), 973–993. doi:10.1007/s10531-015-0906-8

Rodriguez-Caballero, E., Belnap, J., Büdel, B., Crutzen, P. J., Andreae, M. O., Pöschl, U., et al. (2018). Dryland photoautotrophic soil surface communities endangered by global change. *Nat. Geosci.* 11 (3), 185–189. doi:10.1038/s41561-018-0072-1

Rondanelli, R., Molina, A., and Falvey, M. (2015). The Atacama surface solar maximum. *Bull. Am. Meteorological Soc.* 96 (3), 405–418. doi:10.1175/bams-d-13-00175.1

Sánchez-Baracaldo, P., and Cardona, T. (2020). On the origin of oxygenic photosynthesis and Cyanobacteria. *New Phytol.* 225 (4), 1440–1446. doi:10.1111/nph.16249

Schmidt, S. K., Gendron, E. M. S., Vincent, K., Solon, A. J., Sommers, P., Schubert, Z. R., et al. (2018). Life at extreme elevations on Atacama volcanoes: The closest thing to Mars on Earth? *Antonie Leeuwenhoek* 111 (8), 1389–1401. doi:10.1007/s10482-018-1066-0

Schulze-Makuch, D., Wagner, D., Kounaves, S. P., Mangelsdorf, K., Devine, K. G., de Vera, J. P., et al. (2018). Transitory microbial habitat in the hyperarid Atacama Desert. *Proc. Natl. Acad. Sci. U. S. A.* 115 (11), 2670–2675. doi:10.1073/pnas.1714341115

Schweiger, A. H., Ullmann, G. M., Nürk, N. M., Triebel, D., Schobert, R., and Rambold, G. (2022). Chemical properties of key metabolites determine the global distribution of lichens. *Ecol. Lett.* 25 (2), 416–426. doi:10.1111/ele.13930

Shih, P. M. (2015). Photosynthesis and early Earth. *Curr. Biol.* 25 (19), R855–R859. doi:10.1016/j.cub.2015.04.046

Sobron, P., Lefebvre, C., Leveille, R., Koujelev, A., Haltigin, T., Du, H., et al. (20122013). Geochemical profile of a layered outcrop in the Atacama analogue using laser-induced breakdown spectroscopy: Implications for Curiosity investigations in Gale. *Geophys. Res. Lett.* 40 (10), 1965–1970. doi:10.1002/grl.50261

Sterflinger, K. (1998). Temperature and NaCl-tolerance of rock-inhabiting meristematic fungi. *Antonie Leeuwenhoek* 74 (4), 271–281. doi:10.1023/a:1001753131034

Stern, J. C., Sutter, B., Freissinet, C., Navarro-Gonza´lez, R., McKay, C. P., Archer, P. D., et al. (2015). Evidence for indigenous nitrogen in sedimentary and aeolian deposits from the Curiosity rover investigations at Gale crater, Mars. *Proc. Natl. Acad. Sci. U. S. A.* 112 (14), 4245–4250. Epub 20150323. doi:10.1073/pnas.1420932112

Sundarasami, A., Sridhar, A., and Mani, K. (2019). Halophilic archaea as beacon for exobiology: Recent advances and future challenges. *Adv. Biol. Sci. Res.*, 197–214.

Toigo, A. D., and Richardson, M. I. (2000). Seasonal variation of aerosols in the Martian atmosphere. *J. Geophys. Res.* 105, 4109–4121. doi:10.1029/1999je001132

Treves, H., Raanan, H., Finkel, O. M., Berkowicz, S. M., Keren, N., Shotland, Y., et al. (2013). A newly isolated Chlorella sp. from desert sand crusts exhibits a unique resistance to excess light intensity. *FEMS Microbiol. Ecol.* 86 (3), 373–380. doi:10.1111/1574-6941.12162

Verseux, C., Baqué, M., Cifariello, R., Fagliarone, C., Raguse, M., Moeller, R., et al. (2017). Evaluation of the resistance of Chroococcidiopsis spp. to sparsely and densely ionizing irradiation. *Astrobiology* 17, 118–125. doi:10.1089/ast.2015.1450

Verseux, C., Baque´, M., Lehto, K., de Vera, J-P. P., Rothschild, L. J., and Billi, D. (2016). Sustainable life support on Mars—The potential roles of cyanobacteria. *Int. J. Astrobiol.* 15, 65–92. doi:10.1017/s147355041500021x

Viúdez-Moreiras, D., Gómez-Elvira, J., Newman, C. E., Navarro, S., Marin, M., Torres, J., et al. (2019). Gale surface wind characterization based on the Mars Science Laboratory REMS dataset. Part II: Wind probability distributions. *Icarus* 319, 645–656. doi:10.1016/j.icarus.2018.10.010

Wamelink, G. W., Frissel, J. Y., Krijnen, W. H., Verwoert, M. R., and Goedhart, P. W. (2014). Can plants grow on Mars and the moon: A growth experiment on Mars and moon soil simulants. *PLoS One* 9 (8), e103138. Epub 2014/08/27PMID: 25162657; PubMed Central PMCID:PMC4146463. doi:10.1371/journal.pone.0103138

Wang, B., Ye, T., Li, X., Bian, P., Liu, Y., and Wang, G. (2021). Survival of desert algae Chlorella exposed to Mars-like near space environment. *Life Sci. Space Res.* 29, 22–29. doi:10.1016/j.lssr.2021.02.003

Weber, B., Belnap, J., Büdel, B., Antoninka, A. J., Barger, N. N., Chaudhary, V. B., et al. (2022). What is a biocrust? A refined, contemporary definition for a broadening research community. *Biol. Rev.* 97, 1768–1785. doi:10.1111/brv.12862

White, B. R. (1979). Soil transport by winds on Mars. *J. Geophys. Res.* 84, 4643–4651. doi:10.1029/jb084ib09p04643

Xin, Z. H. A. O., Yong, S. H. I., Yang, L. I. U., Rong-Liang, J. I. A., and Xin-Rong, L. I. (2015). Osmotic adjustment of soil biocrust mosses in response to desiccation stress. *Pedosphere* 25 (3), 459–467. doi:10.1016/s1002-0160(15)30013-8

Young, K. E., Ferrenberg, S., Reibold, R., Reed, S. C., Swenson, T., Northen, T., et al. (2022). Vertical movement of soluble carbon and nutrients from biocrusts to subsurface mineral soils. *Geoderma* 405, 115495. doi:10.1016/j.geoderma.2021.115495

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 15 December 2022
DOI 10.3389/fspas.2022.872283

Check for updates

OPEN ACCESS

EDITED BY
Emanuele Contini,
Yonsei University, South Korea

REVIEWED BY
Vladimir Avila-Reese,
Institute of Astronomy, National Autonomous University of Mexico, Mexico
Daniela Bettoni,
Osservatorio Astronomico di Padova (INAF), Italy

*CORRESPONDENCE
Magda Arnaboldi,
marnabol@eso.org

SPECIALTY SECTION
This article was submitted to Extragalactic Astronomy, a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 09 February 2022
ACCEPTED 22 November 2022
PUBLISHED 15 December 2022

CITATION
Arnaboldi M and Gerhard O (2022), Kinematics of the diffuse intragroup and intracluster light in groups and clusters of galaxies in the local universe within 100 Mpc distance.
Front. Astron. Space Sci. 9:872283.
doi: 10.3389/fspas.2022.872283

COPYRIGHT
© 2022 Arnaboldi and Gerhard. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Kinematics of the diffuse intragroup and intracluster light in groups and clusters of galaxies in the local universe within 100 Mpc distance

Magda Arnaboldi [1]* and Ortwin Gerhard [2]

[1]European Southern Observatory (ESO), Garching, Germany, [2]Max-Planck-Institute für extraterrestrische Physik, Garching, Germany

Nearly all intragroup (IGL) and intracluster light (ICL) comes from stars that are not bound to any single galaxy but were formed in galaxies and later unbound from them. In this review we focus on the physical properties–phase space properties, metallicity and age distribution–of the ICL and IGL components of the groups and clusters in the local universe, within 100 Mpc distance. Kinematic–information on these very low surface brightness structures mostly comes from discrete tracers such as planetary nebulae and globular clusters, showing highly unrelaxed velocity distributions. Cosmological hydrodynamical simulations provide key predictions for the dynamical state of IGL and ICL and find that most IC stars are dissolved from galaxies that subsequently merge with the central galaxy. The increase of the measured velocity dispersion with radius in the outer halos of bright galaxies is a physical feature that makes it possible to identify IGL and ICL components. In the local groups and clusters, IGL and ICL are located in the dense regions of these structures. Their light fractions relative to the total luminosity of the satellite galaxies in a given group or cluster are between a few to ten percent, significantly lower than the average values in more evolved, more distant clusters. IGL and ICL in the Leo I and M49 groups, and the Virgo cluster core around M87, has been found to arise from mostly old (≥10 Gyr) metal-poor ([Fe/H] <-1.0) stars of low-mass progenitor galaxies. New imaging facilities such as LSST, Euclid, and the "big eyes" on the sky–ELT and JWST with their advanced instrumentation–promise to greatly increase our knowledge of the progenitors of the IGL and ICL stars, their ages, metal content, masses and evolution, there by increasing our understanding of this enigmatic component.

KEYWORDS

galaxies, halos, evolution, clusters, planetary nebulae, general

1 Introduction

The motivation to learn about the physics of extended stellar halos in galaxies and the diffuse light in larger structures is synthesized by the statement of Eggen, Lynden-Bell and Sandage in their 1962 seminal paper (Eggen et al., 1962), that is "the time required for stars in the galactic system to exchange their energies and momenta is very long compared with the age of the galaxy. Hence the knowledge of the present energies and momenta of individual objects tells us something of the initial dynamic conditions under which they were formed". This statement from 60 years ago is still very relevant today: we study the dynamics and chemical composition of stars in the diffuse outer halos of galaxies and in the intracluster and intragroup light in clusters an groups, in order to understand the build-up and assembly of these extended structures.

The diffuse stellar component (DSC) in local groups and clusters is an extended, low surface brightness distribution of stars with approximate elliptical symmetry around the bright galaxies at the center of these groups and clusters. The DSC can harbor multiple components from galaxy outer haloes to intracluster light, which can be characterized by their radial profiles, and shapes of the isophotes. Thanks to wide field imaging facilities these DSCs have been measured to extend out to hundreds of kiloparsecs. In addition to the smooth components with elliptical symmetry, maze-like webs of tails, plumes and spurs can be present. Exemplary testimony of such intricate luminous substructures is provided by the extremely deep images, reaching surface brightness levels below 1% of the night sky: see for example those obtained for the Virgo (Mihos et al., 2005; Mihos et al., 2013; Spavone et al., 2017), Fornax (Iodice et al., 2016) and the Hydra-I (Arnaboldi et al., 2012; Barbosa et al., 2018; La Marca et al., 2022a; La Marca et al., 2022b) clusters.

Attempts to quantify the kinematics of the DSC within the groups and clusters date back to the work of Alan Dressler in 1979, on the dynamics and structure of the cD galaxy IC1101 in the cluster Abell 2029 (Dressler, 1979). In his work, Dressler measured absorption lines on the continuum from the stars in spectra obtained for the extended halo around IC1101; the relatively ``bright'' surface brightness of this halo allowed the measurement of the broadening function of the MgII and Fe absorption features around 5,200 Å. The width of these absorption lines probes the line-of sight (LOS) velocity dispersion of the stars orbiting in this halo. Dressler found that the measured velocity dispersion increased with central distance from IC1101 implying that the mass-to-light (M/L) ratio of the envelope was increasing rapidly. His assessment of the physical implications of these measurements was that he identified the outermost high M/L component with the stripped stars from satellite galaxies orbiting in the cluster around IC1101. Today we associate such diffuse light around central bright galaxies with the intracluster light (ICL), defined as the stellar component at the centres of groups and clusters which is not bound to individual galaxies.

Kinematic measurements of the stellar motions in extended structures such as the outer halos of galaxies, the ICL and the intragroup light (IGL) are challenging. Absorption line spectroscopy in early-type galaxies (ETGs) is often confined to within two effective radii (R_e, defined as the radius which contains half of the total light in a galaxy). Despite the observational challenges, kinematic studies of ETG halos from integrated-light absorption spectra were undertaken by Kelson et al. (2002), Weijmans et al. (2009), Coccato et al. (2010), Murphy et al. (2011), Bender et al. (2015), and Barbosa et al. (2018) using either long slit spectroscopy or integral field spectroscopy (IFS) out to large radii, and integrating for several hours on large telescopes. More recently the SLUGGS survey (Arnold et al., 2014; Foster et al., 2016; Bellstedt et al., 2017), the MASSIVE survey (Raskutti et al., 2014; Veale et al., 2017), Boardman et al. (2017) and Loubser et al. (2022) generated kinematic data from IFSs for larger samples of ETGs. However, the extended kinematics in these studies do not reach distances beyond 3–4R_e.

The observational limitations hamper the assessment of the complex dynamics of dynamically "hot" stellar systems, dominated by velocity dispersion. "Complex" here means that the inference of the total enclosed mass from the observed velocities is not straightforward as it is for circular orbits. In hot stellar systems, the total enclosed mass can be constrained using the higher moments of the line of sight velocity distribution (LOSVD). Measuring these higher order moments in galaxy halos may resolve the anisotropy-potential-degeneracy (e.g., Gerhard 1993; Rix et al., 1997; Thomas et al., 2005; Napolitano et al., 2009; Morganti et al., 2013), but beyond 3–4 R_e this is not possible with absorption line spectroscopy.

The way forward to map the kinematics of the DSC to large radii is the spectroscopic follow-up of relatively luminous discrete tracers which overcome the drop in surface brightness. Such discrete tracers are globular clusters (GCs, Schuberth et al., 2010; Strader et al., 2011; Pota et al., 2018 and reference therein) or planetary nebulae. Planetary nebulae (PNe) are established kinematics probes of the stellar populations in extended galaxy halos and ICL (e.g., Arnaboldi et al., 1998; Arnaboldi, 2005; Longobardi et al., 2013; Arnaboldi et al., 2017; Hartke et al., 2017; Pulsoni et al., 2018). Their relatively bright [OIII] 5,007 Å line emission stands out against the galaxy faint continuum background, making them relatively easy to detect (Arnaboldi et al., 2002). Since PNe represent the final phase of the Asymptotic Giant Branch (AGB) for low to intermediate mass stars, in old stellar populations such as found in ETGs they trace the bulk of the stars that emit the host-galaxy continuum light. In the overlapping regions, the kinematics of PNe is directly comparable to integrated light measurements (Hui et al., 1995; Arnaboldi et al., 1996; Arnaboldi et al., 1998; Méndez et al., 2001; Coccato et al., 2009; Cortesi et al., 2013; Pulsoni et al., 2018).

From the projected two-dimensional kinematics maps derived with discrete tracers like PNe and GCs it becomes possible to map substructures both in space and velocity (Napolitano et al., 2003; Longobardi et al., 2015b; Hartke et al., 2018; Napolitano et al., 2022). In the last few years evidence was collected for debris of disrupted satellites in halos, groups and clusters, providing observational evidence that these components are still building up (Ventimiglia et al., 2011; Longobardi et al., 2015b; Hartke et al., 2018). To understand the growth of these structures, observations must be compared with models and simulations. This is because we deal with a long sequence of events, where accretion and mergers play important roles (see also Arnaboldi and Gerhard 2010; Pulsoni et al., 2021), leading to the formation of the IGL and ICL that we see today in the nearby universe. Since the first spectroscopic identification of intracluster planetary nebulae (ICPNe) in the Virgo cluster (Arnaboldi et al., 1996), imaging surveys and extended spectroscopic follow-up campaigns have identified these kinematic probes and used them to trace the radial extent and the kinematics of the DSC in groups and clusters in the nearby universe. PNe as discrete tracers also play an important role for galaxy halos, in order to constrain the integrated mass as function of radius, the intrinsic three-dimensional shape of the mass distribution, luminous and dark, and the orbital distribution of halo stars in the nearby universe. See for example the recent comparisons between the observed physical properties of ETGs outer halos (Pulsoni et al., 2018) with their analogs selected from the Illustris TNG simulations (e.g., Pulsoni et al., 2020; Pulsoni et al., 2021).

In this article we describe the IGL and ICL properties in the local groups[1] and clusters, as mapped by the kinematics and spatial distribution of their stars. [2]We begin with a concise overview of the general properties of the IGL and ICL from deep imaging and quantitative photometry, to then move on to utilize them together with the kinematics measured *via* absorption lines on the stellar continuum and/or PN spectroscopy. Important guidance comes from numerical simulations and semi-analytic models, which are used as abenchmark to identify the most robust observables in the group and cluster, but also as a tool to interpret the results of the observations. We then explore ways to investigate observationally the stellar populations of the IGL and ICL, to constrain their progenitors, and compare the measured kinematics of distinct discrete tracers, such as PNe and GCs. We then draw our conclusions and illustrate avenues for future investigations.

2 IGL and ICL properties in local groups and clusters: Their light distribution

The properties of IGL and ICL in groups and clusters measured from deep imaging and surface brightness photometry are the main topics of several other contributions in this journal. In this review, we attempt to collect the basic facts which set the stage for an overview of the IGL and ICL kinematics in local groups and clusters. The presence of ICL is historically assessed as an excess of light with respect to the de Vaucouleurs $R^{1/4}$ profile (de Vaucouleurs 1948), extrapolated from the galaxy at the cluster center to large cluster radii (Melnick et al., 1977; Thuan and Kormendy 1977; Schombert 1986; Bernstein et al., 1995). The central group or cluster galaxy (CCG) is often but not always the brightest cluster galaxy (BCG) and often but not always a cD galaxy, i.e., a giant elliptical with cluster-size extra light at large radii with respect to the outward extrapolation of a Sérsic, (1963) function fitted to the inner profile (Kormendy et al., 2009).

Because of this light excess, either one or more components modeled as Sersic or exponential profiles are commonly used to reproduce the surface brightness profile over the entire range of radii, from a few to hundreds of kiloparsecs (Gonzalez et al., 2005; Spavone et al., 2017). These surface brightness profiles have a range of radial gradients, from steeper to flatter slopes across different clusters (Krick and Berstein 2007; Kluge et al., 2020). The use of several components with different analytical formulae to reproduce the surface brightness profile from the inner to the outer regions is further motivated by a frequent change of the isophote shapes with radius. The ellipticity generally increases with radius, i.e., isophotes become flatter, and the position angle profile of the major axis of the outer isophotes has sharp variations (Gonzalez et al., 2005; Kluge et al., 2021) as function of radius outwards from the CCG to the group and cluster regions.

This behavior signals the transition from the CCG halo to the IGL and ICL. In general, the ICL is more aligned with the host cluster than the CCG in terms of position angle, ellipticity, and centering (Zibetti et al., 2005; Kluge et al., 2021). The quantitative 2D morphological variations (flattening of the isophotes, twist of the positional angle of the major axis with radius) and photometric properties (different radial gradients in different clusters) suggest that the IGL and ICL are dynamically young and separate from the CCG halo.

When mapping the DSC in groups and clusters on the basis of photometry, the question arises whether the separation of the IGL and ICL from the outer halo of the CCG has a physical motivation or whether it is only a taxonomy exercise, focused

1 The closest compact group is at D > 100 Mpc; for a summary of the properties of IGL in compact groups see Poliakov et al. (2021).

2 We do not discuss the Intra Cluster Medium properties, either hot gas or dust.

primarily on assessing changes of the radial gradients or isophotal twists of the surface brightness distribution. The latter approach which unifies both ICL and the halo of the CCG as DSC can be summarized by the statement of Uson et al. (1991): "*Whether this diffuse light is called the cD envelope or diffuse intergalactic light is a matter of* semantics; it is a diffuse component which is distributed with elliptical symmetry about the center of *the cluster potential*". This sets the stage for a statistical study of the CCG and ICL by stacking deep images of clusters after suitable rescaling, such as utilized by Zibetti et al. (2005), see also D'Souza et al. (2014). Zibetti et al. (2005) studied the surface brightness profiles of the DSC components obtained after stacking images of 683 clusters in the redshift range 0.2–0.3, from the first data release of the Sloan Digital Sky Survey. They were able to detect ICL out to 700 kpc from the CCGs as excess with respect to an inner $R^{1/4}$ profile; they determined an ICL faction of 10.9% ± 5%, which seems to be independent of cluster richness and CCG luminosity. While this "global" approach provides the average distribution of the DSC, it erases any differences of shapes and position angles as function of radius.

Investigations of the properties of the DSC in clusters from cosmological simulations can be carried out with a very similar methodology, see discussions in Cooper et al. (2013), Cooper et al. (2015) for particle tagging N-body simulations and in Pillepich et al. (2018), Cañas et al. (2020) for fully hydrodynamical cosmological simulations. In simulations, the stellar particles which remain in a group or cluster once all the sub-halos or satellite galaxies have been removed using binding energy constraints, can be readily interpreted as the close analog of the observed DSC in these structures, including the CCG. From the studies based on the Illustris TNG simulations, Pillepich et al. (2018) obtained that the radial profiles of the DSC at large radii become shallower in the more massive clusters.

A complementary approach to that described by Uson et al. (1991) and in the previous paragraphs seeks to associate major changes in the spatial distribution and surface brightness (radial gradient, ellipticity or position angle profiles) of the DSC in groups and clusters with the transition to a different physical component, e.g., the IGL or ICL. This strategy is similar to that carried out for the structural components in galaxies (bulge *versus* disk decomposition, see Kormendy and Bender 2012).

The fraction of the IGL and ICL in individual groups and clusters is then derived from quantitative photometry. The estimated values depend also on the methodology used to quantify them, i.e., whether the ICL is measured as the light above a fixed surface brightness threshold, or computed as the light excess with respect to the extrapolation of the $R^{1/4}$ law, or as fraction of the light associated with a second, outer Sersic component (see extensive comparisons and discussions in Kluge et al., 2021; Montes 2022). It also depends on the radial range over which this component is measured. Typically measuring light above a certain threshold provides the largest values of the IGL and ICL fractions with respect to the total galaxy light, which is determined by summing up all the light in the CCG and satellite galaxies. Fractions obtained from the different methodologies vary from few percent in the low mass groups (4% in the Leo I group, Watkins et al., 2014; Hartke et al., 2020; 10% in the Virgo subcluster B, see Hartke et al., 2018) and unrelaxed clusters (e.g., 5% in Virgo, see Castro-Rodriguéz et al., 2009), up to an average 48–52% in the more relaxed, dynamically evolved clusters (computed as excess to the extrapolated $R^{1/4}$ profile or associated with the second Sersic component, Kluge et al., 2021). See also Cañas et al. (2020) and Contini and Gu (2020) for predictions on the IGL and ICL fractions in groups and clusters from cosmological simulations. These recent simulations reproduce the increase of the amount of light at 100 kpc with respect to the inner 10 kpc in groups and clusters as function of mass, measured by DeMaio et al. (2020).

An important comparison is that between the surface brightness profile measured for the ICL, and the surface brightness profile obtained from the averaged light of all group and cluster satellite galaxies. Where it has been measured, the ICL surface brightness is steeper than that of the averaged-in-annuli (group or cluster) satellite galaxies surface brightness profile, i.e., the ICL more spatially concentrated to the high density regions (Zibetti et al., 2005; Castro-Rodriguéz et al., 2009; Spiniello et al., 2018). For more exhaustive reviews of these results see the contributions in this volume focused on the quantitative photometric properties of the IGL and ICL.

3 DSC in groups and clusters: How to kinematically disentangle IGL and ICL from the outskirts of the CCG

Another approach to identify IGL and ICL is *via* measurements of the kinematic of stars. This is also motivated by the observed rise of the velocity dispersion in CCGs, which started with the work of Dressler, (1979) on IC 1101. This early work signaled a transition from stellar motions bound to the CCG to motions gravitationally bound to the group or cluster structure as a whole. It was then thefirst kinematic detections of IC PNe in the core of the Virgo cluster (Arnaboldi et al., 1996) and in Coma (Gerhard et al., 2005; Gerhard et al., 2007), which displayed large residual velocities with respect to the closest galaxies (>1,600 kms^{-1} relative to M86 in Virgo and ~700 kms^{-1} relative to NGC 4874 in Coma) that provided the opportunity to distinguish CCG halo and IGL or ICL at the same location. This early work was subsequently expanded to provide LOSVDs and velocity histograms of stellar tracers on larger areas: see Arnaboldi et al. (2004), Doherty et al. (2009) for the Virgo cluster and Ventimiglia et al. (2011) for the Hydra I cluster. When LOS velocities are measured to be significantly in excess of the velocity dispersion measured in the CCG, then the gravitational

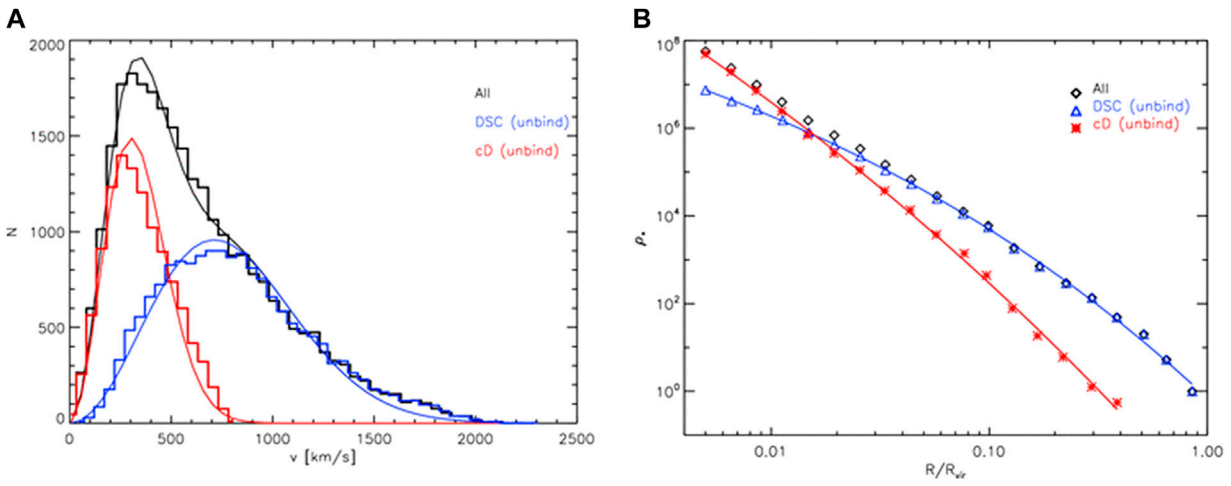


FIGURE 1
Maxwell-Boltzmann distributions of the absolute velocities of stellar particles in simulated clusters and separation of central galaxies and ICL. **(A)**–velocity histogram (black line) of the DSC in a simulated cluster and a double Maxwellian fit to it (grey line) are shown. Red and blue histograms show the velocity distribution of the CCG and ICL stars after the unbinding procedure, together with the two individual Maxwellian distributions from the global fit, respectively. **(B)**–All stars centered on CCG (black symbols), CCG (red symbols) and ICL (blue symbols) stellar density profiles. The red and blue line are Sersic fits to the two distributions, respectively. Image reproduced from Figure 7 and Figure 8 of Dolag et al. (2010) with permission.

potential of the entire group or cluster must provide the pull to bind these high velocity tracers, at the location where they are detected. The kinematically motivated approach to separate the CCG halo from ICL was adopted by Dolag et al. (2010) and Cui et al. (2014) for simulated clusters, using the distribution[3] of absolute velocities of stellar particles. In their cosmological cluster simulations, these authors found that the velocity distribution of the stellar particles associated with the CCG and DSC (after all the stellar particles in the satellite galaxies are removed) is bimodal in the volume occupied by a given cluster and can be fitted by two Maxwell-Boltzmann distributions, see Figure 1. The "narrower", i.e., colder, particle distribution occupies a more limited region of the phase space; these particles are associated with the central galaxy and the mass profile is steeper with radius, see Figure 1. The second ``broader", i.e., dynamically "hotter", particle distribution is more extended in the cluster volume, it has a shallower mass density profile and it is thus responsible for the "light excess" at large cluster-centric radii. The physical properties of these two components of the DSC, correspond to the CCG and ICL stars respectively. These results from simulations provide a consistent framework for the interpretation of the observed increase in the velocity dispersion profiles around IC 1101 (Dressler 1979), M87 (Murphy et al., 2011; Longobardi et al., 2015a), NGC 6166 (Bender et al., 2015), and NGC 3311 (Arnaboldi et al., 2012; Barbosa et al., 2018), whereby the LOSVD for the DSC in these clusters can be considered as the sum of the "colder" CCG halo and the ``hotter" ICL. We shall discuss these observational results more in detail in the next sections.

The physical criterium which is applied to kinematically separate the ICL from the CCG in these simulated clusters is that of the binding energy, i.e., whether the observed velocity component would satisfy the Virial theorem for the mass associated with the closest galaxy. By adopting the physical criteria based on binding energy, we can then identify the ICL stellar particles; whether this may always be possible as function of decreasing group and cluster masses, and dynamical status of the group or cluster, would require further analysis of cosmological simulations.

One can further explore the properties of the kinematically selected ICL particles in simulations. It turns out that the ICL stellar particles in hydro-dynamical simulations selected in the aforementioned way have systematically different star formation histories (Kapferer et al., 2010) and are on average older that the stars in cluster galaxies (Murante et al., 2004) in addition to their distinct phase space distribution. This is a very interesting perspective for the investigation of ICL in clusters. It provides a complementary, potentially more powerful approach to the statistical one described by Uson et al. (1991), which seeks to unify CCG halo and ICL in a unique DSC. The results from simulations stimulate the

3 Histogram of the modulus of velocities.

need for observational strategies to measure these differences and investigate how they were established during the galaxy assembly and cluster formation. Mapping differences in kinematics and stellar population properties thus provides the opportunity to tag CCG halo and ICL stars at the same spatial location. In Sections 4–6 we review highlights and examples where it has been possible to observationally distinguish CCG halo, IGL and ICL *via* kinematic measurements and to link the distinct kinematics to differences in surface brightness distributions and stellar populations, which in turn leads to constraining their origins in a quantitative way, in terms of progenitor masses and timing of the disruption events.

4 CCG and IGL or ICL as distinct kinematics components

The measurement of the surface brightness profiles and of detailed kinematics at large radii from the CCG center would reveal the presence of the accretion substructures that are not yet phase mixed, as well as diffuse ICL components. These measurements would allow one to isolate the smooth CCG halo, constrain its orbit distribution and obtain unbiased estimates of its stellar and enclosed dynamical mass. By smooth CCG halo here we refer to the part of the halo that is approximately phase-mixed, i.e., has approximately Gaussian LOSVD centered about the systemic velocity of the galaxy. This is clearly different from a multi-peaked LOSVD. The working concept of a smooth halo is useful because the approximately Gaussian LOSVD enables one to determine a well-defined halo velocity dispersion profile and combine it with the tracer density to carry out a Jeans analysis of the mass distribution at large radii. The use of these concepts, i.e., smooth CCG halo and less well mixed dynamically hotter IGL and ICL allows one to investigate the transition between the CCG halo and the IGL or ICL.

Since the serendipitous discovery of three ICPNe at $V_{LOS,mean}$ = 1,400 kms^{-1} along the LOS of the galaxy M86 (NGC 4406; v_{sys} = −240 kms^{-1}) in the Virgo cluster core (Arnaboldi et al., 1996), ICPNe have been used to map both spatial distributions and kinematics of the ICL and IGL component in nearby clusters. In what follows, the kinematic properties of the ICPN samples in the clusters within 100 Mpc distance, i.e., Virgo, Fornax, Hydra I and Coma, are outlined and compared, including efforts to physically separate the CCG halo from the cluster ICL or group IGL component.

4.1 Clusters in the nearby universe

4.1.1 Kinematics of ICL around M87 in the Virgo cluster (D = 15 Mpc)

Narrow band imaging surveys carried out to identify ICPNe were performed by Feldmeier et al. (1998), Feldmeier et al. (2004), Arnaboldi et al. (2002), Okamura et al. (2002), Aguerri et al. (2005), Castro-Rodriguéz et al., 2009, Longobardi et al. (2013). These were followed up by efforts to obtain spectra for ICPNe in the Virgo cluster (Arnaboldi et al., 2003; Arnaboldi et al., 2004; Doherty et al., 2009), which provided reliable measurements for 52 single PNe within one degree circular radius around M87 and in the Virgo core (as defined by Binggeli et al., 1987). The spectra for these single stars showed the [OIII] λ4959/5,007 Å double emissions, confirming the identification of these emission line sources as PNe and their physical association with the Virgo cluster. The sparsely sampled projected phase-space, built from the ICPN V_{LOS} *versus* radial distance from the centre of M87, displayed regions with different kinematics. The halo of M87 was clearly identified as a single, relatively narrow ridge centred around the galaxy systemic velocity (V_{sys},M87 ≈ 1,300 kms^{-1}). However, the data showed an additional component to the M87 halo, which was characterised by a broad and extended LOSVD out to distances of 2,000 arcsec, ≈150 kpc, and larger. Such a broad LOSVD is reminiscent of the velocity distribution of the satellite galaxies in the Virgo cluster (Binggeli et al., 1987).

An important step in mapping the IC component around M87 came from the more recent, extended, ICPN imaging and spectroscopic survey carried out by Longobardi et al. (2013), Longobardi et al. (2015a). This work led to a major increase of a factor 15 in the number of measured v_{LOS} from independent spectra, with respect with previous ICPNe samples in the centre of the Virgo subcluster A (Binggeli et al., 1987). The ICPN survey carried out with Subaru SuprimeCam (imaging) and Flames at UT2 on the VLT (spectroscopy) delivered 287 spectra for individual PNe (Longobardi et al., 2015a). This sample was analysed including additional 11 v_{LOS} from independent PN spectra measured by Doherty et al. (2009) in the same area, giving a combined total sample of 298 PN v_{LOS}. Interestingly, the entire PN LOSVD displays broad asymmetric wings that cannot be matched by a single Gaussian, see Figure 2. Instead it can be approximated by two Gaussians, a narrow component centred at the systemic velocity of M87, with σ_{narrow} ≈300 kms^{-1}, and an additional, much broader, Gaussian with an average velocity at v_{broad} ≈1,000 kms^{-1} and σ_{broad} ≈900 kms^{-1}. Each PN in the projected phase-space, V_{LOS} vs. R_{major}, around M87 can be assigned a probability to belong either to the narrow Gaussian, i.e. the M87 halo component, or to the broad component, i.e.,the ICL. The ICPNe in the Virgo cluster core were thus identified by Longobardi et al. (2015a) *via* a multi-Gaussian decomposition of their LOSVD. The ICPN sample was later expanded by Longobardi et al. (2018a), who added the unrelaxed component that overlaps with the M87 smooth halo in the 40–150 kpc radial range. Using the combined ICPN sample, i.e., ICL from the broad Gaussian (σ_{broad} ≈ 900 kms^{-1}) in the Virgo core plus the unrelaxed components at the location of the M87 halo (including the crown, Longobardi et al., 2015b), the resulting ICPN LOSVD is statistically undistinguishable from the LOSVD for satellite galaxies (Longobardi et al., 2018a) in the

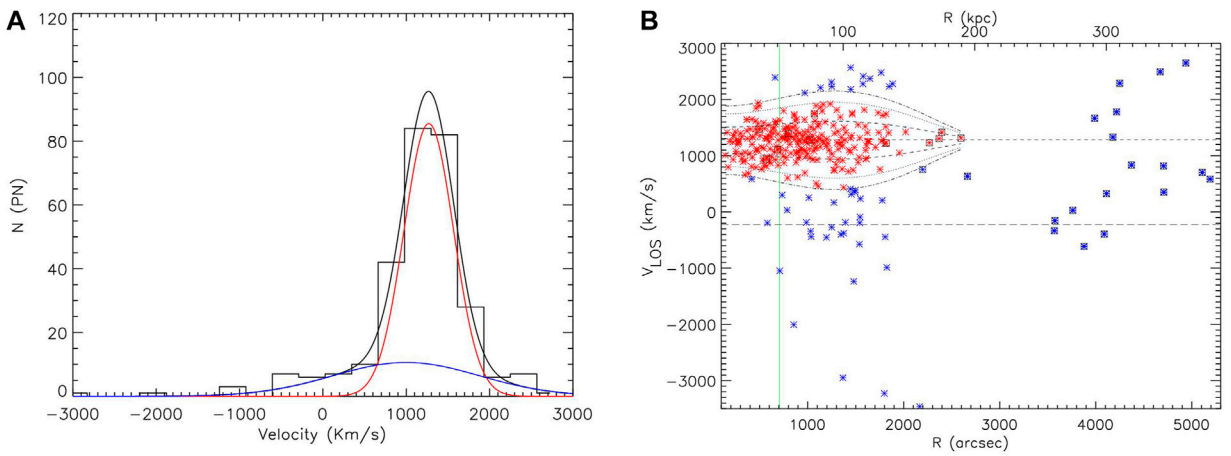


FIGURE 2
PN velocities tracing the diffuse light around M87 in the Virgo cluster. **(A)**–Histogram of the V_{LOS} of the spectroscopically confirmed PNe (black histogram) fitted with a double Gaussian (black curve). Red and blue lines represent the two Gaussians associated with the M 87 halo and the IC components. **(B)**–Projected phase-space diagram, V_{LOS} vs. major axis distance from the centre of M 87, for all spectroscopically confirmed PNe in Longobardi et al. (2015a). The major axis distance is given both in arcsec (bottom axis), and in kpc (top axis), where 73 pc = 1″. The PNe are classified as M 87 halo PNe (red asterisks) and ICPNe (blue asterisks), respectively; see text. Black squares identify spectroscopically confirmed PNe from Doherty et al. (2009). The smoothed 1, 2, and 2.5σ thresholds are represented by the dashed, dotted, and dot-dashed lines, respectively. The dashed horizontal line shows the M 87 systemic velocity Vsys = 1,275 kms^{-1}, while the continuous green vertical line shows the effective radius Re = 704″ determined by Kormendy et al. (2009). At VLOS = −220 kms^{-1}, we plot the M 86 systemic velocity (long dashed line). Image reproduced from Longobardi et al. (2015a) with permission.

Virgo core region (Binggeli et al., 1993). This approach is the implementation in the observable parameter space of the CCG halo/ICL decomposition of the DSC in simulated clusters, as illustrated by Dolag et al. (2010). For the GCs populations in the Virgo core, the separation between a GCs population bound to M87 and a GCs population orbiting in the IC space was determined independently by Longobardi et al. (2018b).

4.1.2 Kinematics of the ICL around NGC 1399 in the Fornax cluster (D = 18 Mpc)

The extended halos of NGC 1399, NGC 1316 and the ICL in the Fornax cluster have been targets of several extended imaging and spectroscopic surveys to measures the V_{LOS} of discrete tracers including PNe (Arnaboldi et al., 1994; McNeil et al., 2010; McNeil-Moylan et al., 2012; Spiniello et al., 2018) and GCs (Schuberth et al., 2010; Pota et al., 2018; Chaturvedi et al., 2022). For the earlier imaging survey to identify ICPNe in Fornax see also Theuns and Warren (1997). The outer envelope of NGC 1399 is characterized photometrically by an exponential profile (Iodice et al., 2016 and reference therein) which dominates the DSC in Fornax out to 200 kpc. Once the different tracers are merged, one wishes to probe the kinematics from the center, using absorption line spectra, then out to 200 kpc using PNe, red and blue GCs. The combined velocity dispersion profile shows a central velocity dispersion value of ~ 380 kms^{-1}, which then decreases rapidly to 250–200 kms^{-1} between 10–20 kpc radius.

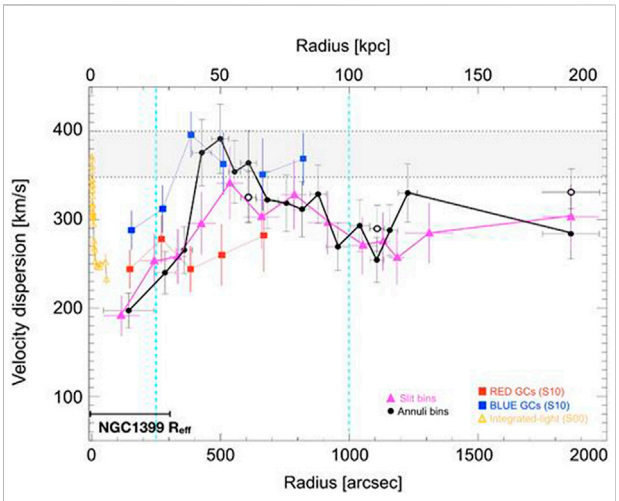


FIGURE 3
Velocity dispersion profile of diffuse light in the Fornax cluster, from the center of NGC 1399 out to 200 kpc. The σ_{LOS} (R) from PNe is plotted as filled black circles. The three empty symbols show values at radii where PNe most likely associated with satellite galaxies were removed. σ_{LOS} (R) from red and blue GCs (see Schuberth et al., 2010; S10 in the figure) are shown as red and blue squares, respectively. The inner velocity dispersion profile from absorption line spectra (Saglia et al., 2000; S00 in the figure) is shown as yellow triangles. The grey shaded horizontal region represents the σ_{LOS}, gal of the Fornax cluster with its 1σ error. Image reproduced from 8 of Spiniello et al. (2018) with permission.

Beyond those radii, σ_{LOS} increases again reaching 350 kms^{-1} at 50 kpc and then decreases to 300–250 kms^{-1} at larger radii. The σ_{LOS} values of the DSC in the Fornax cluster at ~ 200 kpc from NGC 1399 are lower than the velocity dispersion measured from the V_{LOS} of the satellite galaxies at those radii. Figure 3 shows a complete mapping of the velocity dispersion profile in the Fornax cluster from the different tracers and satellite galaxies.

The sample of PNe (Spiniello et al., 2018) and GCs (Pota et al., 2018) was used by Napolitano et al. (2022) to identify substructures in the Fornax cluster projected phase space, for which they provide a catalogue of 13 stellar stream candidates, each with their mean centroid position, associated number of PN and GC tracers, radial velocity, size, total luminosity and surface brightness. Most of these streams are kinematically connected to ultra-compact dwarfs (UCDs), which supports a scenario of their morphological transformation from the dwarf progenitors, after being disrupted in the cluster gravitational field. Some of the additional long coherent substructures connecting cluster members and some isolated clumps of tracers not associated with UCDs may represent fossil records of satellites that have since merged with Fornax cluster member galaxies.

4.1.3 Kinematics of ICL around NGC 3311 in the Hydra I cluster (D = 50 Mpc)

With an optimized observational technique named Multi Slit Imaging Spectroscopy (MSIS: Gerhard et al., 2005; Arnaboldi et al., 2007) with FORS2 at the ESO VLT, Ventimiglia et al. (2011) detected a total of 56 PNe in a single field of 100 × 100 kpc^2, centered on the galaxy NGC 3311, in the core of the Hydra I cluster. The PN LOSVD in this region has several peaks. In addition to a broad symmetric component centered at the systemic velocity of the cluster (V_{Hydra} = 3,683 kms^{-1}; σ_{Hydra} = 784 kms^{-1}), two narrow peaks are detected at 1,800 kms^{-1} and 5,000 kms^{-1}. We address the origin of the broad component and the secondary peaks in turn. In a series of papers, Barbosa et al. (2016), Barbosa et al. (2018) investigated the velocity dispersion profile of the galaxy NGC 3311 along its photometric major axis, and carried out a physical modelling of both the photometry and projected kinematics on the sky from deep IFS observations with MUSE at the ESO VLT. They found that the light in this galaxy can be modelled as a superposition of several non-concentric spheroidal components with increasing scale, slightly different systemic velocities and velocity dispersions. Interestingly, the outermost envelope is offset with respect to the central region by up to 50″ to the North. They found that the central high surface brightness components have a relative motion and displacement from the outer envelope, whose velocity dispersion rises to cluster-core values of ~ 400–500 kms^{-1} already at R > 20”. This outer envelope has an exponential surface brightness profile and is dynamically associated with the cluster (Arnaboldi et al., 2012). The transition to cluster-like velocity dispersion around NGC 3311 is confirmed independently by the LOS velocities of GCs (Hilker et al., 2018). The entire inner region is further displaced from the Hydra I cluster’s center, as defined by the spatial distribution of the outer satellite galaxies, presumably due to a recent sub-cluster merger (Barbosa et al., 2018).

The presence of secondary peaks in the LOSVD of the PNe indicates the existence of un-mixed populations in the Hydra I cluster core. The spatial distribution of the PNe associated with the narrow velocity sub-component at 5,000 kms^{-1} is spatially located on an excess of light in the North-East quadrant of NGC 3311, in the B band surface photometry (Arnaboldi et al., 2012). At the same sky area, a group of dwarf galaxies with $V_{LOS} \approx$ 5,000 kms^{-1} is detected. Deep long-slit spectra at the position of the dwarf galaxy HCC 026 at the center of this light excess show absorption features whose observed wavelengths are consistent with a V_{LOS} ~ 5,000 kms^{-1}. Arnaboldi et al. (2012) concluded that the PNe associated with the LOSVD peak at ≈5,000 kms^{-1}, spatially overlapping with the dwarfs and the light excess in the North-East quadrant of NGC 3311 are physically associated. Given the large residual velocity (~1,400 kms^{-1}) from the systemic velocity of NGC 3311, they are part of the ICL. About the secondary velocity peak at ~1,800 kms^{-1}: within a 20 arcmin distance from NGC 3311 there are eight galaxies with systemic velocities lower than 2,800 kms^{-1} and R band total magnitude < 15.4. Two of these galaxies are spirals, with NGC 3312 being at the boundary of the MSIS regions. These ICPNe may therefore be associated with the infalling group associated with NGC 3312. Hence also in the case of the relaxed Hydra-I cluster, PN kinematics, photometry and deep absorption spectra support the evidence for accretion or infalling of groups, including a sub-cluster merger, with a significant fraction of stars being added to the ICL around NGC 3311.

4.1.4 Kinematics of ICL in the core of the Coma cluster (D = 100 Mpc)

The Coma cluster is the richest and densest of the nearby clusters, yet there is evidence that sub-cluster mergers are taking place. With the MSIS technique using the FOCAS spectrograph at the Subaru telescope, Gerhard et al. (2005) detected the [OIII] 5,007 Å emission for ICPNe and the V_{LOS} for 37 of them were measured (Arnaboldi et al., 2007). These ICPNe are associated with the diffuse light previously detected by Bernstein et al. (1995) near the sky position α(J2000) 12:59: 41.784; δ(J2000) 27:53:25.388. These ICPNe in Coma are the most distant single stars whose spectra can be measured with 8-m class telescopes, apart from cosmological supernovae. Gerhard et al. (2007) detected several peaks in the LOSVD of the ICPNe within a 6′ diameter field, ~5′ away from the cD galaxy NGC 4874. A velocity sub-structure is present at ~ 5,000 kms^{-1}, probably from in-fall of a galaxy group (Adami et al., 2005), while the main IC stellar component has a V_{mean} centered around ~6,500 kms^{-1}, about $\Delta v \approx -700$ kms^{-1} offset from the nearest cD galaxy, NGC4874 (v_{sys} = 7,224 kms^{-1}). The ICPN V_{mean} is consistent with the systemic velocity of the

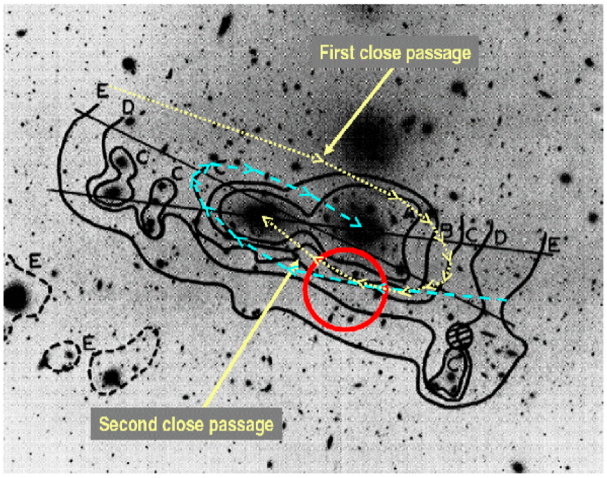


FIGURE 4
Diffuse ICL in the Coma cluster core. The position of the MSIS field studied in Gerhard et al. (2005), Gerhard et al. (2007) is indicated by the red circle. It is superposed on the diffuse light isodensity contours drawn by Thuan and Kormendy (1977). The MSIS field is about 5′ south of NGC 4874. The second Coma cD galaxy NGC 4889 is 7′ east (to the left) of NGC 4874. The bright object north of NGC 4874 is a star. Note the strong elongation of the ICL distribution in the Coma cluster core. The likely orbits of NGC 4889 and NGC 4874 up to their present positions are sketched by the yellow dotted and magenta dashed lines, respectively; image reproduced from Gerhard et al. (2007) with permission.

other D galaxy in the Coma core, NGC 4889. The result of these observations connects the stars and their motions at the location of the MSIS field with a thick plume or tail stripped off NGC 4889 during an on-going binary sub-cluster merger in the Coma core (Fitchett and Webster 1987; Colless and Dunn 1996). Both the kinematics and elongated morphology of the ICL in Coma (see Thuan and Kormendy 1977) show that the cluster core is in a highly dynamically evolving state; this is independently supported by galaxy redshift measurements (Adami et al., 2005) and X-ray data (Neumann et al., 2003). The sub-cluster associated with NGC 4889 is likely to have fallen into Coma from the eastern A2199 filament, in a direction nearly in the plane of the sky, meeting the NGC 4874 sub-cluster arriving from the west. The two inner sub-cluster cores are presently beyond their first and second close passage, during which the elongated distribution of diffuse light was created, see Figure 4. The absence of a cooling core in the Coma cluster is consistent with near head-on collision of the two sub-clusters. Gerhard et al. (2007) also argued that the extended western X-ray arc (see Neumann et al., 2003) traces the arc shock generated by the collision between the two sub-cluster gaseous halos. The high-density mass concentrations associated with the sub-clusters around NGC 4874 and NGC 4889 in the Coma cluster's central region were confirmed independently by weak lensing measurements (Okabe et al., 2010).

4.1.5 Summary kinematics of ICL in local clusters

Kinematics measurements from discrete tracers clearly support the ICL to be a distinct physical component from the CCG halo, which can be offset spatially from the central CCG. The ICL LOSVD has 1) width approaching values expected from the cluster velocity dispersion and 2) may have additional multiple peaks. Those multiple peaks, when present, are related with sub-clusters and their on-going mergers, see for example the binary mergers in the central region of the massive Coma cluster, or the displaced/offset envelope surrounding NGC 3311. Kinematics substructures signaled by independent peaks in the ICL LOSVD correlate with the presence of photometric substructures and satellites at similar LOS velocities.

4.2 Properties of the IGL in local galaxy groups

In this section we present a concise overview of the kinematics of the IGL in nearby local groups. We shall look at Virgo sub-cluster B with its CCG, M49, and the Leo I group with M105.

4.2.1 Kinematics of IGL around M49 in the Virgo sub-cluster B (D = 15 Mpc)

M49 (NGC 4472) is the BCG in the Virgo sub-cluster B (Binggeli et al., 1987). Because of its smooth X-ray halo (Böhringer et al., 1994), the temperature map around M49 can be used to trace the gravitational potential within 30 kpc (Das et al., 2010) and thus alleviate the anisotropy-potential degeneracy. From the extended Planetary Nebula Spectrograph (PN.S) ETG survey (Douglas et al., 2007; Arnaboldi et al., 2017; Pulsoni et al., 2018) and the deep photometric sample obtained with the Subaru SuprimeCam by Hartke et al. (2017), a cumulative PN catalogue was assembled to map the kinematic properties out to 95 kpc radius. Hartke et al. (2017), Hartke et al. (2018) implemented a multi-Gaussian analysis for the LOSVD in order to identify sub-populations with distinct kinematics, which showed the presence of three kinematically distinct components around M49 (Hartke et al., 2018). These authors found 1) a kinematic substructure associated with the on-going accretion of the dwarf irregular galaxy VCC 1249 onto the M49 halo. The two additional kinematically distinct PN populations found around M49 are 2) the main M49 halo and 2) the IGL, respectively. These two components have velocity dispersions $\sigma_{halo} \simeq 170\,kms^{-1}$ and $\sigma_{IGL} \simeq 400\,kms^{-1}$ in the 10–80 kpc radial range, respectively. The kinematically selected PNe which are part of the M49 halo have different stellar population properties from the intra group (IG) PNe which we shall present in the following sections. The velocity dispersion σ_{IGL} of the IGPN profile at 80 kpc radial distance from the M49 center is consistent with a flat circular velocity curve extended outwards from

the region with the X-ray observations. The σ_{IGL} measured profile joins onto that of the satellite galaxies in the Virgo sub-cluster B at 100 kpc radius. This was the first time that the transition from a CCG halo to the surrounding IGL is observed based on the velocities of individual stars. For M49 this transition was later confirmed independently from the LOSVD of blue GCs (Taylor et al., 2021).

4.2.2 Kinematics of IGL around M105 in the Leo I group (D = 10 Mpc)

M105 (NGC 3379) is an ETG in the nearby Leo I group. While classified as a poor group in a low mass environment, the Leo I group is the closest which contains all galaxy types. It thus allows one to explore the physical properties of IGL at the low-mass end of the group mass function. Hartke et al. (2020; 2022) present an extended, imaging and kinematic survey of PNe in M105 and its surrounding, covering 0.25 deg^2 area in the Leo I group with SuprimeCam Narrow band imaging and the PN.S. For previous results from a narrow band survey in [OIII] centered on the HI gas cloud in Leo I group see (Castro-Rodriguéz et al., 2003). Hartke et al. (2022) employ Gaussian mixture models to first separate those PNe that are gravitationally bound to the companion galaxy NGC 3384. Then the M105 PNe (169 in total) are further classified as belonging either to the galaxy's inner Sersic halo or to the surrounding outer exponential envelope, which is traced by an excess of metal poor stars (see Hartke et al., 2020; for reference to the HST observations and analysis see Harris et al., 2007; Lee and Jang 2016). Hartke et al. (2022) constructed the smooth velocity and velocity dispersion fields, and determined projected rotation, velocity dispersion, and λ_R profile for each component. The M105 halo PNe exhibit a rapidly declining velocity dispersion from $\sigma_0 \sim 220$ kms^{-1} at the center to as low as 50 kms^{-1} at about $6R_e$, with $R_e = 57$" (Hartke et al., 2022). Such a negative gradient of the velocity dispersion profile is steeper than that predicted by the $R^{-1/2}$ Keplerian decrease (Romanowsky et al., 2003; De Lorenzi et al., 2009). Differently from the PN subsample linked to the inner Sersic component, the velocity dispersion and rotation traced by the PNe associated with the outer exponential envelope around M105 remain approximately constant at $\sigma_{exp} \sim 160$ kms^{-1} and $V_{rot} = 75$ kms^{-1} at 12 R_e. These velocity dispersion and rotation values traced by the PNe associated with the exponential envelope are consistent with those measured for the dwarf satellite galaxies in the Leo I group. This indicates that the exponential envelope traces the IGL in the Leo I group. The consistency between the kinematic transition from M105 halo and IGL in the Leo I group and photometric components has been recently confirmed independently by the deep photometry carried out with the VST by Ragusa et al. (2022).

4.2.3 Summary kinematics of IGL in local groups

The kinematics observations of the IGL support that the stars in these components originate from the group satellite galaxies. IGL stars are accumulated around the central group galaxies. The IGL in local groups is more relaxed and phase mixed than ICL in clusters. The kinematically selected IGL stars follow exponential profiles away from the central galaxies of these groups.

5 Constraints on the orbits of outer halo, IGL, and ICL stars, and their dynamical origin

From their kinematic properties discussed in the previous sections, several components such as the ICPNe in Virgo or Coma appear far from dynamical equilibrium. Others such as the IGL stars around M49 and M105 or the ICPNe in Fornax have smoother photometric and kinematic properties, suggesting that they are more well-mixed in the gravitational potential. In this section we discuss for several of these populations the constraints that can be obtained on their orbits, their accretion history, or the mass distributions in their host systems. We begin by reviewing some general results that we draw on.

5.1 Some results from dynamical modelling and cosmological simulations

From dynamical modelling (e.g., Gerhard et al., 2001; Poci et al., 2017) and strong lensing analysis (e.g., Koopmans et al., 2009; Li et al., 2019) the inner high surface brightness regions of massive ETGs have near-isothermal total mass density profiles, i.e., power laws $\rho \propto R^{-\gamma}$ with $\gamma \approx 2.0$–2.2. To measure the total mass distribution from stellar kinematics at radii where dark matter contributes substantially, requires information on the shape of the stellar LOSVD, often parametrized through its third (h3) and fourth (h4) and possibly higher-order Gauss-Hermite moments (Gerhard 1993; van der Marel and Franx 1993). The parameter h3 is proportional to the skewness and measures the asymmetric deviations of the LOSVD from a Gaussian, often related to rotation when the low-velocity tail of distribution dominates. The parameter h4 is proportional to the kurtosis of the LOSVD; it probes symmetric deviations of the LOSVD, indicating either a more peaked ($h4 > 0$) or top-hat ($h4 < 0$) shape with respect to a Gaussian. Even in spherical potentials, h4 values depend on the orbital anisotropy, the stellar density profile, and the gravitational potential (Gerhard 1993); therefore determining anisotropy and mass distribution from projected velocity dispersion and h4-profiles generally requires dynamical modelling relating all of these variables.

For ETGs with relatively steep stellar density profiles in typical dark matter halos, radial anisotropy in the outer regions leads to radially decreasing projected LOS velocity dispersion profiles and positive h4, caused by an overabundance of stars at zero LOS velocity (Gerhard et al., 1998). This has been analyzed in several lower-mass ETGs (de Lorenzi et al., 2009; Morganti et al., 2013). On the other hand,

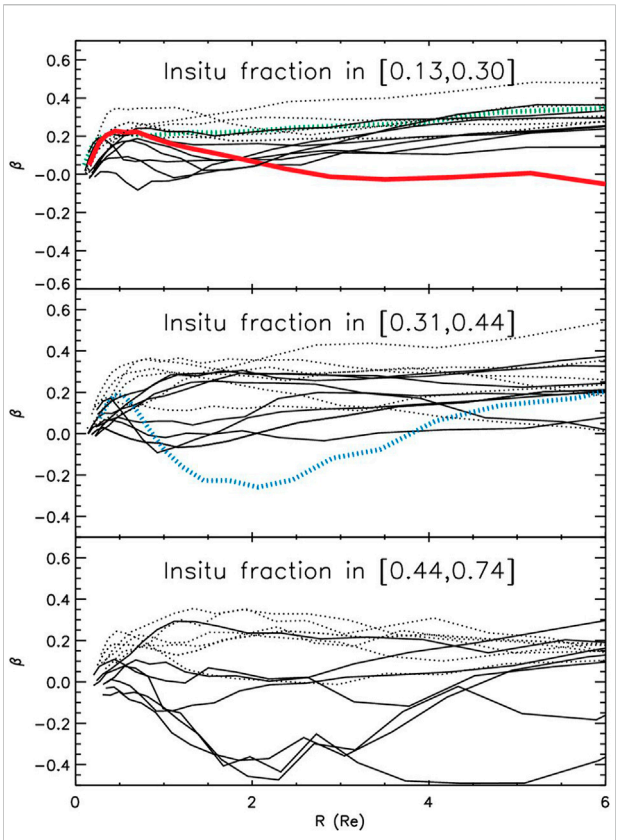


FIGURE 5
Velocity anisotropy profiles for the simulated galaxies from Wu et al. (2014) in bins of *in situ* fraction of stars within 10 per cent r_{vir}. Solid lines are for fast rotators while dotted lines are for slow rotators. Colored curves depict the three fiducial galaxies: green dotted line for a slow rotator, blue dotted line for a slow rotator with peaked λ(r) profile at $2R_e$, and red solid line for a fast rotator. Galaxies with low *in situ* fractions have more radially anisotropic orbit distributions, while tangential anisotropy is seen only for systems with high *in situ* fraction. Image reproduced from Figure 11 of Wu et al. (2014) with permission.

kinematic studies of the most luminous ETGs ($M_K < -25.7$; Veale et al., 2017) measuring h3 and h4 values out to two effective radii with extended IFS (the MASSIVE survey, Veale et al., 2017) or deep long-slit (Bender et al., 2015) data, found high mean values <h4 >≈ 0.05 at 2 Re while values at the centers are near-zero (e.g., van de Sande et al., 2017). The high h4 values correlate with increasing projected velocity profiles and both are likely caused by mass distributions increasing outwards faster than isothermal (Veale et al., 2018; Wang et al., 2022). We use this to interpret some of the IGL kinematic data below.

In cosmological simulations, the stellar particles found in the simulated galaxies are divided according to two possible origins. They may form out of gas within the host galaxy, either at early times during a period of rapid star formation, or from gas accreted later; in this case they are tagged as the "*in situ*" component. Or they are stellar particles formed in satellite galaxies which are later accreted or merged with the host; this second set is labelled "*ex-situ*". The accreted stars may be predominantly found at large radii (Oser et al., 2010) but for the most massive galaxies accreted stars dominate at all radii (Pulsoni et al., 2021). The accreted stellar component is characterized by a radial orbit anisotropy (Abadi et al., 2006; Hilz et al., 2012), because the satellites that dissolved and merged with the host come in on predominantly radial orbits. This suggests that the kinematics of simulated galaxies should be more or less radially anisotropic depending on the relative fractions of *in situ versus* accreted (*ex situ*) stars. In Figure 5 we reproduce the anisotropy profiles $\beta(r) \equiv 1-(\sigma^2\theta + \sigma^2\varphi)/2\sigma_r^2$ for the simulated galaxies in Wu et al. (2014), divided into bins of *in situ* fraction. They indeed find that tangentially anisotropic simulated galaxies can only be found in the group with large fraction of *in situ* stars. Almost all simulated galaxies with low *in situ* star fraction (upper panel) are radially anisotropic with β ~ 0–0.3. However, many of the systems with higher *in situ* fraction have similar anisotropies. As a result, Wu et al. (2014) find no correlation between both quantities for their whole galaxy sample. For their radially anisotropic galaxies the β(r) profiles are nearly constant with radius, typically with β ~0.1–0.3 for R > $2R_e$. They also find that most slow rotators are radially anisotropic, whereas simulated fast rotator galaxies have both radially andtangentially biased anisotropy profiles. Wang et al. (2022) studied the distribution of β values around 1.5 effective radii in ETGs from the TNG simulation (Springel et al., 2018). They find that galaxies with flat or rising outer velocity dispersion profiles are generally radially anisotropic, whereas galaxies with falling outer dispersion profiles can have a wide range of anisotropies at this fiducial radius. These results appear broadly consistent with the findings of Wu et al. (2014).

5.2 Case studies of nearby groups and clusters

5.2.1 Example of increasing circular rotation curve with radius: M49 in Virgo sub-cluster B

In M49 the rising σ_{LOS} profile with radius reaching the velocity dispersion of the Virgo Subcluster B (Binggeli et al., 1993), from 300–500 kms^{-1} at 100–250 kpc radial distance from the center of M49 arises from the transition between the CCG halo and an outer stellar component with a larger velocity dispersion. These two components overlap along the same LOS following different surface brightness distribution (Spavone et al., 2017). Hartke et al. (2018) constructed the σ_{LOS} profile for the faint subsample of PNe in M49, which most likely trace the relaxed smooth component. They used the v_{LOS} of the faint PNe in the radial range 10–80 kpc around M49 to build the global LOSVD; this LOSVD turned out to have strong high-velocity wings. The h4 value is h4 = 0.11 +/-0.03 and a much

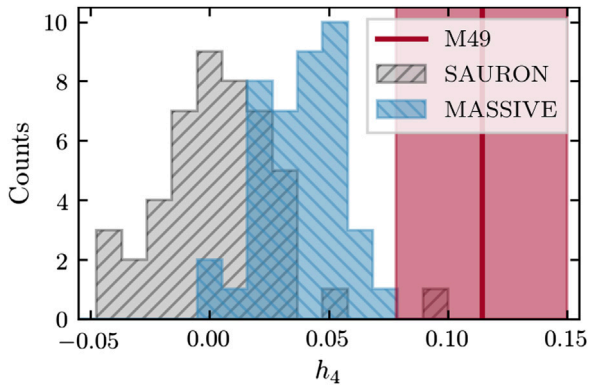


FIGURE 6
Best-fit h_4 value and error band (red vertical line and shaded region) derived from the LOSVD of the M49 faint PN sample compared to the luminosity averaged h_4 distribution of 47 and 41 ETGs from the SAURON (Emsellem et al., 2004) and MASSIVE (Veale et al., 2017) surveys (grey and blue histograms). Image reproduced from Hartke et al. (2018) with permission.

smaller h3 = 0.01 +/- 0.03 (Hartke et al., 2018). For ordinary massive ETGs, h3 has usually larger values than h4 (Bender et al., 1994). To further illustrate the uncommonly large value of h4 for M49, Hartke et al. (2018) compared the h4 value for the M49 smooth halo to the distribution of luminosity-averaged h4 values of ETGs in the SAURON[4] (Emsellem et al., 2004) and MASSIVE (Veale et al., 2017) surveys. As shown in Figure 6, the large h4 value for the extended PN velocity field in M49 is larger than the h4 for any of these galaxies at smaller radii, including M49 itself (at 2 R_e, Veale et al., 2017). The implied positive radial h4 gradient and the outward dispersion gradient both point to a significant increase of the circular velocity in the transition from the galaxy to the surrounding group.

We note that the core-winged structure of the LOSVD associated with large positive values of h4 at large radii around M49 can emerge quite naturally from the superposition of two LOSVDs, a narrower and a broader one as predicted by the simulation of CCG and ICL by Dolag et al. (2010).

5.2.2 Rotation and anisotropy in outer halo

According to cosmological simulations, both the CCG halos and ICL in clusters result from hierarchical accretion of stars; however, they differ in their spatial distribution, degree of dynamical relaxation and/or phase-space mixing (Cooper et al., 2015).

4 The SAURON survey covers the central 0.5–1 Re, where the h4 values are generally small or even slightly negative (see also Bender et al., 1994).

In this section we focus on the outer regions of M87 as an example where ICL and CCG stars can be tagged on the basis of their kinematics at the same spatial location.

The IFS VIRUS-P data of Murphy et al. (2011), Murphy et al. (2014) measure a steep increase of σ_{LOS} at R ≥ 40 kpc in M87. The comparison of these measurements with the V_{LOS} distribution mapped by PNe suggests that the cause for their steep rise in the σ_{LOS}(R) profile is the contribution from the ICL. At R ≥ 40 kpc, the M87 CCG surface brightness decreases rapidly, while the ICL becomes significant, see Figure 7 which shows the ICL cospatial with the outskirts of M87 and with a large dispersion characteristic of the Virgo cluster core (Longobardi et al., 2015a; 2018a). Once the broad ICL LOSVD and the contribution from the localized substructure (e.g., the crown, Longobardi et al., 2015b) are subtracted, the remaining smooth M87 halo has a LOSVD which is centered on the galaxy's systemic velocity (1,307 kms^{-1}). The mean rotation velocity of the CCG halo is ~25 kms^{-1}, and the velocity dispersion profile, after rising slowly to 300 kms^{-1} at R_{avg} ~50–70 kpc, decreases rapidly to 100 kms^{-1} at R_{avg} ~135 kpc (Longobardi et al., 2018a). In the same region from R > 40 kpc the halo surface brightness profile drops below the Sersic fit to a steeper $I(R) \propto R^{-\gamma}$ with γ=(2–2.5) and the circular rotation curve obtained from X-ray data (Churazov et al., 2010; Simionescu et al., 2017) turns from flat to steeply rising. Local Jeans models (Churazov et al., 2010; Longobardi et al., 2018a) show that, with this configuration, the measured σ_{LOS}(R) profile out to 135 kpc can be reproduced by an isotropic stellar orbital distribution at R ≤ 60 kpc which then become strongly radially anisotropic in the radial range 70–135 kpc.

5.2.3 NGC 3311 and the Hydra I cluster

An interesting case where the increasing velocity dispersion profile σ_{LOS}(R) correlates with higher moments of the LOSVD is that of the halo in NGC 3311 and the ICL in the Hydra-I core. For NGC 3311, Barbosa et al. (2018) were able to measure the h3 and h4 profiles from the MUSE IFS data, in addition to V_{LOS} and σ_{LOS} profiles which are shown in their Figure 16. The skewness h3 profile has about null values out to 2 kpc and then becomes negative at distances larger than 10 kpc, where both V_{LOS} and σ_{LOS} increase. The kurtosis h4 profile has also near null values at the center, increases to ≈0.1 at 2–4 kpc, and then decreases to smaller positive values at radii >10 kpc. The positive h4 values measured for NGC 3311 are similar to those measured in other galaxies of the similar luminosity class (Veale et al., 2017). Positive h4 and increasing σ_{LOS}(R), as observed in NGC 3311 (see also Hilker et al., 2018) suggest a mass distribution increasingly outwards faster than isothermal (Section 4.1), and are reminiscent of the properties of simulated massive ETGs in cosmological simulations analyzed by Wu et al. (2014). The most massive galaxies in these simulations have radially increasing circular velocity curves, radial anisotropy in the 2–5 R_e region and a large fraction of accreted *ex-situ* stars (Wu et al., 2014).

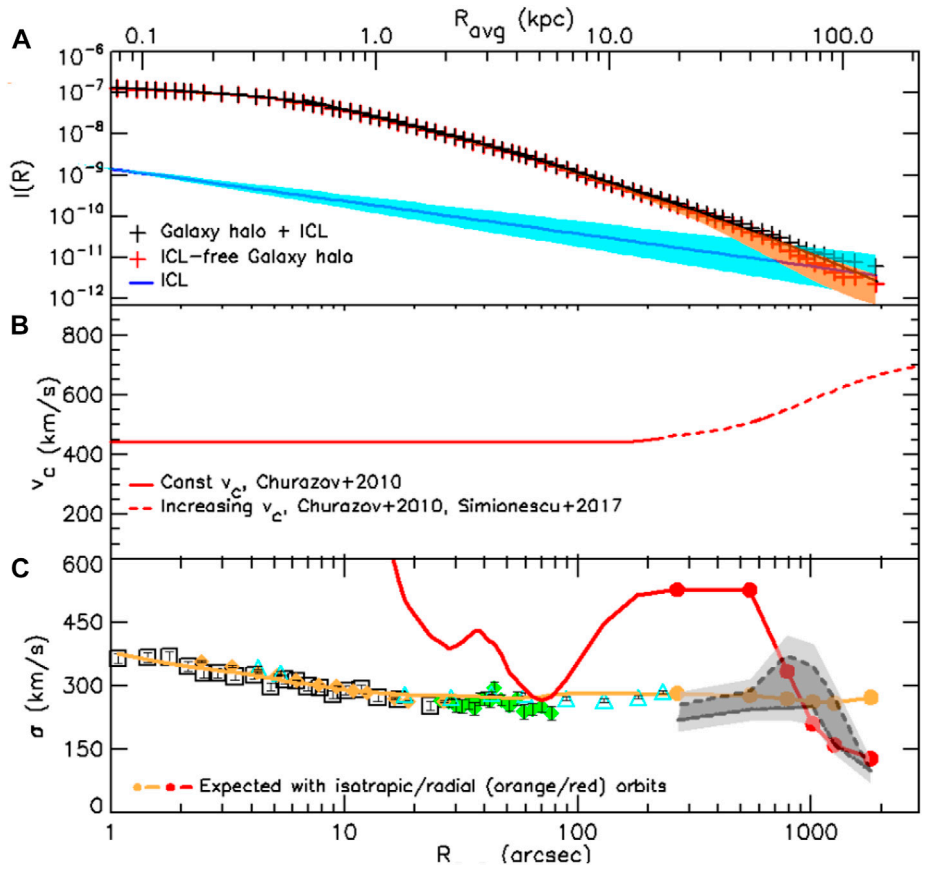


FIGURE 7
Dynamics of the smooth M87 halo out to 2000″ compared with predictions from Jeans equations. Panel **(A)** surface brightness profile of the M 87 smooth halo (red crosses and shaded area) obtained from the extended photometry (Kormendy et al., 2009), after subtracting the ICL contribution (blue line and shaded area). The black continuous line shows the Sersic fit with n = 11 to the total halo + ICL extended photometry. The adopted circular velocity profile based on X-ray data is shown in panel **(B)**. Panel **(C)** velocity dispersion profile $\sigma(R)$ from IFS (squares, triangles) and long slit (diamonds) data (source: Longobardi et al., 2018a). Large dots represent the expected σ values at the average radii of the radial bins, for isotropic (orange) and completely radial (red) orbital anisotropy. The comparison with the range of σ values for halo PNe (shaded area) suggests that the distribution of orbits changes from near-isotropic at ~200″ to strongly radial at ~2000″. Image reproduced from Longobardi et al. (2018a) with permission.

5.2.4 The radial orbital anisotropy of the ICL stars in the Fornax cluster

In Figure 3, we showed the velocity dispersion $\sigma_{LOS}(R)$ profile of the NGC 1399 halo and the ICL in the Fornax cluster out to 200 kpc. Between 100 and 200 kpc, the $\sigma_{LOS}(R)$ profile is flat at a value $\sigma_{LOS} \approx 300$ kms^{-1}, beyond the peak at 400 kms^{-1} reached at ~ 50 kpc from the center of NGC 1399. When comparing with the velocity dispersion of Fornax cluster galaxies, obtained by averaging in annuli their V_{LOS} at about similar distances from the cluster center, Spiniello et al. (2018) found a negative difference of $\Delta\sigma$~80 kms^{-1} between the velocity dispersion of Fornax satellite galaxies (σ_{cl}) and that measured for the ICPNe. The lower value for σICPNe in this radial range is consistent with a dynamical model where the ICPNe orbit in the same potential as the Fornax cluster satellite galaxies, have constant radial anisotropy parameter (β = 0.3) but have a different, steeper, density profile than that of the Fornax cluster galaxies (averaged in circular annuli). When both distributions are fitted with a power law $R^{-\gamma}$, $\gamma_{ICPNe} = 3.0$ while $\gamma_{cl} = 2.0$ at R = 110 kpc, supporting the view that the stars building the ICL are those that were stripped from the satellites on radial orbits that plunge deep into the cluster potential.

6 Stellar populations and progenitors of IGL and ICL in nearby groups and clusters

6.1 Learning from simulations

Murante et al. (2004), Murante et al. (2007) investigated the spatial distribution and origin of the DSC in simulations in more than hundred galaxy clusters, with masses M > 10^{14} h^{-1}, which

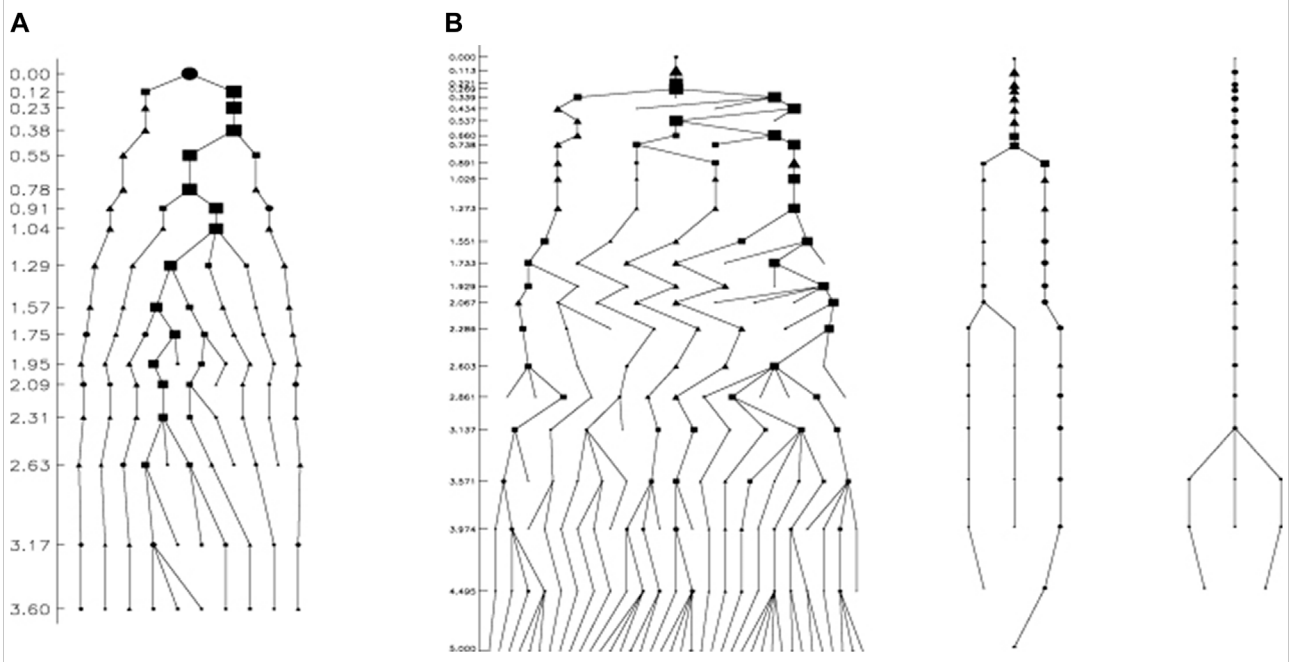


FIGURE 8
(A): family tree of the BCG galaxy in a low-resolution simulation. **(B)**: family trees of the BCG galaxy, the third most massive galaxy and a lower mass galaxy in the high-resolution re-simulation of the same cluster. The size of symbols is proportional to the logarithm of the mass of the galaxies at the corresponding redshift. Shown on the vertical axis on the left-hand side are the output redshifts used to reconstruct the family trees; these are different in both simulations. A galaxy in these trees is considered a progenitor of another galaxy if at least 50 per cent of its stars are bound to its daughter galaxy, according to the SKID algorithm. Many more galaxies can be identified in the high-resolution simulation at similar redshift. The BCG family tree is characterized by one dominant branch with a number of other branches merging into it, at both resolutions. Squares and triangles are assigned according to the classified 'merging' or "stripping" events. Circles correspond to redshift at which the galaxy is not releasing stars to the DSC. Image reproduced from Figure 3 of Murante et al. (2007) with the permission.

were selected from a cosmological hydrodynamical simulation. Murante et al. (2004) applied a Friend-of-Friend algorithm to remove all satellite galaxies, and then studied the DSC which included CCG and ICL stellar particles. They found evidence for a DSC in the cluster volume once the cluster satellite galaxies (but the CCG) were removed. For each DSC star particle at z = 0 in these clusters, Murante et al. (2007) looked for the satellite cluster galaxy to which this particle once belonged at an earlier redshift, thus linking the presence of the DSC stars to the formation history[5] of the cluster. Examples of merger trees are given in Figure 8. The main results of their analysis are that on average, 50% of the DSC star particles come from galaxies associated with the family tree of the most massive galaxy, i.e., the BCG. For the other half, 25% comes from the family trees of other massive galaxies, and the remaining 25% from dissolved galaxies: clearly the formation of the DSC is parallel to the build-up of the BCG and other massive galaxies. Most DSC star particles become unbound during mergers in the formation history of the BCGs and of other massive galaxies, independent of cluster mass.

Murante et al.'s (2007) results further suggest that the tidal stripping mechanism is responsible only for a minor fraction of the DSC[6]. At cluster radii larger than 250 h^{-1} kpc, the DSC fraction from the BCG is reduced and the largest contribution comes from the other massive galaxies. In the cluster outskirts, galaxies of all masses contribute to the DSC and most of the star particles come from stripping. An important aspect of ICL formation is the so called preprocessing (Rudick et al., 2006; Contini et al., 2014), by which is meant the

5 Galaxies were identified in the simulated clusters at several redshifts, starting with z = 3.5, and then the family trees were built for all the z = 0 cluster galaxies in Murante et al. (2007). The most massive cluster galaxies show complex family trees, resembling the merger trees of dark matter haloes, while the majority of other cluster galaxies experience only one or two major mergers during their entire life history. Fraction of DSC particles associated with the merger part of the family trees, and with the stripping part of the family trees, were computed as follows. A DSC star particle to arise from a merger at redshift zj if the galaxy it was last bound to has more than one progenitor at z j–1. The DSC particles coming from the progenitors at z j–1 are also defined as arising from a merger. A DSC star particle is classified to be unbound through stripping if the galaxy it was last bound to at redshift zj has only one progenitor at zj–1.

6 The classification of DSC and ICL stars as originating from either "merger" or "stripping" in simulations depend on the adopted "operational" definition. See discussion in the review by Contini (2021).

formation of diffuse DSC in a smaller entity, e.g., galaxy group, before this later merges with the cluster and releases its DSC to the cluster's ICL when disrupted. For a more extensive discussion of ICL formation processes see the review by Contini (2021).

The DSC does not have a preferred redshift of formation: however, most DSC stars are unbound at $z < 1$ (see also similar results from semi-analytical model predictions by Contini et al., 2014). The amount of DSC stars at $z = 0$ does not correlate strongly with the global dynamical history of clusters and increases weakly with cluster mass.

Cosmological simulations like those discussed above show that different galaxies in a cluster have distinct accretion histories with alternative channels contributing to the gravitationally unbound stars which populate the cluster volume. If these different progenitors contribute tracers of different nature, PNe, red or blue GCs, to the DSC with distinct frequencies (which may depend on the stellar masses, hence absolute B band magnitudes of the progenitors and also on galaxy types, see Coccato et al., 2013), the different origin or progenitors may bring about different velocity fields and distinct spatial distributions for the different discrete tracers. Diversity in either the spatial distribution and/or kinematics may provide interesting opportunities to reconstruct the mass assembly, i.e., to map which progenitors brought which stars to the DSC and at what radii.

6.2 PNe as tracers of stellar populations: PN specific frequency and PNLF morphology

In Section 4 we argued that the CCG halo, IGL and ICL in groups and clusters can be identified according to their different LOSVDs, and thus it is possible to tag tracers according to which kinematic component they belong to at the same spatial position. In this section we discuss the difference in the stellar population properties of the stars kinematically tagged as linked either to the CCG halo or to the IGL and ICL, as from their PN population properties.

In M105, M49, and M87 the kinematically tagged IC and IG PNe show different stellar population properties from the CCG halo PNe (Longobardi et al., 2013; Longobardi et al., 2015a; Hartke et al., 2017; Longobardi et al., 2018a; Hartke et al., 2020). As a general trend, the inner PN number density profile follows the Sersic fit to the CCG V-band surface brightness profile, while the ICPN number density profile is consistent with a shallower power law, ($N_{ICPN}(R) \propto R^{-\gamma}$) or exponential. In detail, the number density profiles of PNe become "flatter" than the surface brightness V band profiles of these bright galaxies in the cluster/group region at radii smaller than those at which the change of slope associated with the excess of light from the IGL or ICL occurs. This distinct outer behavior of IC PNe with respect to the CCG halo population is linked to the different values of the PN specific yields, or α parameter, where $\alpha = N_{tot,PN}/L_{bol}$. The α parameter values of the PN populations associated with CCG, IGL or ICL, are measured empirically and turn out to be different. Current observations show that the IGL and ICL PN populations have several times larger PN yield values per bolometric luminosity than the inner CCG halo light, in the range 3–7 time larger. The measured larger values of the PN yield values for the IGL and ICL ($1.6{\cdot}10^{-8}$ 10^{-7} PN L^{-1}_{BOL}) are similar to the α-parameter values measured for the PN populations detected in dwarf irregular galaxies of the Local Group, such as Leo I, Sextans A, and Sextans B (Buzzoni et al., 2006). The association of high PN yields with metal poor stellar populations is further explored and supported by the studies of the resolved stellar population in the Leo I group, which is summarized in Section. 6.4.

Another important observational result that supports differences among parent stellar populations is linked to the morphology of the PN luminosity function (PNLF) for these CCG, ICL or IGL PN populations. In the case of M105 and M49, according to the generalized PNLF formula (Longobardi et al., 2015a), the PNLF for the IGL component has a steeper gradient quantified by a larger value of the c_2 parameter (in the range 1–1.5) than that measured for the CCG halo PNLF (c_2 value in the range 0.5–0.7, Hartke et al., 2017, Hartke et al., 2020) or the Ciardullo's original PNLF analytical formula (c_2 value equal to 0.302, Ciardullo et al., 1989). The PNLF steeper slope is related with a larger number of core stars with smaller masses, as observed in older stellar populations like those in bulges of nearby galaxies (Ciardullo et al., 2004). In the Virgo cluster, the IC PNLF displays a "dip" (Longobardi et al., 2018a) which is reminiscent of a similar morphological feature detected for the SMC/LMC PNLFs (Jacoby and De Marco 2002; Reid and Parker 2010) and a value of the c_2 parameter equal to 0.66.

In summary, the larger PN yield per bolometric luminosity measured for the IGL and ICL and these measured similarities of the IC PNLFs with those observed for low-luminosity, low metallicity, star forming irregulars support the inference that stellar population in the IGL and ICL traced by PNe come from lower mass progenitors than those which deposited the accreted stars at smaller radii within the CCGs.

6.3 Resolved stellar population of the ICL in the Virgo cluster

Direct detections of IC red giant branch (RGB) stars in the IC regions around M87 come from the measurement of an excess of unresolved number counts in deep I band (F814W) HST images, with respect to similarly deep cosmological fields (see for example the Hubble UDF). The first case reported of such an

excess was measured by Ferguson et al. (1998) in a field 44 arcmin east of M87. Subsequent measurements at independent positions were performed by Durrell et al. (2002) and Williams et al. (2007), in two distinct fields ~40 arcmin north-west of M87. These number count excesses display a sharp edge/increase at the apparent magnitude m_{F814W} ~ 27, which corresponds to the apparent magnitude of the tip of the RGBs at a distance modulus (m-M)≈31. From the fit with theoretical stellar evolution models to the color magnitude diagram (CMD) of these sources, it was determined that about 70% of the stars in the WFP2/ACS IC fields are at a common distance of ~16 Mpc (+/-0.1 mag). Adopting a technique which is similar to both Durrell et al. (2002) and Williams et al. (2007), Yan et al. (2008) serendipitously detected an excess of point-like sources in the I band (F814W) HST images available for two ACS fields, at 7 arcmin and 15 arcmin respectively from M60 (NGC 4649). The number density of the excess I band field population shows a decreasing gradient with increasing radial distance from M60. This population of stars around M60 occurs at apparent magnitude m_{F814W} ~ 26.2; its inferred average surface brightness is 26.6 mag arcsec^{-2} in the I band in the field at 7 arcmin from M 60. These values are very similar to those measured by Durrell et al. (2002) observations of the IC RGBs in the Virgo core. Yan et al. (2008) argued that they detected an IG stellar population associated with the M60/M59 galaxy pair. The results linked to the detected excess of I band counts may indicate that the M60/M59 sub-group is yet another region of the Virgo cluster where IGL is being assembled.

The deep HST/ACS images acquired at the IC positions in the Virgo cluster at 40 arcmin (190 kpc) from M87, and 37 arcmin (170 kpc) from M86 in the wide V (F606W, ~63 k secs) and I band (F814W, ~27 k secs) filters, were analyzed by Williams et al. (2007). These authors analyzed the I band luminosity function for the ~5300 IC RGB stars (once corrected for background unresolved contaminants) and their CMD distribution. At this sky location, the excess number count population has a corresponding mean V-band surface brightness (excluding background galaxies) of μ_v ~27.7 mag arcsec^{-2}, in agreement with the deep image of the Virgo core from Mihos et al. (2005). The stellar isochrone fits to the CMD showed that the Virgo's IC RGB population is dominated by old, metal-poor stars. The best-fit isochrone models by Williams et al. (2007) indicate that 70%–80% of the stars have ages that are greater than 10 Gyr; these stars have a median metallicity of [M/H] ~ -1.3 and fall in the range −2.5 < [M/H]< −0.7. However, there is some evidence for the existence of a younger, metal-rich component. Integrated over the entire mass function, the younger component contributes 20%–30% of the total population. The V-I data by Williams et al. (2007) are the deepest CMD available for resolved RGB stars at the Virgo cluster distance to date.

6.4 Progenitors of IGL in the Leo I group and Virgo sub-cluster B

The V band profile of M105 in the Leo I group is described by a superposition of a Sersic profile (n = 4) and an outer exponential envelope (Watkins et al., 2014; Ragusa et al., 2022). The PNe distribution from Hartke et al. (2020) shows a flattening with radius at large radii, corresponding to an higher PN yield α value in the outer regions. Harris et al. (2007) and Lee and Jang (2016) analyzed HST deep images acquired at two distinct locations, respectively at 4 arcmin and 15 arcmin from the M105 (NGC 3379) center. They assumed a 10 Gyr old age for the resolved stellar population in these fields. This assumption is supported by the general old age of M105 and independently by the steep slope of the PNLFs in the inner and outer regions (Section 6.2). The CMDs show evidence for a metal poor population whose fraction increases outwards, with the number density of metal poor stars with [M/H] < −1.0 exceeding that of the metal richer ones ([M/H]>-1), at approximately ~12 R_e from M105. Using these results, Hartke et al. (2020) constructed a photometric model which conclusively established the association of the exponential envelope with the high PN yield value and the [M/H] < -1.0 metal poor resolved population. In Section 4.2, we already discussed the kinematics association of the exponential envelope with the IGL around M105, which has a σ = 160 kms^{-1}. Thus the IGL around M105 must be dominated by metal poor stars with [M/H] < −1.0.

In M49, the CCG of Virgo sub-cluster B, the V band profile (Kormendy et al., 2009; Capaccioli et al., 2015) follows an inner Sersic profile (n = 6.0), with an additional flatter, outer distribution. The PN distribution from Hartke et al. (2017) shows a flattening at large radii, corresponding to an outer component with high PN yield α as established with a similar photometric model as in M105. In Section 4.2 we presented the evidence that this outer high α component is associated kinematically with the IGL component surrounding M49, with σ = 400 km^{-1}.

In M49, no resolved stellar populations data are available, however Mihos et al. (2013), 2017) published a deep B-V color profile and 2D color map of the halo and IGL out to 100 kpc from the galaxy's center. The profile shows a strong gradient to bluer color in the outer parts reaching B−V ≈ 0.6 at the largest radius (Mihos et al., 2013). For a 10-Gyr old stellar population, the corresponding metallicity would be [Fe/H] < −1.5, see the simple stellar population tracks in Figure 3 in Mihos et al. (2013). The assumption of an ``old" age for this component is independently supported by the smooth spatial appearance of the PN (Hartke et al., 2018) and light distributions (Capaccioli et al., 2015) which suggest a dynamical age for the IGL significantly older than the orbital precession time, which is at least 5 Gyr around M49. The blue color, together with the age constraint, and the high α value suggest a similarly metal poor IGL population surrounding M49 as in the Leo I group.

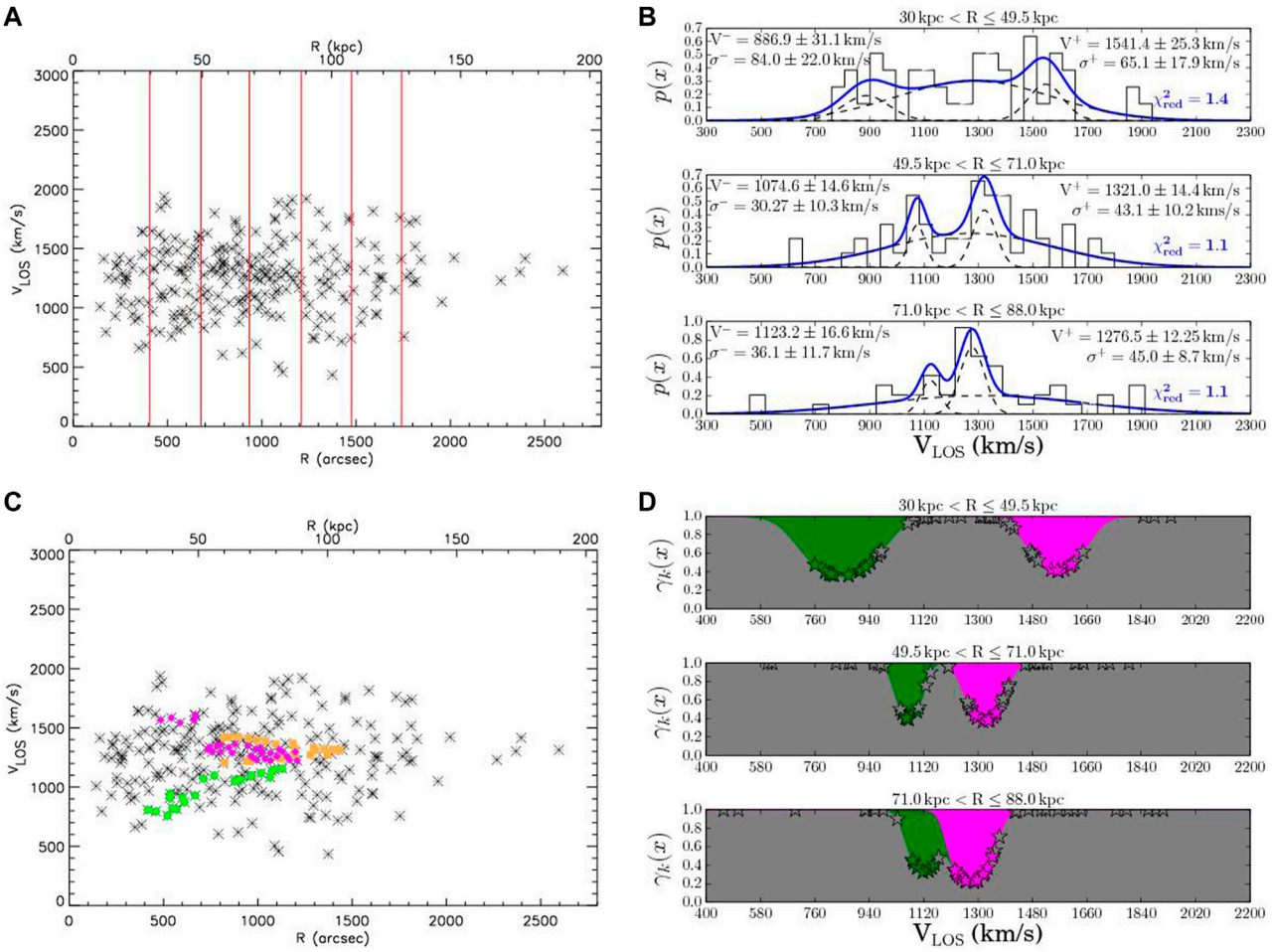


FIGURE 9
(A): projected phase space, V_{LOS} vs. major axis distance (R), for all spectroscopically confirmed PNe (black asterisks) in the halo of M 87 from Longobardi et al. (2015a). Red lines separate the elliptical annuli to isolate the cold components associated to the substructure. **(B)**: LOSVD histograms in the three elliptical annuli. In each panel, the blue lines show the best-fit model computed as a combination of three Gaussians. Black dashed lines show the relative contribution of each component to the LOSVD, with parameters for the cold components given in the plot. **(C)**: as in the panel **(A)**, however, the green circles and magenta diamonds show the PNe associated with the cold secondary peaks in the LOSVD. Orange squares show a kinematically selected GC substructure from Romanowsky et al. (2012). **(D)**: probability that a PN is drawn from the halo component (dark grey area) or from the chevron (green, magenta areas). Stars represent PN probabilities at their measured V_{LOS}. Image reproduced from Longobardi et al. (2015b) with permission.

6.5 The crown in M87: An example of a CCG family tree and extended mass assembly

A vivid example of an on-going accretion event in a halo of a CCG is the "crown" in M87 (Longobardi et al., 2015b). This event is traced by a sub sample of M87 PNe whose projected position-velocity phase space displays an over-density with a distinct chevron-like feature, being the result of the incomplete phase-space mixing of a disrupted galaxy, see Figure 9. For numerical simulations of accretion events with incomplete phase mixing in the M87 halo see Weil et al. (1997). A deep optical image in Longobardi et al. (2015b); processed with unsharped masking) shows the presence of a crown-like substructure that contributes ~60% of the total light at major axis distances of R ~ 60–90 kpc. The on-going disruption of a recently accreted satellite in the halo of M87 is supported by the azimuthal variation which is observed in the (B-V) color map of M87, with a higher fraction of tracers where the galaxy color is bluer and the velocity substructure has the largest number of tracers. In turn there is a deficit of tracers along the photometric minor axis, where the galaxy is redder, see Figure 2 in Longobardi et al. (2015b). An investigation for kinematic substructures in the projected phase-space of the M87 GCs

was also conducted by Romanowsky et al. (2012), which resulted in the discovery of a chevron-like structure also in the projected phase space. Though the morphology in the projected phase spaces for PNe and GCs is alike, they differ in a number of quantitative properties. The width of the chevron for the GCs projected V_{LOS} versus major axis distance (R) phase space goes to zero velocity at $R_{GC} \sim 1{,}500''$ with $V_{LOS,GC} = 1{,}307$ kms^{-1}, see orange squares in Figure 9, while the width of the chevron for the PNe goes to zero at $R \sim 1{,}200''$ with $V_{LOS,PN} = 1{,}250 \pm 21$ kms^{-1}, see green circle and magenta diamonds in Figure 9. Furthermore, the GCs associated with the chevron in the GC projected phase space show a different spatial distribution from the selected chevron PNe. The highest density of chevron GCs occur on the NE photometric minor axis of M87 (Romanowsky et al., 2012), while only a few GCs are found near the M87 major axis, where the PNe associated with the ``crown'' substructure are most frequent in number. The identification of a disrupted satellite *via* the observed "chevron" feature in the M87 outer halo's phase-space supports the size growth and late assembly of giant elliptical galaxies' halos.

6.6 Stellar populations of the ICL around M87 in the Virgo cluster core

In M87, the CCG of Virgo sub-cluster A, the V band profile (Kormendy et al., 2009) follows a Sersic profile with high index (n = 10.0), out to 150 kpc major axis distance, with the two furthest photometric points reflecting the additional contribution from the ICL (Mihos et al., 2005; Mihos et al., 2017). Longobardi et al. (2013) showed that PN distribution can be decomposed into an inner halo component following the Sersic profile and an outer ICL component with 3 times larger PN yield α. From the kinematic decomposition described in Section 4.2, Longobardi et al. (2018a) showed that the number density distribution of the PNe identified as the "smooth" M87 halo follow the same Sersic profile as the light, and the ICPN component, which includes the unrelaxed component of the M87 halo and the broad σ = 900 km^{-1} component, follows a power law $R^{-\gamma}$ with γ = 0.79 ± 0.15.

A 2D color map for M87 and its immediate surroundings was published by Mihos et al. (2017) which showed a strong color gradient in the outer M87 halo, reaching colors consistent with B-V = 0.6 at 135 kpc major axis distance. The higher PN yield α ICL component dominates the surface brightness at these distances around M87. We know from the previous IGL discussion in Section 6.4 that this is associated with metal poor stellar population. This inference is independently supported by the best-fit isochrone models by Williams et al. (2007) to the ACS field CMD between M87 and M86. These indicate that 70%–80% of the stars in this field have ages greater than 10 Gyr and a median metallicity of [M/H] ~ −1.3, in the range −2.5 < [M/H] < −0.7. In addition Williams et al. (2007) found a 20% contribution from a population with [M/H] ~ -0.3 and younger ages (2–3 Gyr). The HST-ACS data were later re-analyzed by Lee and Jang (2016) assuming a 10 Gyr age, α-enhanced ([α/Fe] = 0.2) population using Dartmouth isochrones (Dotter et al., 2008). In their analysis they found a larger contribution from subsolar metallicity stars, about 50%, similar to their results for the outermost field from M105 in the Leo I group.

In the case of the M87 halo and surrounding ICL, the dynamical age is not constrained. However the PNLF of the IC PNe has steep slope consistent with an older age but at the same time shows the "dip" observed for star forming populations making a case for a contribution from stars with younger ages in this area. Thus the combined properties of the Virgo ICPN population near M87 appear to be most readily explained if this population derives from a faded population of low-luminosity, low-metallicity, star-forming or irregular galaxies, such as M 33 or the LMC, which are very different from M87 itself, possibly with additional stars from more massive galaxies that later merged with M87. Longobardi et al. (2018a) found that the IC stars sampled in their survey fields correspond to ~1.5 M 33- or ~4 LMC-like galaxies.

6.7 Stellar population constraints on ICL in massive clusters from absorption line spectra

Within the limit imposed by long integration times (to reach suitable S/N) because of the steeply decreasing surface brightness profiles, efforts are carried out to constrain the metal content of stars in the CCG halos and ICL, using stellar absorption line features of the stellar continuum. In general stars in massive ETGs at 2 R_e or further out have high [α/Fe] (Greene et al., 2013; Bender et al., 2015) and produce measured profiles with declining metallicity gradients (Zibetti et al., 2020) and flat-to- positive alpha element gradients (Coccato et al., 2010, 2011; Kuntschner et al., 2010). The high [α/Fe] is associated with stellar populations formed during a short, intense burst of star formation (Thomas et al., 2005). Negative metallicity gradients which extend to several R_e are found in most ETGs (e.g., Baes et al., 2007; Spolaor et al., 2010; La Barbera et al., 2012).

As an example, the outer envelope of NGC 3311 provides evidence of un-mixed stellar populations with different [α/Fe] values and metallicity (Barbosa et al., 2016; Barbosa et al., 2021). At the radii where the σ increases, the stars in the off-centered exponential envelope are younger that those in the central regions, with lower metallicity but α-enhanced. Barbosa et al. (2021) stated the because of the central location of NGC 3311 in the Hydra I cluster the accreted stars which build the ICL come mostly from rapidly quenched satellites.

7 Concluding remarks and forward look

7.1 What have we learnt on ICL and IGL from studies in the local groups and clusters?

The study of the ICL and IGL in the local groups and clusters in the nearby Universe allows us a privileged view on these components and their phase space properties, which is complementary to the photometric studies carried out for clusters at higher redshifts. In the nearby Universe it is possible to measure the projected phase space–radius and velocity–for a sparse but still significant sample of individual IGL and ICL tracers. Nearby clusters are also within reach of HST, and pointed observations provide specific information on the age and metallicity distribution function of the resolved stellar populations in these clusters and groups.

The IGL and ICL surface brightness distributions in the nearby clusters are more centrally concentrated than the azimuthally averaged light of group and cluster galaxies excluding the CCG: by itself it is an indication that stars are gravitationally unbound and added to the group or cluster volume from those galaxies that plunge deeper into the group or cluster potential. The ICL and IGL fractions are a few to ten percent of the light in cluster and group galaxies. These fractions are lower than those measured in more concentrated, dynamically evolved clusters at larger distances. The ICL stars are not bound to a single galaxy, but by the cluster or group potential. The binding energy of the ICL may be in fact be a continuum depending on the whether the cluster is relaxed/compact or affected by sub-cluster mergers. From the measured CMD in deep HST images, the theoretical stellar evolution models suggest that the IC population is dominated by stars with old ages (≥10 Gyr) with a significant number of very metal-poor stars ([Fe/H] < −1.5). See also the very recent results on the color gradients in the ICL towards bluer color in intermediate redshift (0.2–0.8) clusters (Golden-Marx et al., 2022).

Both the observations and the simulations show that the ICL *in situ* star formation from stripped gas (Puchwein et al., 2010) is not an important channel, see discussion in Contini (2021). But it does happen: see the tidally stripped gas in the Virgo core (Oosterloo and van Gorkom 2005) and the isolated HII around NGC 4388 (Gerhard et al., 2002). Hence the majority of IC stars in the nearby groups and clusters were formed in galaxies and then unbound in the cluster space.

7.2 What are the likely masses of the satellite progenitors of the ICL and IGL stars?

From the arguments presented in Section 6, the picture that emerges for the local IGL/and ICL components is that their stars are mostly old and metal poor implying low mass satellite progenitors. The stellar masses of the satellites that built the IGL component around M105 or M49 would have to be a few 10^8 $M_\odot$, according to the low-mass end of the mass-metallicity relation (Kirby et al., 2013). In more massive clusters, we learn from the NGC 3311 case in the Hydra I cluster that a larger percentage of stars will come from more massive, metal richer progenitors with faster star formation time scale (high [α/Fe]).

An evolutionary path for the formation of a blue, smooth, and extended IGL around M49/M105/M87 may start with an early (before z ~ 0.5) accretion of a number of low-mass satellites with young (~1 Gyr) and blue (B− V = 0.1) stellar populations at that redshift. Since their accretion, more than 5 Gyr ago, these stars would have evolved passively and now reached a color of B-V = 0.65 at z = 0. The presence of stars that originate from low-mass systems is related to the presence of large amounts of dark matter in groups and clusters. Due to dynamical friction, low-mass systems will stay roughly at the radius at which they were disrupted, while the more massive and therefore metal-rich/redder satellites would sink towards the CCG center (e.g., Ostriker and Tremaine 1975; Amorisco 2017). We note that the contribution of such low-mass systems with small mass ratios with the host galaxy to the ICL and IGL is expected to be small in current hydrodynamical cosmological simulations. While several groups/authors are able to carry out simulations with mass resolution equivalent to mass merger ratios 10^{-4}: 1, the cumulative effect from mergers with stellar mass ratios below 10^{-2}: 1 is found to be negligible with respect to stellar halo growth (see Figure 1 in Rodriguez-Gomez et al., 2016). The observations of gravitationally unbound IGL and ICL in the nearby groups and clusters suggest instead that the IGL and ICL formed early from the accretion of many low-mass satellites. Therefore this channel which adds old metal poor stars to the IGL and ICL is implied in the nearby universe.

7.3 Observations of higher redshift clusters

The results illustrated in the previous sections on the origin of stars in the ICL refer to those stars identified because of their 1) large relative velocities which differentiate them from CCG halo stars and 2) distinct properties in the CMD. In clusters at higher redshifts/distances, measurements of the detailed kinematics and stellar population properties are challenging, hence the photometric approach is the one most often implemented observationally. For observations and characterization of the DSC in higher redshift clusters see Montes et al. (2021), (DeMaio et al., 2015; DeMaio et al., 2020) and references therein. Hence because of observational limitations, only comparatively higher surface brightness regions of the DSC are accessible in these clusters. These are then related to the stars stripped from more massive companions or liberated during massive merger events (Purcell et al., 2007; Contini et al., 2014; Contini et al., 2019).

7.4 Future outlook

As photometry becomes deeper with the LSST, or carried out in space with Euclid, it may be possible to detect the outer boundary of the ICL, i.e., the splashback/truncation radius (Gonzalez et al., 2021.) and thereby constrain the redshift of the earliest infall. An additional strong motivation to map and trace the DSC, i.e., CCG halo light, IGL and ICL, comes from the observational results showing that the (dark) matter distribution from weak lensing and the DSC from deep photometry map each other (Montes and Trujillo 2019) and the results from cosmological simulations indicating that the kinematic properties of stars and dark matter in ETG halos become similar at large radii where the accretion factions are largest (Pulsoni et al., 2021).

The new facilities ELT/JWST/Euclid will revolutionize the study of the IGL and ICL. With integral field spectroscopy on the ELT it will be possible to constrain the ages, metallicity and [α/Fe] in the Virgo core. An opportunity may come directly from the nebular phases of the stellar evolution, PNe and HII regions: from their oxygen and argon abundances (as in Arnaboldi et al., 2022 and Bhattacharya et al., 2022), one will be able to constrain the star formation efficiencies of the progenitors that released stars to the ICL. With ELT and JWST it will be possible to study further the resolved stellar populations in clusters and groups, to constrain the CMDs of the ICL stars in different regions of the nearby clusters, thus leading to a better understanding of this enigmatic component.

Author contributions

MA lead the observations and interpretations of the results. OG lead the dynamical modelling of the kinematics and mass determination.

Acknowledgments

The quest for discrete tracers in the nearby clusters have benefitted from the contribution and discussion with several colleagues. Special thanks to Ken C. Freeman, Sadanori Okamura, and Matthew Colless for advice and support at the earlier stages of this project. Many Ph.D students and Post docs (now colleagues) contributed to these investigations: J.A.L. Aguerri, C. Barbosa, S. Bhattacharya, N. Castro-Rodriguez, L. Coccato, A. Cortesi, J. Hartke, A. Longobardi, N.R. Napolitano, M. Pannella, C. Pulsoni, C. Spiniello, G. Ventimiglia. We acknowledge collaboration with A. Buzzoni, M. Capaccioli, R. Ciardullo, N. Douglas, J. Feldmeier, G. Jacoby, K. Kujken, R. Kudritzki, R. H. Méndez, M. Merrifield, J.C. Mihos, G.M. Murante, A. Romanowsky, M. Roth, N. Yasuda. Finally we thank the referees and the editor, E. Contini, for comments and suggestions that improved the content of this review.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abadi, M. G., Navarro, J. F., and Steinmetz, M. (2006). Stars beyond galaxies: The origin of extended luminous haloes around galaxies. *Mon. Not. R. Astron. Soc.* 365, 747. doi:10.1111/j.1365-2966.2005.09789.x

Adami, C., Biviano, A., Durret, F., and Mazure, A. (2005). The build-up of the Coma cluster by infalling substructures. *Astron. Astrophys.* 443, 17–27. doi:10.1051/0004-6361:20053504

Amorisco, N. C. (2017). Contributions to the accreted stellar halo: An atlas of stellar deposition. *Mon. Not. R. Astron. Soc.* 464, 2882–2895. doi:10.1093/mnras/stw2229

Arnaboldi, M., and Gerhard, O. (2010). JD2 - diffuse light in galaxy clusters. *Proc. Int. Astron. Union* 15, 97–110. doi:10.1017/S174392131000846X

Arnaboldi, M., Freeman, K. C., and Hui, X. (1994). The kinematics of the planetary nebulae in the outer regions of NGC 1399. *Messenger* 76, 40–44.

Arnaboldi, M., Freeman, K. C., Mendez, R. H., Capaccioli, M., Ciardullo, R., Ford, H., et al. (1996). The kinematics of the planetary nebulae in the outer regions of NGC 4406. *Astrophys. J.* 472, 145–152. doi:10.1086/178050

Arnaboldi, M., Freeman, K. C., Gerhard, O., Matthias, M., Kudritzki, R. P., Mendez, R. H., et al. (1998). The stellar dynamics and mass of NGC 1316 using the radial velocities of planetary nebulae. *Astrophys. J.* 507, 759–773. doi:10.1086/306359

Arnaboldi, M., Aguerri, J. A. L., Napolitano, N. R., Gerhard, O., Freeman, K. C., Feldmeier, J., et al. (2002). Intracluster planetary nebulae in Virgo: Photometric selection, spectroscopic validation, and cluster depth. *Astron. J.* 123, 760–771. doi:10.1086/338313

Arnaboldi, M., Freeman, K. C., and Okamura, S. (2003). Narrowband imaging in [O III] and hα to search for intracluster planetary nebulae in the Virgo cluster. *Astrophys. J.* 125, 514. doi:10.1086/345962

Arnaboldi, M., Gerhard, O., Aguerri, J. A. L., Freeman, K. C., Napolitano, N. R., Okamura, S., et al. (2004). The line-of-sight velocity distributions of intracluster planetary nebulae in the Virgo cluster core. *Astrophys. J.* 614, 33–L36. doi:10.1086/425417

Arnaboldi, M., Gerhard, O., Okamura, S., Kashikawa, N., Yasuda, N., and Freeman, K. C. (2007). Multi-slit imaging spectroscopy technique: Catalog of intracluster planetary nebulae in the coma cluster. *Publ. Astron. Soc. Jpn. Nihon. Tenmon. Gakkai.* 59, 419–425. doi:10.1093/pasj/59.2.419

Arnaboldi, M., Ventimiglia, G., and Iodice, E. (2012). A tale of two tails and an off-centred envelope: Diffuse light around the cD galaxy NGC 3311 in the Hydra I cluster. *A&A* 545, 37. doi:10.1051/0004-6361/20111675

Arnaboldi, M., Pulsoni, C., and Gerhard, O. (2017). The PN.S ETG survey. *Proc. Int. Astron. Union* 323, 279–283. doi:10.1017/S1743921317000771

Arnaboldi, M., Bhattacharya, S., Gerhard, O., Kobayashi, C., Freeman, K. C., Caldwell, N., et al. (2022). The survey of planetary nebulae in Andromeda (M 31) V. Chemical enrichment of the thin and thicker discs of Andromeda. Oxygen to argon abundance ratios for planetary nebulae and H II regions. *Astron. Astrophys.* 666, 109. doi:10.1051/0004-6361/202244258

Arnaboldi, M. (2005). Intracluster planetary nebulae as dynamical probes of the diffuse light in galaxy clusters. *AIP Conf. Proc.* 804, 301. doi:10.1063/1.2146298

Arnold, J. A., Romanowsky, A. J., Brodie, J. P., Forbes, D. A., Strader, J., Spitler, L. R., et al. (2014). The SLUGGS survey: Wide- field stellar kinematics of early-type galaxies. *Astrophys. J.* 791, 80. doi:10.1088/0004-637X/791/2/80

Baes, M., Sil'chenko, O. K., Moiseev, A. V., and Manakova, E. A. (2007). Metallicity and age gradients in round elliptical galaxies. *Astron. Astrophys.* 467, 991–1001. doi:10.1051/0004-6361:20066758

Barbosa, C. E., Arnaboldi, M., Coccato, L., Hilker, M., Mendes de Oliveira, C., and Richtler, T. (2016). The Hydra I cluster core. *Astron. Astrophys.* 589, 139. doi:10.1051/0004-6361/201628137

Barbosa, C. E., Arnaboldi, M., Coccato, L., Gerhard, O., Mendes de Oliveira, C., Hilker, M., et al. (2018). Sloshing in its cD halo: MUSE kinematics of the central galaxy NGC 3311 in the Hydra I cluster. *Astron. Astrophys.* 609, 78. doi:10.1051/0004-6361/201731834

Barbosa, C. E., Spiniello, C., Arnaboldi, M., Coccato, L., Hilker, M., and Richtler, T. (2021). A preserved high-z compact progenitor in the heart of NGC 3311 revealed with MUSE 2D stellar population analysis. *Astron. Astrophys.* 649, 93. doi:10.1051/0004-6361/202039809

Bellstedt, S., Forbes, D. A., Foster, C., Romanowsky, A. J., Brodie, J. P., Pastorello, N., et al. (2017). The SLUGGS survey: Using extended stellar kinematics to disentangle the formation histories of low-mass S0 galaxies. *Mon. Not. R. Astron. Soc.* 467, 4540–4557. doi:10.1093/mnras/stx418

Bender, R., Saglia, R. P., and Gerhard, O. E. (1994). Line-of-sight velocity distributions of elliptical galaxies. *Mon. Not. R. Astron. Soc.* 269, 785–813. doi:10.1093/mnras/269.3.785

Bender, R., Kormendy, J., and Cornell, M. E. (2015). Structure and formation of cD galaxies: NGC 6166 in ABELL 2199. *Astrophys. J.* 807, 56. doi:10.1088/0004-637X/807/1/56

Bernstein, G. M., Nichol, R. C., and Tyson, J. A. (1995). The luminosity function of the coma cluster core for -25<M/R<-9.4. *Astrophys. J.* 110, 1507. doi:10.1086/117624

Bhattacharya, S., Arnaboldi, M., Caldwell, N., Gerhard, O., Kobayashi, C., Hartke, J., et al. (2022). The survey of planetary nebulae in Andromeda (M31) – IV. Radial oxygen and argon abundance gradients of the thin and thicker disc. *Mon. Not. R. Astron. Soc.* 517, 2343–2359. doi:10.1093/mnras/stac2703

Binggeli, B., Tammann, G. A., and Sandage, A. (1987). Studies of the Virgo cluster. VI. Morphological and kinematical structure of the Virgo cluster. *Astron. J.* 94, 251. doi:10.1086/114467

Binggeli, B., Popescu, C. C., and Tammann, G. A. (1993). The kinematics of the Virgo cluster revisited. *Astronomy Astrophysics, Suppl. Ser.* 98, 275.

Boardman, N. F., Weijmans, A.-M., van den Bosch, R., Kuntschner, H., Emsellem, E., Cappellari, M., et al. (2017). Integral-field kinematics and stellar populations of early-type galaxies out to three half-light radii. *Mon. Not. R. Astron. Soc.* 471, 4005–4026. doi:10.1093/mnras/stx1835

Böhringer, H., Briel, U. G., Schwarz, R. A., Voges, W., Hartner, G., and Trumper, J. (1994). The structure of the Virgo cluster of galaxies from Rosat X-ray images. *Nature* 368, 828–831. doi:10.1038/368828a0

Buzzoni, A., Arnaboldi, M., and Corradi, R. L. M. (2006). Planetary nebulae as tracers of galaxy stellar populations. *Mon. Not. R. Astron. Soc.* 368, 877–894. doi:10.1111/j.1365-2966.2006.10163.x

Cañas, R., Lagos, C., del, P., Elahi, P. J., Welker, C., Dubois, Y., et al. (2020). From stellar haloes to intracluster light: The physics of the intra-halo stellar component in cosmological hydrodynamical simulations. *Mon. Not. R. Astron. Soc.* 494, 4314–4333. doi:10.1093/mnras/staa1027

Capaccioli, M., Spavone, M., Grado, A., et al. (2015). VEGAS: A VST early-type GAlaxy Survey. I. Presentation, wide-field surface photometry, and substructures in NGC 4472. *A&A* 581, 10. doi:10.1051/0004-6361/201526252

Castro-Rodriguéz, N., Aguerri, J. A. L., Arnaboldi, M., Gerhard, O., Freeman, K. C., Napolitano, N. R., et al. (2003). Narrow band survey for intragroup light in the Leo HI cloud. Constraints on the galaxy background contamination in imaging surveys for intracluster planetary nebulae. *Astron. Astrophys.* 405, 803–812. doi:10.1051/0004-6361:20030588

Castro-Rodriguéz, N., Arnaboldi, M., and Aguerri, J. A. L. (2009). Intracluster light in the Virgo cluster: Large scale distribution. *A&A* 507, 621. doi:10.1051/0004-6361/200809774

Chaturvedi, A., Hilker, M., and Cantiello, M. (2022). The Fornax Cluster VLT Spectroscopic Survey. III. Kinematical characterisation of globular clusters in the Fornax galaxy cluster. *A&A* 657, 93. doi:10.1051/0004-6361/202141334

Churazov, E., Tremaine, S., Forman, W., Gerhard, O., Das, P., Vikhlinin, A., et al. (2010). Comparison of approximately isothermal gravitational potentials of elliptical galaxies based on X-ray and optical data. *Mon. Not. R. Astron. Soc.* 404, 1165. doi:10.1111/j.1365-2966.2010.16377.x

Ciardullo, R., Jacoby, G. H., Ford, H. C., and Neill, J. D. (1989). Planetary nebulae as standard candles. II. The calibration in M31 and its companions. *Astrophys. J.* 339, 53. doi:10.1086/167275

Ciardullo, R., Durrell, P., Laychak, M. B., Herrmann, K. A., Moody, K., Jacoby, G. H., et al. (2004). The planetary nebula system of M33. *Astrophys. J.* 614, 167–185. doi:10.1086/423414

Coccato, L., Gerhard, O., Arnaboldi, M., Das, P., Douglas, N. G., Kuijken, K., et al. (2009). Kinematic properties of early-type galaxy haloes using planetary nebulae. *Mon. Not. R. Astron. Soc.* 394, 1249–1283. doi:10.1111/j.1365-2966.2009.14417.x

Coccato, L., Gerhard, O., and Arnaboldi, M. (2010). Distinct core and halo stellar populations and the formation history of the bright Coma cluster early-type galaxy NGC 4889. *Mon. Not. R. Astron. Soc.* 407, 26–L30. doi:10.1111/j.1745-3933.2010.00897.x

Coccato, L., Gerhard, O., Arnaboldi, M., and Ventimiglia, G. (2011). Stellar population and the origin of intra-cluster stars around brightest cluster galaxies: The case of NGC 3311. *Astron. Astrophys.* 533, 138. doi:10.1051/0004-6361/201117546

Coccato, L., Arnaboldi, M., and Gerhard, O. (2013). Signatures of accretion events in the haloes of early- type galaxies from comparing PNe and GCs kinematics. *Mon. Not. R. Astron. Soc.* 436, 1322–1334. doi:10.1093/mnras/stt1649

Colless, M., and Dunn, A. M. (1996). Structure and dynamics of the coma cluster. *Astrophys. J.* 458, 435. doi:10.1086/176827

Contini, E., and Gu, Q. (2020). On the mass distribution of the intracluster light in galaxy groups and clusters. *Astrophys. J.* 901, 128. doi:10.3847/1538-4357/abb1aa

Contini, E., De Lucia, G., Villalobos, Á., and Borgani, S. (2014). On the formation and physical properties of the intracluster light in hierarchical galaxy formation models. *Mon. Not. R. Astron. Soc.* 437, 3787–3802. doi:10.1093/mnras/stt2174

Contini, E., Yi, S. K., and Kang, X. (2019). Theoretical predictions of colors and metallicity of the intracluster light. *Astrophys. J.* 871, 24. doi:10.3847/1538-4357/aaf41f

Contini, E. (2021). On the origin and evolution of the intra-cluster light: A brief review of the most recent developments. *Galax* 9, 60C. doi:10.3390/galaxies9030060

Cooper, A. P., D'Souza, R., Kauffmann, G., Wang, J., Boylan-Kolchin, M., Guo, Q., et al. (2013). Galactic accretion and the outer structure of galaxies in the CDM model. *Mon. Not. R. Astron. Soc.* 434, 3348–3367. doi:10.1093/mnras/stt1245

Cooper, A. P., Gao, L., and Guo, Q. (2015). Surface photometry of brightest cluster galaxies and intracluster stars in ΛCDM. *Mon. Not. R. Astron. Soc.* 451, 2703. doi:10.1093/mnras/stv1042

Cortesi, A., Arnaboldi, M., Coccato, L., Merrifield, M. R., Gerhard, O., Bamford, S., et al. (2013). The Planetary Nebula Spectrograph survey of S0 galaxy kinematics. *Astron. Astrophys.* 549, 115. doi:10.1051/0004-6361/201220306

Cui, W., Murante, G., Monaco, P., Borgani, S., Granato, G. L., Killedar, M., et al. (2014). Characterizing diffused stellar light in simulated galaxy clusters. *Mon. Not. R. Astron. Soc.* 437, 816–830. doi:10.1093/mnras/stt1940

Das, P., Gerhard, O., Churazov, E., and Zhuravleva, I. (2010). Steepening mass profiles, dark matter and environment of X-ray bright elliptical galaxies. *Mon. Not. R. Astron. Soc.* 409, 1362–1378. doi:10.1111/j.1365-2966.2010.17417.x

de Lorenzi, F., Gerhard, O., Coccato, L., Arnaboldi, M., Capaccioli, M., Douglas, N. G., et al. (2009). Dearth of dark matter or massive dark halo? Mass-shape-anisotropy degeneracies revealed by NMAGIC dynamical models of the elliptical galaxy NGC 3379. *Mon. Not. R. Astron. Soc.* 395, 76–96. doi:10.1111/j.1365-2966.2009.14553.x

de Vaucouleurs, G. (1948). Recherches sur les Nebuleuses Extragalactiques. *Ann. d'Astrophysique* 11, 247.

DeMaio, T., Gonzalez, A. H., and Zabludoff, A. (2015). On the origin of the intracluster light in massive galaxy clusters. *Mon. Not. R. Astron. Soc.* 448, 1162. doi:10.1093/mnras/stv033

DeMaio, T., Gonzalez, A. H., and Zabludoff, A. (2020). the growth of brightest cluster galaxies and intracluster light over the past 10 billion years. *Mon. Not. R. Astron. Soc.* 491, 3751. doi:10.1093/mnras/stz3236

Doherty, M., Arnaboldi, M., Das, P., Gerhard, O., Aguerri, J. A. L., Ciardullo, R., et al. (2009). The edge of the M 87 halo and the kinematics of the diffuse light in the Virgo cluster core. *Astron. Astrophys.* 502, 771–786. doi:10.1051/0004-6361/200811532

Dolag, K., Murante, G., and Borgani, S. (2010). Dynamical difference between the cD galaxy and the diffuse, stellar component in simulated galaxy clusters. *Mon. Not. R. Astron. Soc.* 405, 1544. doi:10.1111/j.1365-2966.2010.16583.x

Dotter, A., Chaboyer, B., Jevremović, D., Kostov, V., Baron, E., and Ferguson, J. W. (2008). The Dartmouth stellar evolution database. *Astrophys. J. Suppl. Ser.* 178, 89–101. doi:10.1086/589654

Douglas, N. G., Napolitano, N. R., Romanowsky, A. J., Coccato, L., Kuijken, K., Merrifield, M. R., et al. (2007). The PN.S elliptical galaxy survey: Data reduction, planetary nebula catalog, and basic dynamics for NGC 3379. *Astrophys. J.* 664, 257–276. doi:10.1086/518358

Dressler, A. (1979). The dynamics and structure of the cD galaxy in Abell 2029. *Astrophys. J.* 231, 659. doi:10.1086/157229

D'Souza, R., Kauffman, G., Wang, J., and Vegetti, S. (2014). Parametrizing the stellar haloes of galaxies. *Mon. Not. R. Astron. Soc.* 443, 1433–1450. doi:10.1093/mnras/stu1194

Durrell, P. R., Ciardullo, R., Feldmeier, J. J., Jacoby, G. H., and Sigurdsson, S. (2002). Intracluster red giant stars in the Virgo cluster. *Astrophys. J.* 570, 119–131. doi:10.1086/339735

Eggen, O. J., Lynden-Bell, D., and Sandage, A. R. (1962). Evidence from the motions of old stars that the Galaxy collapsed. *Astrophys. J.* 136, 748. doi:10.1086/147433

Emsellem, E., Cappellari, M., Peletier, R. F., McDermid, R. M., Bacon, R., Bureau, M., et al. (2004). The SAURON project - III. Integral-field absorption-line kinematics of 48 elliptical and lenticular galaxies. *Mon. Not. R. Astron. Soc.* 352, 721–743. doi:10.1111/j.1365-2966.2004.07948.x

Feldmeier, J. J., Ciardullo, R., and Jacoby, G. H. (1998). Intracluster planetary nebulae in the Virgo cluster. I. Initial results. *Astrophys. J.* 503, 109–117. doi:10.1086/305981

Feldmeier, J. J., Ciardullo, R., Jacoby, G. H., and Durrell, P. R. (2004). Intracluster planetary nebulae in the Virgo cluster. III. Luminosity of the intracluster light and tests of the spatial distribution. *Astrophys. J.* 615, 196–208. doi:10.1086/424372

Ferguson, H. C., Tanvir, N. R., and von Hippel, T. (1998). Detection of intergalactic red-giant-branch stars in the Virgo cluster. *Nature* 391, 461–463. doi:10.1038/35087

Fitchett, M., and Webster, R. (1987). Substructure in the coma cluster. *Astrophys. J.* 317, 653. doi:10.1086/165310

Foster, C., Pastorello, N., Roediger, J., Brodie, J. P., Forbes, D. A., Kartha, S. S., et al. (2016). The SLUGGS survey: Stellar kinematics, kinemetry and trends at large radii in 25 early-type galaxies. *Mon. Not. R. Astron. Soc.* 457, 147–171. doi:10.1093/mnras/stv2947

Gerhard, O., Jeske, G., Saglia, R. P., and Bender, R. (1998). Breaking the degeneracy between anisotropy and mass: The dark halo of the E0 galaxy NGC6703. *Mon. Not. R. Astron. Soc.* 295, 197–215. doi:10.1046/j.1365-8711.1998.29511341.x

Gerhard, O., Kronawitter, A., and Saglia, R. P. (2001). Dynamical family properties and dark halo scaling relations of giant elliptical galaxies. *Astrophys. J.* 121, 1936. doi:10.1086/319940

Gerhard, O., Arnaboldi, M., Freeman, K. C., and Okamura, S. (2002). Isolated star formation: A compact H [CSC]ii[/CSC] region in the Virgo cluster. *Astrophys. J.* 580, 121–L124. doi:10.1086/345657

Gerhard, O., Arnaboldi, M., Freeman, K. C., Kashikawa, N., Okamura, S., and Yasuda, N. (2005). Detection of intracluster planetary nebulae in the coma cluster. *Astrophys. J.* 621, 93–L96. doi:10.1086/429221

Gerhard, O., Arnaboldi, M., Freeman, K. C., Okamura, S., Kashikawa, N., and Yasuda, N. (2007). The kinematics of intracluster planetary nebulae and the on-going subcluster merger in the Coma cluster core. *Astron. Astrophys.* 468, 815–822. doi:10.1051/0004-6361:20066484

Gerhard, O. E. (1993). Line-of-sight velocity profiles in spherical galaxies: Breaking the degeneracy between anisotropy and mass. *Mon. Not. R. Astron. Soc.* 265, 213–230. doi:10.1093/mnras/265.1.213

Golden-Marx, J. B., Zhang, Y., and Ogando, R. L. C. (2022). Characterising the intracluster light over the redshift range 0.2<z<0.8 in the DES-ACT overlap. 2022arXiv220905519G

Gonzalez, A. H., Zabludoff, A. I., and Zaritsky, D. (2005). Intracluster light in nearby galaxy clusters: Relationship to the halos of brightest cluster galaxies. *Astrophys. J.* 618, 195. doi:10.1086/425896

Gonzalez, A. H., George, T., Connor, T., Deason, A., Donahue, M., Montes, M., et al. (2021). Discovery of a possible splashback feature in the intracluster light of MACS J1149.5+2223. *Mon. Not. R. Astron. Soc.* 507, 963–970. doi:10.1093/mnras/stab2117

Greene, J. E., Murphy, J. D., Graves, G. J., Gunn, J. E., Raskutti, S., Comerford, J. M., et al. (2013). The stellar halos of massive elliptical galaxies. II. Detailed abundance ratios at large radius. *Astrophys. J.* 776, 64. doi:10.1088/0004-637X/776/2/64

Harris, W. E., Harris, G. L. H., Layden, A. C., and Wehner, E. M. H. (2007). The Leo elliptical NGC 3379: A metal- poor halo emerges. *Astrophys. J.* 666, 903–918. doi:10.1086/520799

Hartke, J., Arnaboldi, M., Longobardi, A., Gerhard, O., Freeman, K. C., Okamura, S., et al. (2017). The halo of M 49 and its environment as traced by planetary nebulae populations. *Astron. Astrophys.* 603, 104. doi:10.1051/0004-6361/201730463

Hartke, J., Arnaboldi, M., Gerhard, O., Agnello, A., Longobardi, A., Coccato, L., et al. (2018). Three dynamically distinct stellar populations in the halo of M49. *Astron. Astrophys.* 616, 123. doi:10.1051/0004-6361/201832711

Hartke, J., Arnaboldi, M., and Gerhard, O. (2020). The halo of M 105 and its group environment as traced by planetary nebula populations. I. Wide-field photometric survey of planetary nebulae in the Leo I group. *A&A* 642, 46. doi:10.1051/0004-6361/202038009

Hartke, J., Arnaboldi, M., Gerhard, O., Coccato, L., Merrifield, M., Kuijken, K., et al. (2022). The halo of M105 and its group environment as traced by planetary nebula populations: II. Using kinematics of single stars to unveil the presence of intragroup light around the Leo I galaxies NGC 3384 and M105. *Astron. Astrophys.* 663, 12. doi:10.1051/0004-6361/202243117

Hilker, M., Richtler, T., Barbosa, C. E., Arnaboldi, M., Coccato, L., and Mendes de Oliveira, C. (2018). The Hydra I cluster core. *Astron. Astrophys.* 619, 70. doi:10.1051/0004-6361/201731737

Hilz, M., Naab, T., Ostriker, J. P., Thomas, J., Burkert, A., and Jesseit, R. (2012). Relaxation and stripping - the evolution of sizes, dispersions and dark matter fractions in major and minor mergers of elliptical galaxies. *Mon. Not. R. Astron. Soc.* 425, 3119–3136. doi:10.1111/j.1365-2966.2012.21541.x

Hui, X., Ford, H. C., Freeman, K. C., and Dopita, M. A. (1995). The planetary nebula system and dynamics of NGC 5128. III. Kinematics and halo mass distributions. *Astrophys. J.* 449, 592. doi:10.1086/176082

Iodice, E., Capaccioli, M., and Grado, A. (2016). The Fornax deep survey with VST. I. The extended and diffuse stellar halo of NGC 1399 out to 192 kpc. *Astrophys. J.* 820,42. doi:10.3847/0004-637X/820/1/42

Jacoby, G. H., and De Marco, O. (2002). A survey for very faint planetary nebulae in the SMC. I. Identification, confirmation, and preliminary analysis. *Astron. J.* 123, 269–278. doi:10.1086/324737

Kapferer, W., Schindler, S., Knollmann, S. R., and van Kampen, E. (2010). The evolution of an intra-cluster and intra-group stellar population. Their contribution to the stellar mass, their age, and their dynamics. *Astron. Astrophys.* 516, 41. doi:10.1051/0004-6361/200913563

Kelson, D. D., Zabludoff, A. I., Williams, K. A., Trager, S. C., Mulchaey, J. S., and Bolte, M. (2002). Determination of the dark matter profile of A2199 from integrated starlight. *Astrophys. J.* 576, 720–737. doi:10.1086/341891

Kluge, M., Neureiter, B., Riffeser, A., Bender, R., Goessl, C., Hopp, U., et al. (2020). Structure of brightest cluster galaxies and intracluster light. *Astrophys. J. Suppl. Ser.* 247, 43. doi:10.3847/1538-4365/ab733b

Kluge, M., Bender, R., and Riffeser, A. (2021). Photometric dissection of intracluster light and its correlations with host cluster properties. *Astrophys. J. Suppl. Ser.* 252, 27. doi:10.3847/1538-4365/abcda6

Koopmans, L. V. E., Bolton, A., Treu, T., Czoske, O., Auger, M. W., Barnabe, M., et al. (2009). The structure and dynamics of massive Early-Type Galaxies: On homology, isothermality, and isotropy inside one effective radius. *Astrophys. J.* 703, 51–L54. doi:10.1088/0004-637X/703/1/L51

Kormendy, J., and Bender, R. (2012). A revised parallel-sequence morphological classification of galaxies: Structure and formation of S0 and spheroidal galaxies. *Astrophys. J. Suppl. Ser.* 198, 2. doi:10.1088/0067-0049/198/1/2

Kormendy, J., Fisher, D. B., Cornell, M. E., and Bender, R. (2009). Structure and formation of elliptical and spheroidal galaxies. *Astrophys. J. Suppl. Ser.* 182, 216–309. doi:10.1088/0067-0049/182/1/216

Krick, J. E., and Bernstein, R. A. (2007). Diffuse optical light in galaxy clusters. II. Correlations with cluster properties. *Astrophys. J.* 134, 466. doi:10.1086/518787

Kuntschner, H., Emsellem, E., Bacon, R., Cappellari, M., Davies, R. L., De Zeeuw, P. T., et al. (2010). The SAURON project - XVII. Stellar population analysis of the absorption line strength maps of 48 early-type galaxies. *Mon. Not. R. Astron. Soc.* 408, 97–132. doi:10.1111/j.1365-2966.2010.17161.x

La Barbera, F., Ferreras, I., de Carvalho, R. R., Bruzual, G., Charlot, S., Pasquali, A., et al. (2012). Spider - VII. Revealing the stellar population content of massive early-type galaxies out to 8Re. *Mon. Not. R. Astron. Soc.* 426, 2300–2317. doi:10.1111/j.1365-2966.2012.21848.x

La Marca, A., Peletier, R., Iodice, E., Paolillo, M., Choque Challapa, N., Venhola, A., et al. (2022a). Galaxy populations in the Hydra I cluster from the VEGAS survey. I. Optical properties of a large sample of dwarf galaxies. *Astron. Astrophys.* 659, 92. doi:10.1051/0004-6361/202141901

La Marca, A., Iodice, E., Cantiello, M., Forbes, D. A., Rejkuba, M., Hilker, M., et al. (2022b). Galaxy populations in the Hydra I cluster from the VEGAS survey. II. The ultra-diffuse galaxy population. *Astron. Astrophys.* 665, 105. doi:10.1051/0004-6361/202142367

Lee, M. G., and Jang, I. S. (2016). Dual stellar halos in the standard elliptical galaxy M105 and formation of massive early-type galaxies. *ApJ* 822, 70. doi:10.3847/0004-637X/822/2/70

Li, R., Li, H., Shao, S., Lu, S., Zhu, K., Wang, C., et al. (2019). SDSS-IV MaNGA: The inner density slopes of nearby galaxies. *Mon. Not. R. Astron. Soc.* 490, 2124–2138. doi:10.1093/mnras/stz2565

Longobardi, A., Arnaboldi, M., Gerhard, O., Coccato, L., Okamura, S., and Freeman, K. C. (2013). The planetary nebula population in the halo of M 87. *Astron. Astrophys.* 558, 42. doi:10.1051/0004-6361/201321652

Longobardi, A., Arnaboldi, M., and Gerhard, O. (2015a). The outer regions of the giant Virgo galaxy M 87 Kinematic separation of stellar halo and intracluster light. *A&A* 579, 135. doi:10.1051/0004-6361/201525773

Longobardi, A., Arnaboldi, M., Gerhard, O., and Mihos, J. C. (2015b). The build-up of the cD halo of M 87: Evidence for accretion in the last gyr. *Astron. Astrophys.* 579, 3. doi:10.1051/0004-6361/201526282

Longobardi, A., Arnaboldi, M., Gerhard, O., Pulsoni, C., and Soldner-Rembold, I. (2018a). Kinematics of the outer halo of M87 as mapped by planetary nebulae. *Astron. Astrophys.* 620, 111. doi:10.1051/0004-6361/201832729

Longobardi, A., Peng, E. W., Côté, P., Mihos, J. C., Ferrarese, L., Puzia, T. H., et al. (2018b). The next generation Virgo cluster survey (NGVS). XXXI. The kinematics of intracluster globular clusters in the core of the Virgo cluster. *Astrophys. J.* 864, 36. doi:10.3847/1538-4357/aad3d2

Loubser, S. I., Lagos, P., Babul, A., et al. (2022). Merger histories of brightest group galaxies from MUSE stellar kinematics. *MNRAS* 515, 1104. doi:10.1093/mnras/stac1781

McNeil, E. K., Arnaboldi, M., Freeman, K. C., Gerhard, O. E., Coccato, L., and Das, P. (2010). Counter-dispersed slitless-spectroscopy technique: Planetary nebula velocities in the halo of NGC 1399. *Astron. Astrophys.* 518, 44. doi:10.1051/0004-6361/201014428

McNeil-Moylan, E. K., Freeman, K. C., Arnaboldi, M., and Gerhard, O. E. (2012). Planetary nebula kinematics in NGC 1316: A young sombrero. *Astron. Astrophys.* 539, 11. doi:10.1051/0004-6361/201117875

Melnick, J., White, S. D. M., and Hoessel, J. (1977). Photoelectric surface photometry of the Coma cluster. *Mon. Notices R. Astronomical Soc.* 180, 207. doi:10.1093/mnras/180.2.207

Méndez, R. H., Riffeser, A., Kudritzki, R. -P., Matthias, M., Freeman, K. C., Arnaboldi, M., et al. (2001). Detection, photometry, and slitless radial velocities of 535 planetary nebulae in the flattened elliptical galaxy NGC 4697. *Astrophys. J.* 563, 135–150. doi:10.1086/323794

Mihos, J. C., Harding, P., Feldmeier, J. J., and Morrison, H. (2005). Diffuse light in the Virgo cluster. *Astrophys. J.* 631, 41–L44. doi:10.1086/497030

Mihos, J. C., Harding, P., Rudick, C. S., et al. (2013). Stellar populations in the outer halo of the massive elliptical M49. *ApJL* 764, 20. doi:10.1088/2041-8205/764/2/L20

Mihos, J. C., Harding, P., Feldmeier, J. J., et al. (2017). The Burrell schmidt deep virgo survey: Tidal debris, galaxy halos, and diffuse intracluster light in the virgo cluster. *ApJ* 16. doi:10.3847/1538-4357/834/1/1

Montes, M., and Trujillo, I. (2019). Intracluster light: A luminous tracer for dark matter in clusters of galaxies. *Mon. Not. R. Astron. Soc.* 482, 2838–2851. doi:10.1093/mnras/sty2858

Montes, M., Brough, S., Owers, M. S., and Santucci, G. (2021). The buildup of the intracluster light of A85 as seen by subaru's hyper suprime-cam. *Astrophys. J.* 910, 45. doi:10.3847/1538-4357/abddb6

Montes, M. (2022). The faint light in groups and clusters of galaxies. *Nat. Astron.* 6, 308–316. doi:10.1038/s41550-022-01616-z

Morganti, L., Gerhard, O., Coccato, L., Martinez-Valpuesta, I., and Arnaboldi, M. (2013). Elliptical galaxies with rapidly decreasing velocity dispersion profiles: NMAGIC models and dark halo parameter estimates for NGC 4494. *Mon. Not. R. Astron. Soc.* 431, 3570–3588. doi:10.1093/mnras/stt442

Murante, G., Arnaboldi, M., Gerhard, O., Borgani, S., Cheng, L. M., Diaferio, A., et al. (2004). The diffuse light in simulations of galaxy clusters. *Astrophys. J.* 607, 83–L86. doi:10.1086/421348

Murante, G., Giovalli, M., and Gerhard, O. (2007). The importance of mergers for the origin of intracluster stars in cosmological simulations of galaxy clusters. *Mon. Notices R. Astronomical Soc.* 377, 2. doi:10.1111/j.1365-2966.2007.11568.x

Murphy, J. D., Gebhardt, K., and Adams, J. J. (2011). Galaxy kinematics with VIRUS-P: The dark matter halo of M87. *Astrophys. J.* 729, 129. doi:10.1088/0004-637X/729/2/129

Murphy, J. D., Gebhardt, K., and Cradit, M. (2014). The rising stellar velocity dispersion of M87 from integrated starlight. *Astrophys. J.* 785, 143. doi:10.1088/0004-637X/785/2/143

Napolitano, N. R., Pannella, M., Arnaboldi, M., Gerhard, O., Aguerri, J. A. L., Freeman, K. C., et al. (2003). Intracluster stellar population properties from N-body cosmological simulations. I. Constraints at z = 0. *Astrophys. J.* 594, 172–185. doi:10.1086/376860

Napolitano, N. R., Romanowsky, A. J., Coccato, L., Capaccioli, M., Douglas, N. G., Noordermeer, E., et al. (2009). The planetary nebula spectrograph elliptical galaxy survey: The dark matter in NGC 4494. *Mon. Not. R. Astron. Soc.* 393, 329–353. doi:10.1111/j.1365-2966.2008.14053.x

Napolitano, N. R., Gatto, M., Spiniello, C., Cantiello, M., Hilker, M., Arnaboldi, M., et al. (2022). The Fornax cluster VLT spectroscopic survey. IV. Cold kinematical substructures in the Fornax core from COSTA. *Astron. Astrophys.* 657, 94. doi:10.1051/0004-6361/202141872

Neumann, D. M., Arnaud, M., Gastaud, R., Aghanim, N., Lumb, D., Briel, U. G., et al. (2003). The NGC 4839 group falling into the Coma cluster observed by XMM-Newton. *Astron. Astrophys.* 400, L74–L79. doi:10.1051/0004-6361:20000182

Okabe, N., Okura, Y., and Futamase, T. (2010). Weak-lensing mass measurements of substructures in coma cluster with subaru/suprime-cam. *Astrophys. J.* 713, 291–303. doi:10.1088/0004-637X/713/1/291

Okamura, S., Yasuda, M., Arnaboldi, M., Freeman, K. C., Ando, H., Doi, M., et al. (2002). Candidates for intracluster planetary nebulae in the Virgo cluster based on the suprime-cam narrow-band imaging in [O III] and hα *Publ. Astron. Soc. Jpn. Nihon. Tenmon. Gakkai.* 54, 883–889. doi:10.1093/pasj/54.6.883

Oosterloo, T., and van Gorkom, J. (2005). A large H I cloud near the centre of the Virgo cluster. *Astron. Astrophys.* 437, 19–L22. doi:10.1051/0004-6361:200500127

Oser, L., Ostriker, J. P., and Naab, T. (2010). The two phases of galaxy formation. *Astrophys. J.* 725, 2312. doi:10.1088/0004-637X/725/2/2312

Ostriker, J. P., and Tremaine, S. D. (1975). Another evolutionary correction to the luminosity of giant galaxies. *Astrophys. J.* 202, 113. doi:10.1086/181992

Pillepich, A., Nelson, D., Hernquist, L., Springel, V., Pakmor, R., Torrey, P., et al. (2018). First results from the IllustrisTNG simulations: The stellar mass content of groups and clusters of galaxies. *Mon. Not. R. Astron. Soc.* 475, 648–675. doi:10.1093/mnras/stx3112

Poci, A., Cappellari, M., and McDermid, R. M. (2017). Systematic trends in total-mass profiles from dynamical models of early-type galaxies. *Mon. Not. R. Astron. Soc.* 467, stx101. doi:10.1093/mnras/stx101

Poliakov, D., Mosenkov, A. V., Brosch, N. al., Koriski, S., and Rich, R. M. (2021). Quantified diffuse light in compact groups of galaxies. *Mon. Not. R. Astron. Soc.* 503, 6059–6077. doi:10.1093/mnras/stab853

Pota, V., Napolitano, N. R., Hilker, M., Spavone, M., Schulz, C., Cantiello, M., et al. (2018). The Fornax Cluster VLT Spectroscopic Survey - I. VIMOS spectroscopy of compact stellar systems in the Fornax core region. *Mon. Not. R. Astron. Soc.* 481, 1744–1756. doi:10.1093/mnras/sty2149

Puchwein, E., Springel, V., Sijacki, D., and Dolag, K. (2010). Intracluster stars in simulations with active galactic nucleus feedback. *Mon. Not. R. Astron. Soc.* 406, 936. doi:10.1111/j.1365-2966.2010.16786.x

Pulsoni, C., Gerhard, O., Arnaboldi, M., Coccato, L., Longobardi, A., Napolitano, N. R., et al. (2018). The extended Planetary Nebula Spectrograph (ePN.S) early-type galaxy survey: The kinematic diversity of stellar halos and the relation between halo transition scale and stellar mass. *Astron. Astrophys.* 618, 94. doi:10.1051/0004-6361/201732473

Pulsoni, C., Gerhard, O., Arnaboldi, M., Pillepich, A., Nelson, D., Hernquist, L., et al. (2020). The stellar halos of ETGs in the IllustrisTNG simulations: The photometric and kinematic diversity of galaxies at large radii. *Astron. Astrophys.* 641, 60. doi:10.1051/0004-6361/202038253

Pulsoni, C., Gerhard, O., Arnaboldi, M., Pillepich, A., Rodriguez-Gomez, V., Nelson, D., et al. (2021). The stellar halos of ETGs in the Illustris TNG simulations. II. Accretion, merger history, and dark halo connection. *Astron. Astrophys.* 647, 95. doi:10.1051/0004-6361/202039166

Purcell, C. W., Bullock, J. S., and Zentner, A. R. (2007). Shredded galaxies as the source of diffuse intrahalo light on varying scales. *ApJ* 666, 20. doi:10.1086/519787

Ragusa, R., Mirabile, M., Spavone, M., et al. (2022). The Intra-Group Baryons in the LEO I Pair From the VST Early-Type GAlaxy Survey. *Front. Astron. Space Sci.* 9 (852810). doi:10.3389/fspas.2022.852810

Raskutti, S., Greene, J. E., and Murphy, J. D. (2014). The stellar halos of massive elliptical galaxies. III. Kinematics at large radius. *Astrophys. J.* 786, 23. doi:10.1088/0004-637x/786/1/23

Reid, W. A., and Parker, Q. A. (2010). A new population of planetary nebulae discovered in the Large Magellanic Cloud - III. The luminosity function. *Mon. Not. R. Astron. Soc.* 405, 1349. doi:10.1111/j.1365-2966.2010.16635.x

Rix, H.-W., de Zeeuw, P. T., Cretton, N., van der Marel, R. P., and Carollo, C. M. (1997). Dynamical modeling of velocity profiles: The dark halo around the elliptical galaxy NGC 2434. *Astrophys. J.* 488, 702–719. doi:10.1086/304733

Romanowsky, A. J., Douglas, N. G., Arnaboldi, M., Kuijken, K., Merrifield, M. R., Napolitano, N. R., et al. (2003). A dearth of dark matter in ordinary elliptical galaxies. *Science* 3015640, 1696–1698. doi:10.1126/science.1087441

Romanowsky, A. J., Strader, J., Brodie, J. P., Mihos, J. C., Spitler, L. R., Forbes, D. A., et al. (2012). The ongoing assembly of a central cluster galaxy: Phase-space substructures in the halo of M87. *Astrophys. J.* 748, 29. doi:10.1088/0004-637X/748/1/29

Rodriguez-Gomez, V., Pillepich, A., Sales, L., et al. (2016). The stellar mass assembly of galaxies in the Illustris simulation: growth by mergers and the spatial distribution of accreted stars. *MNRAS* 458, 2371. doi:10.1093/mnras/stw456

Rudick, C. S., Mihos, J. C., and McBride, C. (2006). The Formation and evolution of intracluster light. *Astrophys. J.* 648, 936–946. doi:10.1086/506176

Saglia, R. P., Kronawitter, A., Gerhard, O., and Bender, R. (2000). The orbital structure and potential of NGC 1399. *Astron. J.* 119, 153–161. doi:10.1086/301153

Schombert, J. M. (1986). The structure of brightest cluster members. I. Surface photometry. *Astrophys. J. Suppl. Ser.* 60, 603. doi:10.1086/191100

Schuberth, Y., Richtler, T., Hilker, M., Dirsch, B., Bassino, L. P., Romanowsky, A. J., et al. (2010). The globular cluster system of NGC 1399. *Astron. Astrophys.* 513, 52. doi:10.1051/0004-6361/200912482

Sérsic, J. L. (1963). Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Bol. Asoc. Argent. Astron.* 6, 41–43.

Simionescu, A., Werner, N., Mantz, A., Allen, S. W., and Urban, O. (2017). Witnessing the growth of the nearest galaxy cluster: Thermodynamics of the Virgo cluster outskirts. *Mon. Not. R. Astron. Soc.* 469, 1476–1495. doi:10.1093/mnras/stx919

Spavone, M., Capaccioli, M., Napolitano, N. R., Iodice, E., Grado, A., Limatola, L., et al. (2017). Vegas: A VST early-type GAlaxy survey. II. Photometric study of giant ellipticals and their stellar halos. *Astron. Astrophys.* 603, 38. doi:10.1051/0004-6361/201629111

Spiniello, C., Napolitano, N. R., Arnaboldi, M., Tortora, C., Coccato, L., Capaccioli, M., et al. (2018). The Fornax cluster VLT spectroscopic survey II - planetary nebulae kinematics within 200 kpc of the cluster core. *Mon. Not. R. Astron. Soc.* 477, 1880–1892. doi:10.1093/mnras/sty663

Spolaor, M., Kobayashi, C., Forbes, D. A., Couch, W. J., and Hau, G. K. T. (2010). Early-type galaxies at large galactocentric radii - II. Metallicity gradients and the [Z/H]-mass, [α/Fe]-mass relations. *Mon. Not. R. Astron. Soc.* 408, 272–292. doi:10.1111/j.1365-2966.2010.17080.x

Springel, V., Pakmor, R., Pillepich, A., Weinberger, R., Nelson, D., Hernquist, L., et al. (2018). First results from the IllustrisTNG simulations: Matter and galaxy clustering. *Mon. Not. R. Astron. Soc.* 475, 676–698. doi:10.1093/mnras/stx3304

Strader, J., Romanowsky, A. J., Brodie, J. P., Spitler, L. R., Beasley, M. A., Arnold, J. A., et al. (2011). Wide-field precision kinematics of the M87 globular cluster system. *Astrophys. J. Suppl. Ser.* 197, 33. doi:10.1088/0067-0049/197/2/33

Taylor, M. A., Ko, Y., Côté, P., Ferrarese, L., Peng, E. W., Zabludoff, A., et al. (2021). Fresh insights on the kinematics of M49's globular cluster system with MMT/hectospec spectroscopy. *Astrophys. J.* 915, 83. doi:10.3847/1538-4357/abffcf

Theuns, T., and Warren, S. J. (1997). Intergalactic stars in the Fornax cluster. *Mon. Not. R. Astron. Soc.* 284L, 11–L14. doi:10.1093/mnras/284.3.L11

Thomas, D., Maraston, C., Bender, R., and de Oliveira, C. M. (2005). The epochs of early-type galaxy formation as a function of environment. *Astrophys. J.* 621, 673–694. doi:10.1086/426932

Thuan, T. X., and Kormendy, J. (1977). Photographic measurements of the diffuse light in the Coma cluster. *PASP* 89, 466. doi:10.1086/130146

Uson, J. M., Boughn, S. P., and Kuhn, J. R. (1991). Diffuse light in dense clusters of galaxies. I. R-Band observations of Abell 2029. *Astrophys. J.* 369, 46. doi:10.1086/169737

van de Sande, J., Bland-Hawthorn, J., Fogarty, L. M. R., Cortese, L., d'Eugenio, F., Croom, S. M., et al. (2017). The sami galaxy survey: Revisiting galaxy classification through high-order stellar kinematics. *Astrophys. J.* 835, 104. doi:10.3847/1538-4357/835/1/104

van der Marel, R. P., and Franx, M. (1993). A new method for the identification of non-Gaussian line profiles in elliptical galaxies. *Astrophys. J.* 407, 525. doi:10.1086/172534

Veale, M., Ma, C.-P., Greene, J. E., Thomas, J., Blakeslee, J. P., McConnell, N., et al. (2017). The MASSIVE Survey - VII. The relationship of angular momentum, stellar mass and environment of early-type galaxies. *Mon. Not. R. Astron. Soc.* 471, 1428–1445. doi:10.1093/mnras/stx1639

Veale, M., Ma, C.-P., Greene, J. E., Thomas, J., Blakeslee, J. P., Walsh, J. L., et al. (2018). The MASSIVE survey - VIII. Stellar velocity dispersion profiles and environmental dependence of early-type galaxies. *Mon. Not. R. Astron. Soc.* 473, 5446–5467. doi:10.1093/mnras/stx2717

Ventimiglia, G., Arnaboldi, M., and Gerhard, O. (2011). The unmixed kinematics and origins of diffuse stellar light in the core of the Hydra I cluster (Abell 1060). *Astron. Astrophys.* 528, 24. doi:10.1051/0004-6361/201015982

Wang, Y., Mao, S., Vogelsberger, M., Springel, V., Hernquist, L., and Wechsler, R. H. (2022). Early-type galaxy density profiles from IllustrisTNG - III. Effects on outer kinematic structure. *Mon. Not. R. Astron. Soc.* 513, 6134. doi:10.1093/mnras/stac1375

Watkins, A. E., Mihos, J. C., and Harding, P. (2014). Searching for diffuse light in the M96 galaxy group. *Astrophys. J.* 791, 38. doi:10.1088/0004-637X/791/1/38

Weijmans, A.-M., Cappellari, M., Bacon, R., de Zeeuw, P. T., Emsellem, E., Falcon-Barroso, J., et al. (2009). Stellar velocity profiles and line strengths out to four effective radii in the early-type galaxies NGC3379 and 821. *Mon. Not. R. Astron. Soc.* 398, 561–574. doi:10.1111/j.1365-2966.2009.15134.x

Weil, M. L., Bland-Hawthorn, J., and Malin, D. F. (1997). Diffuse stellar light at 100 kiloparsec scales in M87. *Astrophys. J.* 490, 664–681. doi:10.1086/304886

Williams, B. F., Ciardullo, R., Durrell, P. R., Vinciguerra, M., Feldmeier, J. J., Jacoby, G. H., et al. (2007). The metallicity distribution of intracluster stars in Virgo. *Astrophys. J.* 656, 756–769. doi:10.1086/510149

Wu, X., Gerhard, O., Naab, T., Oser, L., Martinez-Valpuesta, I., Hilz, M., et al. (2014). The mass and angular momentum distribution of simulated massive early-type galaxies to large radii. *Mon. Not. R. Astron. Soc.* 438, 2701–2715. doi:10.1093/mnras/stt2415

Yan, H., Hathi, N. P., and Windhorst, R. A. (2008). An overdensity of i-dropouts among a population of excess field objects in the Virgo cluster. *Astrophys. J.* 675, 136–145. doi:10.1086/527349

Zibetti, S., White, S. D. M., Schneider, D. P., and Brinkmann, J. (2005). Intergalactic stars in z 0.25 galaxy clusters: Systematic properties from stacking of sloan digital sky survey imaging data. *Mon. Not. R. Astron. Soc.* 358, 949–967. doi:10.1111/j.1365-2966.2005.08817.x

Zibetti, S., Gallazzi, A. R., Hirschmann, M., Consolandi, G., Falcon-Barroso, J., van de Ven, G., et al. (2020). Insights into formation scenarios of massive early-type galaxies from spatially resolved stellar population analysis in CALIFA. *Mon. Not. R. Astron. Soc.* 491, 3562–3585. doi:10.1093/mnras/stz3205

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 22 December 2022
DOI 10.3389/fspas.2022.1017103

Check for updates

OPEN ACCESS

EDITED BY
Jonathan Eastwood,
Imperial College London,
United Kingdom

REVIEWED BY
Scott William McIntosh,
National Center for Atmospheric Research (UCAR), United States
Joseph E. Borovsky,
Space Science Institute, United States

*CORRESPONDENCE
Natalia Buzulukova,
nbuzulukova@gmail.com

SPECIALTY SECTION
This article was submitted to Space Physics,a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 11 August 2022
ACCEPTED 13 December 2022
PUBLISHED 22 December 2022

CITATION
Buzulukova N and Tsurutani B (2022), Space Weather: From solar origins to risks and hazards evolving in time. *Front. Astron. Space Sci.* 9:1017103. doi: 10.3389/fspas.2022.1017103

COPYRIGHT
© 2022 Buzulukova and Tsurutani. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Space Weather: From solar origins to risks and hazards evolving in time

Natalia Buzulukova[1,2]* and Bruce Tsurutani[3]

[1]NASA GSFC, Heliophysics Division, Geospace Physics Laboratory, Greenbelt, MD, United States, [2]University of Maryland, Department of Astronomy, College Park, MD, United States, [3]Heliospheric Physics and Astrophysics Section, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, United States

Space Weather is the portion of space physics that has a direct effect on humankind. Space Weather is an old branch of space physics that originates back to 1808 with the publication of a paper by the great naturalist Alexander von Humboldt (Von Humboldt, Ann. Phys. 1808, 29, 425–429), first defining a "Magnetische Ungewitter" or magnetic storm from auroral observations from his home in Berlin, Germany. Space Weather is currently experiencing explosive growth, because its effects on human technologies have become more and more diverse. Space Weather is due to the variability of solar processes that cause interplanetary, magnetospheric, ionospheric, atmospheric and ground level effects. Space Weather can at times have strong impacts on technological systems and human health. The threats and risks are not hypothetical, and in the event of extreme Space Weather events the consequences could be quite severe for humankind. The purpose of the review is to give a brief overall view of the full chain of physical processes responsible for Space Weather risks and hazards, tracing them from solar origins to effects and impacts in interplanetary space, in the Earth's magnetosphere and ionosphere and at the ground. In addition, the paper shows that the risks associated with Space Weather have not been constant over time; they have evolved as our society becomes more and more technologically advanced. The paper begins with a brief introduction to the Carrington event, arguably the greatest geomagnetic storm in recorded history. Next, the descriptions of the strongest known Space Weather processes are reviewed, tracing them from their solar origins. The concepts of geomagnetic storms and substorms are briefly introduced. The main effects/impacts of Space Weather are also considered, including geomagnetically induced currents (GICs) which are thought to cause power outages. The effects of radiation on avionics and human health, ionospheric effects and impacts, and thermosphere effects and satellite drag will also be discussed. Finally, we will discuss the current challenges of Space Weather forecasting and examine some of the worst-case scenarios.

KEYWORDS

Carrington event, geomagnetic storms, substorms, geomagnetically induced currents, radiation storms, ionospheric disturbances, space weather

1 Introduction

The fields of Space Weather and space physics were strongly stimulated by the publication of a paper by the great naturalist Alexander von Humboldt (1808). Von Humboldt observed that oscillations of magnetic needles occurred for over 4 h when there were auroras overhead at his home in Berlin, Germany. He called this a "Magnetische Ungewitter" or "magnetic storm". Beginning in the mid-19th century, many technological systems on Earth were impacted by Space Weather. Movements of the telegraph magnetic needles of the Midland Railroad (England) were observed to be coincident with auroral sightings (Barlow, 1849). On 1 September 1859, R. C. Carrington and R. Hodgson both observed and reported an optical solar flare (Carrington, 1859; Hodgson, 1859). Carrington also noted that a large magnetic storm on Earth occurred some 17 h and 40 min after the flare. During the magnetic storm, fires were started in telegraph stations (Loomis, 1861; see also Shea and Smart, 2006 for a compilation of 8 works by Loomis). Details of this extreme magnetic storm have been reported by Kimball (1960), Tsurutani et al. (2003) and Lakhina et al. (2012). The special issue of Advances in Space Research (Clauer and Siscoe, 2006) brings together 23 articles on the Carrington event.

People were aware about geomagnetic activity long before the Carrington event (Maden, 2020). Although sailors of the past did not know what geomagnetic storms or substorms were, geomagnetic activity was of great interest to nations with sailing ships. Geomagnetic storms are worldwide events that can last from hours to tens of hours (Chapman and Bartels, 1940; Gonzalez et al., 1994). The Earth's external magnetic fields (the magnetosphere) becomes filled with energetic particles. The circulation of these particles, called the "ring current" can change the Earth's surface magnetic field directions by up to ~5°. S.-I. Akasofu, as a student of S. Chapman, described a sequence of spatial auroral evolution from all-sky camera data (Akasofu, 1964)[1]. Because he believed this sequence to be a fundamental sub-element of a magnetic storm, Akasofu called this a "substorm".

A substorm is a sequence of auroral features lasting from ~30 min to an hr. Not only are intense auroras associated with substorms, but strong ionospheric currents are as well. The relationship between substorms and magnetic storms still remains a major topic of study today. After decades of intense research, there is still no definite answer.

More recently in the space age satellite operators and designers have had to address effects of Space Weather on a daily basis. Power grid operators closely monitor geomagnetically induced currents (GICs) that could damage/destroy transformers and grid infrastructure. Space Weather has many adverse effects on human health. Future crewed missions to Mars and the Moon view particle radiation storms as potentially life-threatening events. If an extreme Space Weather event occurs, the consequences for Earth could be truly problematic for humankind. Cannon et al. (2013) and Hapgood et al. (2021) estimated that the time it would take to repair damage to the UK power grid system from an extreme geomagnetic storm would be weeks to months. Oughton et al. (2017) have estimated that the worldwide losses due to an extreme magnetic storm would be between $7B and $42B per day.

Because of the diverse and complex nature of Space Weather effects, it is impossible to provide detailed descriptions of all the (presently known) primary facets of Space Weather here. The purpose of the review is to briefly give an overview of the full chain of physical processes responsible for Space Weather risks and hazards. We will trace them from solar origins to particular effects and impacts in interplanetary space and at Earth. Further reading is encouraged to understand the multiple and sometimes interrelated nature of Space Weather effects/impacts. Helpful reviews and in-depth analyses can be found in the following books: "Collisionless Shocks in the Heliosphere: A Tutorial Review" by Stone and Tsurutani (1985), "Collisionless Shocks in the Heliosphere: Reviews of Current Research" by Tsurutani and Stone (1985), "From the Sun: Auroras, Magnetic Storms, Solar Flares, Cosmic Rays" by Suess and Tsurutani (1998), "Magnetic Storms" by Tsurutani et al. (1997), "Storm-Substorm Relationship" by Sharma et al. (2004), "Space Weather: The Physics Behind a Slogan" by Scherer et al. (2005), "Recurrent Magnetic Storms: Corotating Solar Wind Streams" by Tsurutani et al. (2006a), "The Sun and Space Weather" by Hanslmeier (2007), "Space Weather - Physics and Effects" by Bothmer and Daglis (2007), "Physics of Space Storms: From the Solar Surface to the Earth" by Koskinen (2011), "Extreme Events in Geospace: Origins, Predictability and Consequences" by Buzulukova (2018), "Space Weather Effects and Applications" by Coster et al. (2021), and a fairly comprehensive review of space plasma physics by Tsurutani et al. (2022a). We encourage the reader to explore the literature for the details of all physical processes.

2 Carrington event as a first encounter of the widespread effect of Space Weather on human technologies

The best-known Space Weather event is a chain of physical processes in August-September 1859, which began with the

1 Kristian Birkeland published a two volume book entitled "The Norwegian Aurora Polaris Expedition 1902–1903" where he introduced "Polar Elementary Storms" (Birkeland, 1908). Birkeland used his terrella experiments to show that "corpuscular rays" from the Sun could penetrate into the polar regions of the Earth. Analysis of Birkeland's contribution can be found in Egeland and Burke (2010). Akasofu's 1964 work was the first to introduce substorm morphology and provide a modern definition of substorm phases.

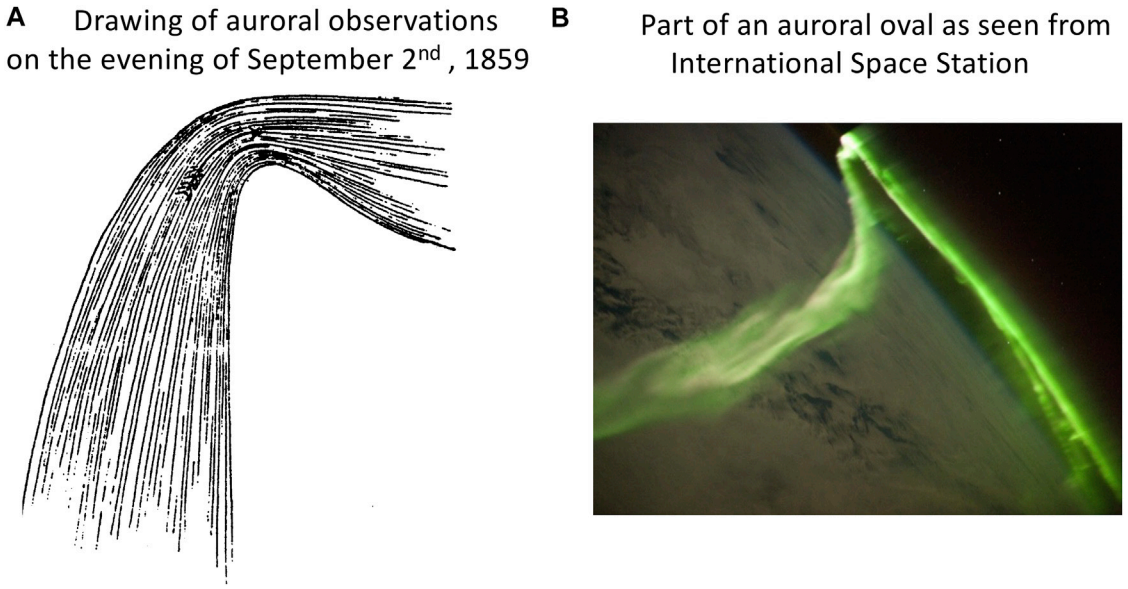


FIGURE 1
(A) A picture of the auroral display on the evening of 2 September 1859, Rome, N.Y. Lat. 43 deg 13 min designed by Mr. Edward Huntington. The picture was originally published in a "fifth article on Great Auroral Exhibition" by Prof. Elias Loomis, in the American Journal of Sciences, 1859–1861. Credit: COSPAR and Advances in Space Research. Source: Shea and Smart, 2006. **(B)** Part of an auroral oval as seen from the International Space Station, taken during a geomagnetic storm in 2010. Credit: NASA. Source: https://commons.wikimedia.org/wiki/File:Aurora_Australis_From_ISS.JPG.

appearance on the Sun's surface of a group of unusually large sunspots (Carrington, 1859). The event culminated in arguably the largest geomagnetic storm in recorded history on September 2nd (Chapman and Bartels, 1940; Kimball, 1960; Tsurutani et al., 2003; Lakhina et al., 2005, 2012). This is now known as the "Carrington event" or "Carrington storm."

The appearance of bright and vibrant auroral displays such as in Figure 1A across much of the globe during the Carrington magnetic storm must have been a memorable if not terrifying experience. For many people (Kimball, 1960). There were dozens of eyewitness accounts describing many auroral forms that lasted for days. We now understand that magnetospheric and auroral ionospheric currents are parts of the giant current system, the so called Global Electric Circuit, that connects the magnetosphere/ionosphere/upper atmosphere and the ground. The charged particles that carry out these currents (electrons and protons, with electrons being a primarily source of visible auroral forms) are accelerated at the distances 4,000–12,000 km above the Earth's surface and collide with neutral atoms in the topside atmosphere. These collisions mainly take place at 80–600 km altitude depending on the energy of the incident interacting charged particles (Shelley, 1995). As a result of electron collisional excitation of oxygen and nitrogen atoms and molecules, visible light is emitted with green, red, purple and blue colors, giving an observer a spectacular view of the dancing light of the Aurora Borealis in the Northern hemisphere, and the Aurora Australis in the Southern hemisphere (Figure 1B). A good review of auroral processes can be found in Paschmann et al. (2003).

Auroras also emit radiation in the ultraviolet and X-ray parts of the spectrum, but because these emissions are absorbed efficiently by the Earth's atmosphere, ultraviolet and X-ray auroras could only be observed from space or rocket experiments (Meier, 1991; Torr et al., 1995; Petrinec et al., 1999; Frey et al., 2004) and high-altitude balloon experiments (Winckler et al., 1959; Anderson and Milton, 1964; Brown et al., 1965; Millan, 2011), respectively.

Evidence of unusually high auroral activity is scattered through mythology and ancient documents, but it is hard to decode the details from these old writings. At the time people did not understand the aurora and its causes (Maden, 2020). It was also difficult to understand that auroral lights might have significant impacts on human society in the future, other than psychological effects. Examples could be found in the essay "Auroral Omens of the American Civil War" (Love, 2014).

However, it is not just great auroral displays that draw the attention of scientists and the general public to the Carrington event. There were numerous documented anomalies related primarily to telegraph communications, the top technological system of the day. These included large voltages at the ends of telegraph wires and fires were started by arcing. More

importantly, there were multiple interruptions in communication in Europe and North America, lasting from a few hours to most of the day on 2 September 1859. These disturbances were correctly attributed to the effect of "auroral electricity flows", although at that time scientists were unaware that strong ionospheric currents flowed parallel to Earth's surface at auroral latitudes at approximately 100 km altitude. These "invisible" ionospheric currents are today called the "auroral electrojets". These currents can be as intense as one million Amperes. The precipitating auroral particles provide high conductivity by producing ionospheric ionization for the electrojets and the currents close the giant current loops/circuits connecting the ionosphere and the magnetosphere.

A detailed description of the telegraph effects such as fires set in telegraph offices in August/September 1859 can be found in the review by Boteler (2006). Thus, one could describe the Carrington event as a *Space Weather* event, where a blob of plasma and embedded magnetic fields called a Coronal Mass Ejection (CME) came from the Sun and impacted the Earth's magnetosphere (CMEs will be described in more detail later) and caused a series of adverse effects on human technologies: in this case and time, on the global system of telegraph communications. In modern terminology, the physical effects that were responsible for communication interruptions are called Geomagnetically Induced Currents (GICs). GICs are widely recognized today as one of the greatest dangers associated with Space Weather (see Section 5.1 for the details). As some observers of the Carrington event correctly recognized, the changes in auroral emission intensities were related to variations in the magnetic field direction and strength at the Earth's surface. The ionospheric auroral electrojets resulted in strong induced currents in Earth-bound conducting communication devices, such as telegraph wires. It is the strength of the effects and the widespread interruptions of services that make the Carrington event a special one, thus defining it as an example of Extreme Space Weather. A review by Lanzerotti (2017) presented other examples of Extreme Space Weather related to the occurrence of GICs, including the May 15 of 1921, March 24 of 1940, February 10 of 1958 magnetic storms and other events.

In 1859, humanity encountered the negative manifestations of Space Weather effects with the high technology at the time: telegraph communications. With the invention of the first radio devices in 1895 (independently by the Italian scientist G. Marconi and Russian scientist A. Popov), the demonstration of the first transatlantic transmission of radio waves in 1901 by Marconi and his team brought humanity into a new era of wireless communications. Since the propagation of radio waves over long distances depends on ionospheric conditions, this new technology became sensitive to new classes of Space Weather effects (see Section 5.4 of this Chapter for details).

The launch of the first Sputnik in 1957 marked the start of the Space Age. Today, the functioning of many government services and private sector enterprises, depend heavily on modern satellite telecommunication systems. A well-known example is the Global Navigation Satellite System (GNSS) which uses satellites to provide positioning, navigation, and timing (PNT) services with high accuracy. An incomplete list of GNSS applications includes airplane, ship and land navigation systems, mapping, surveying, emergency location-based services, and high-speed financial trading systems. In addition to government owned and operated GNSS systems, there is a diverse private sector of telecommunications companies around the world specialized in providing worldwide voice and data communication.

As satellite telecommunication services are closely integrated into all aspects of the life of modern society, it is necessary to understand the effects and negative impact of Space Weather events on the functionality and reliability of these systems. Factors that can affect the performance of a single satellite or deteriorate the whole system strongly depend on the satellite orbits and in general include:

- Particle radiation and plasma effects on satellites (Section 5.2);
- Ionospheric Space Weather effects (Section 5.4);
- Thermosphere (upper atmosphere) effects and satellite drag (Section 5.5).

How many satellites are there in Earth orbit at the moment? The US Government Accountability Office (GAO) website (https://www.gao.gov/products/gao-22-105166) estimates the number of operational satellites orbiting Earth at nearly 5,500 in the spring of 2022. The UCS database (https://www.ucsusa.org/resources/satellite-database) indicates that the number of currently operational satellites is 5,465 as of 1 May 2022. It appears that most of these satellites are privately owned. As of 28 May 2022, the American company SpaceX has launched 2,653 Starlink satellites (https://spaceflightnow.com/2022/05/18/falcon-9-starlink-4-18-live-coverage/). In 2018, the United States Federal Communication Commission (FCC) authorized SpaceX to launch 4,425 satellites in Low Earth Orbit (LEO). As stated in the FCC press release, this authorization and similar authorizations are granted to a number of companies "... to provide broadband services using satellite technology that holds promise to expand Internet access, particularly in remote and rural areas across the country" (https://www.fcc.gov/document/fcc-authorizes-spacex-provide-broadband-satellite-services). It is worth noting that Starlink Low Earth Orbit (LEO) trajectories cover the critical altitudes of 500–600 km where satellite drag effects are important (to be discussed in Section 5.5). A recent example is SpaceX's loss of approximately 40 Starlink spacecraft (s/c) launched on 3 February 2022 and hit by two geomagnetic storms on February 3–4, 2022 (see SpaceX "Updates" at https://www.spacex.com/updates/from 8 February 2022; Dang et al., 2022; Fang et al., 2022 and Tsurutani et al., 2022b). With the daily increase in demand for high-speed data communications, the number of satellites will likely increase substantially each year. Society's exposure to the risks associated with Space Weather will obviously increase as well.

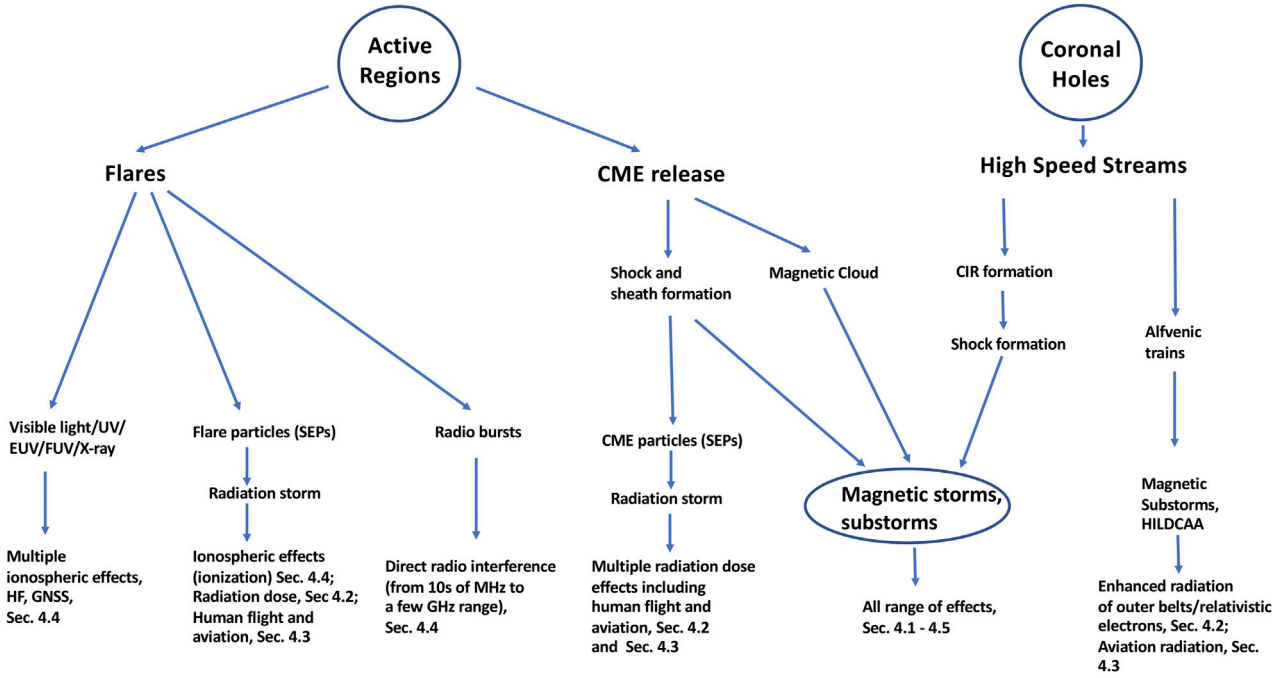


FIGURE 2
Relationship between the main Space Weather processes (solar eruptive processes, and high-speed solar wind) and Space Weather effects/impacts (it is noted that CME/flare events can also be due to filament eruptions in quiet region and the boundaries of polar coronal holes, but most of the energetic eruptions originate from active regions). Different Space Weather effects/impacts could occur at the same time due to close association with eruptive processes, but could also occur independently, e.g. radiation storms and geomagnetic storms.

3 Solar origins of Space Weather

Space Weather is now a growing multi-disciplinary field that studies the origins of the variability of solar-terrestrial processes and links these processes with particular impacts on technological systems and human health. Because of Space Weather's diversity, it is important to understand the physical processes responsible for the most significant effects and impacts of Space Weather. Thus we need to understand space plasma physics even better because of Space Weather effects.

Figure 2 highlights the relationships between diverse Space Weather processes originating from the Sun and Space Weather impacts. Some of these processes are created and evolve in interplanetary space. The two predominant solar sources of Space Weather are active regions (ARs, see review by van Driel-Gesztelyi and Green, 2015), decayed ARs and coronal holes (CHs, see review by Cranmer, 2009). ARs are a complex set of varying polarity sunspot magnetic fields in the photosphere of the Sun. The varying magnetic polarities are interconnected with the magnetic loops protruding above the surface. Due to evolution of the AR substructure and underlying subsurface dynamics, changes in the magnetic field interconnections (magnetic reconnection) lead to magnetic annihilation and sudden energy release in the form of a "solar flare" (Shibata and Magara, 2011; Toriumi and Wang, 2019). The solar flare photons cover the whole spectral scale from X-rays to UV to visible light photons (X-rays have to be detected by detectors onboard satellites or rockets above the Earth's atmosphere - thus only known during the space age). The visible light from a solar flare noted by Carrington (1859) and Hodgson (1859) is reasonably rare and thus as far as we know, was not reported in ancient documents. The magnetic reconnection at the Sun could also release a coronal mass ejection (CME), which will be described later. ARs occur primarily during the ~11-year solar cycle maximum sunspot number phase, called "solar maximum" for short (for details see the review by Hathaway, 2015 on solar cycle).

CHs are "dark regions" in the Sun seen in soft X-rays (Zirker, 1977; Harvey and Sheeley, 1979; Suess, 1979). Therefore, before the space age there was no visible feature on the Sun that could explain a possible 27 day periodicity in geomagnetic activity at the Earth (Maunder, 1905; Maunder, 1906) connecting the geomagnetic activity to the solar rotation[2]. However, Chree

2 Although Edward E. Maunder is listed as the sole author of these two works, there is growing appreciation that his wife, astronomer and mathematician Annie S.D. Maunder, also contributed to the discovery of the statistical association between recurrent magnetic storms and solar activity. Please see for more details: https://mathshistory.st-andrews.ac.uk/Biographies/Maunder/.

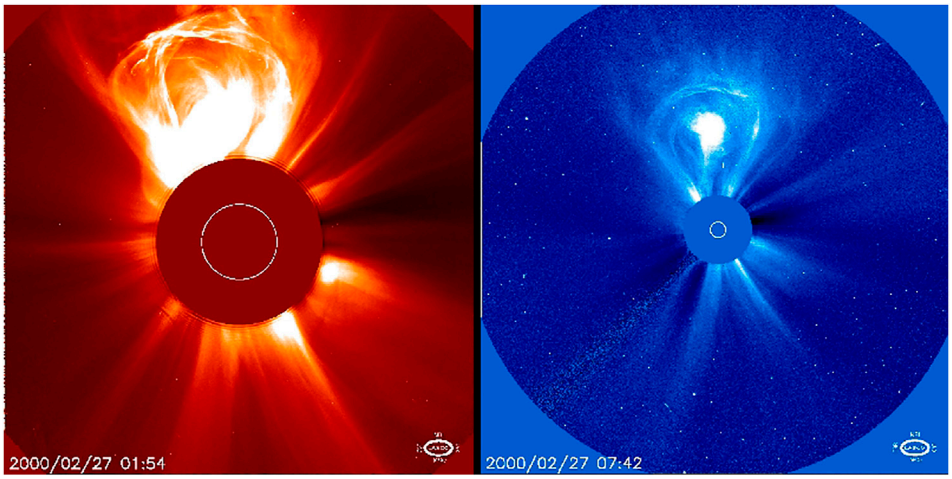


FIGURE 3
An example of a Coronal Mass Ejection (CME). The CME grows larger as it moves away from the Sun and becomes an Interplanetary CME (ICME). Images are taken by NASA's SOHO on 27 Feb 2000. Credit: SOHO ESA and NASA. Source: https://www.nasa.gov/content/goddard/what-is-a-coronal-mass-ejection.

(1913) showed mathematically that the apparent periodicity was statistically significant and there was something on the Sun that was causing this. CHs are regions that have open magnetic fields (see review by Cranmer, 2009). It is known from Ulysses observations (Phillips et al., 1995) that the fast solar wind with speeds of ~750–800 km/s emanate from CHs. It is also known that Alfvén wave magnetic field fluctuations in the high speed streams (HSSs) cause substorm geomagnetic activity at Earth (Tsurutani et al., 2006).

The solar origins of most Space Weather effects are mainly related to a few broad classes of phenomena: i) Solar eruptive phenomena (CMEs, solar flares and erupt filaments) which are linked to the dynamic processes of magnetic field reconfiguration (magnetic reconnection) and energy release in ARs and decayed ARs; ii) variability of solar irradiance due to magnetic convection and 27-day solar rotation, including important spectral bands of extreme ultraviolet (EUV, 10–120 nm), and ultraviolet (UV, 120–400 nm) emissions; iii) solar coronal holes from which high speed streams (HSSs) of solar wind emanate. The HSSs interact with slow speed streams in interplanetary space forming large scale compressive structures called Corotating Interaction Regions (CIRs) (Smith and Wolfe, 1976).

CIRs are responsible for CIR-driven types of geomagnetic storms. These are lower intensity storms than CME storms. The HSSs of ~750–800 km/s and embedded Alfvén waves cause auroral substorms and enhancement of magnetospheric electron fluxes with MeV energies (see Section 3.2 for details). CIRs and HSSs tend to occur during the declining phase of the solar cycle (Tsurutani et al., 2006; Hajra et al., 2013).

CMEs are giant magnetized plasma structures expelled from the Sun (Figure 3). CMEs could eject billions of tons of the Sun's material, and travel at speeds ranging from a few hundreds to a few thousands of km/sec (Forbes, 2000; Chen, 2011). The size of a single CME expands as it propagates from the Sun through the interplanetary medium and could reach ~0.5 astronomical unit (au) in width or larger by the time it reaches the Earth (Byrne et al., 2010). When a CME propagates away from the Sun, it interacts with the interplanetary medium and can be strongly transformed. Therefore, CMEs in interplanetary space are called interplanetary CMEs, or ICMEs. ICMEs are primarily magnetic clouds, the regions of high magnetic field intensities in the form of giant flux ropes (Burlaga et al., 1981; Burlaga et al., 1982). High magnetic field intensities in ICMEs can cause geomagnetic storms on Earth[3] if the magnetic fields have strong southward components (Gonzalez et al., 1994).

When an interplanetary CME (ICME) propagates at a speed faster than the ambient upstream solar wind by more than the upstream fast wave (magnetosonic) speed, a shock wave is formed in front of the ICME. (This is analogous to a jet fighter plane flying faster than the sound speed. A shock forms upstream of the plane wings. This shock can propagate down through the atmosphere and be heard and felt as a "sonic boom" by people on Earth.) The properties of the solar wind are significantly altered between the shock wave and the ejected material. This region is called the sheath. The shock serves to compress the solar wind plasmas and magnetic fields, slow plasma flow to submagnetosonic speeds in the frame of the ICME/piston, and convert bulk kinetic energy into heat (Kennel

3 The list of near-Earth ICMEs for the period 1996–2022 could be found at https://izw1.caltech.edu/ACE/ASC/DATA/level3/icmetable2.html.

et al., 1985; Tsurutani et al., 2011). The higher magnetic field intensities in the sheaths can cause intense magnetic storms, similar to ICMEs (Meng et al., 2019).

CMEs are closely connected to intense *solar flares,* or electromagnetic radiation bursts[4]. As far as we know, a significant fraction of X-class flares[5] are accompanied with CME releases and conversely, extremely fast CMEs (solar wind velocity V_{sw} > 2,000 km/s) have been associated with X-class solar flares. Yashiro et al. (2005) reported that the CME association rate for the flares in X1-X2 classes are 82–91%, and above X3 level all flares had associated CMEs. Less powerful flares (M- and C-class flares) are not always accompanied by CMEs (Burlaga et al., 1981; Klein and Burlaga, 1982). Yashiro et al. (2005) reported a ~20% CME association rate for C-class flares, and ~49% CME association rate for M-class flares. Both flares and CMEs are thought to be the result of magnetic field reconfiguration and reconnection at the Sun. The annihilated field energy goes into the flare electromagnetic energy, acceleration of charged particles and the release and kinetic energy of the CME. Although CMEs and solar flares are typically strongly related, there are exceptions. Yashiro et al. (2005) reported 4 X1-class flares (out of 49 X-class flares) without CMEs. Furthermore, some CMEs are related to "disappearing filaments" and some have had no obviously related flare/disappearing filament (Tsurutani et al., 1988; Tang et al., 1989; Tang and Tsurutani, 1990; Kamide and Kusano, 2015; Lakhina and Tsurutani, 2018).

It is thought that around 10–20% of the energy stored in the magnetic field associated with a group of sunspots could be released as electromagnetic energy of flares, including UV to FUV, soft X-ray emissions and radio emissions (Leka and Barnes, 2018). Different parts of the electromagnetic spectrum produce different Space Weather effects; see Section 5.4 for the description of different ionospheric effects.

Reconnection, solar flares and CMEs are associated with *Solar Energetic Particles (SEP).* SEP are an increased flux of energized protons (up to GeV energies), heavy ions, and relativistic electrons, either coming directly from near-reconnection locations in the solar corona, or from CME-driven shock waves. Currently, the two main means of SEPs reaching Earth are considered to be: 1) interplanetary magnetic field lines traced backwards from Earth that connect to the flare site in the lower corona; 2) interplanetary magnetic field lines directly connected from Earth to the shock waves formed by propagation of the CMEs through interplanetary space. The first type of SEPs is called 'impulsive SEP events' and the second type is called 'gradual SEP events' because of the temporal profile of the particle radiation events observed near Earth (e. g. Desai and Giacalone, 2016; Tsurutani et al., 2009). The particle radiation storm from a gradual SEP event could last a few days or longer depending upon the transit time of the shock in front of the CME that propagates from the Sun to the Earth and beyond. Gradual SEPs are believed to be more important for Space Weather because impulsive SEPs are shorter in duration and induce less intense radiation levels. For more details concerning flares, CMEs and energetic particle acceleration, we refer the reader to recent review by Tsurutani et al. (2022a).

For the sake of completeness, we note that there are other sources of energetic particles, including Galactic Cosmic Rays (GCRs), particle acceleration at planetary magnetospheric bow shocks, CIR shocks, and the heliospheric termination shock (Reames, 1999). Aside from SEPs, GCRs are also important for Space Weather, especially when estimating the radiation danger for human flight (Lockwood and Hapgood, 2007). The level of GCRs changes normally in antiphase with the solar cycle, as the varying interplanetary magnetic field strength modulates GCRs entering the heliosphere. On the Earth's surface, modulation of GCRs by ICMEs cause "Forbush decreases", rapid decreases in GCR intensities due to particle deflection from the intense ICME magnetic fields.

SEPs cause particle *radiation storms* in the near-Earth environment and are responsible for numerous Space-Weather effects: Single Event Effects (SEE) and radiation impacts on avionics in general, increased ionization levels and variations of ionospheric densities in polar regions leading to radio wave absorption, increased radiation risks for astronauts and airplane crew members and passengers on polar flights, and Ground Level Enhancement (GLE) radiation at the Earth's surface (see also Section 5.2, Section 5.3, Section 5.4 for specific Space Weather effects of SEPs.)

Prediction of radiation storm intensities is an important and not completely solved problem in the Space Weather community. It is impossible to predict SEP fluxes at present - on any meaningful timescale. This is partly due to yet unpredictable particle acceleration efficiencies and partly due to the complex geometry/magnetic connectivity issues. Energetic particles are guided by the interplanetary magnetic field lines. If the energetic particles accelerated at the flare site are released onto magnetic fields that connect to the Earth, those particles will be detected in the near-Earth environment. Particles accelerated at ICME shocks will reach the Earth if the field lines from that acceleration regions connect to the Earth. A complicating factor in the latter mechanism is that a CME shock has different particle acceleration properties at different locations along its surface (Kennel et al., 1984; Kennel et al., 1985; Tsurutani and Lin, 1985).

4 Curious readers may enjoy the Grand Archive of flare and CME cartoons as well as the list of relevant references at https://www.astro.gla.ac.uk/cartoons/index.html.

5 The standard flare classification is based on a peak flux intensity (W/m2) in 1–8 A X-ray band. The three largest classes are X, M, C. The scale is logarithmic: a X1-class flare is 10 times more intense than a M1-class flare and a M1-class flare is 10 times more intense than a C1-class. Classification of X flares goes beyond X9 to X11, X28 (on 4 November 2003, the Sun's strongest X-ray flare on record), and to X100, X1,000 and up for superflares from solar-like stars.

Since solar flares (in particular the most powerful ones) are easy to detect using X-ray detectors onboard Earth orbiting satellites, and also because of the strong statistical links between energetic flares and large CMEs/SEPs, forecasting of solar flares is becoming a growing branch of research in Space Weather (Leka and Barnes, 2018). With the accumulation of new data from imagers, solar flare forecasting is beginning to use modern methods of statistical analyses, including machine learning (ML) and artificial intelligence (AI) techniques (Jiao et al., 2020).

Solar eruptive processes are linked to the most dangerous and disruptive effects/impacts and cause development of CME-driven *geomagnetic storms, radiation storms, substorms, and ionospheric effects*. Statistically, the most energetic eruptive phenomena tend to occur at solar maximum and slightly thereafter (Chen, 2011).

Variability of solar irradiance over the solar cycle and over solar rotation in the extreme ultraviolet (EUV, 30–120 nm) and far ultraviolet (FUV, 120–200 nm) spectral bands is one of the major factors defining the variability of ionospheric ionization and thermospheric neutral densities[6]. Thus, this is an important Space Weather variable (Woods and Rottman, 2002). FUV/EUV variability closely follows what is known as the International sunspot Number (ISN). Statistical studies reveal a close correlation between ISN, solar radiance in the FUV/EUV range and solar radio flux at 2.8 GHz (10.7 cm wavelength), known as the F10.7 index[7] (e.g., Lukianova and Mursula, 2011). The underlying dynamics are thought to be complex and different for different spectral lines, showing the dependence from multiple processes in the photosphere, chromosphere and lower coronal (Floyd et al., 2005).

4 Geomagnetic storms, substorms and HILDCAA intervals

Since many Space Weather effects are related to the occurrence of *geomagnetic storms and substorms,* it is instructive to consider these processes in detail.

4.1 Geomagnetic storms

The interplanetary structures responsible for the formation of intense and large magnetic storms have been shown to be magnetic clouds (MCs, one of three parts of an ICME) and the upstream shocked sheath magnetic fields. CIR structures are responsible for milder geomagnetic storms. When the southward directed magnetic fields of the CIR, ICME or the sheath upstream of ICME interact with the terrestrial magnetic field, they cause an increased influx of energy into the magnetosphere. Solar wind energy is transported into the magnetosphere through magnetic reconnection at the dayside, the so called 'merging' of the solar wind magnetic fields with magnetospheric magnetic fields (Dungey, 1961). Magnetic field lines that are connected by one end to the solar wind and by another end to the Earth's ionosphere are dragged from the dayside to the nightside, causing accumulation of magnetic energy in the tail. After some threshold is reached, magnetic energy in the tail is released in magnetic reconnection and it causes the reversed plasma motion from nightside to the dayside. The whole process of plasma motion is called a global convection process. In the polar and mid-latitude ionospheres, this process creates a characteristic two-cell pattern of plasma motion called the Dungey convection cycle (Axford and Hines, 1961; Dungey, 1961).

The increased energy input during geomagnetic storms is related to the opening of the magnetosphere to the interplanetary magnetic field, B (From here and below, we assume Geocentric Solar Magnetospheric, GSM coordinate system for the magnetic field). For the interplanetary magnetic field B_z oriented in the same direction as the Earth's dayside magnetopause magnetic field lines, the magnetosphere is said to be 'closed' with the minimal energy input from the solar wind (Tsurutani et al., 1992; Du et al., 2008). In the case where B_z is oriented in the opposite direction to the Earth's magnetic dipole and magnetic reconnection occurs, the magnetosphere is said to be 'open'.

Although the interplanetary magnetic field B_z and solar wind velocity V_{sw} are the main factors contributing to the energy input into the Earth's magnetosphere, they are not the only solar wind parameters discussed in the literature (see e.g., Newell et al., 2007 and references therein). But in general, V_{sw} x $B_{z,south}$ is the main contributor to the energy inflow.

When there is a prolonged interval of southward B_z in the solar wind (a few hours), there is the formation of a torus-like region of enhanced fluxes of H^+, O^+ and electrons with energies ~10–300 keV. These particles which are initially injected into the nightside magnetosphere by Dungey's convection, drift around Earth under the combined effect of magnetic field curvature and magnetic field gradient drifts, forming a giant 'ring current' flowing in the space around the Earth (Dessler and Parker 1959; Gonzalez et al., 1994; Daglis, 2001).

The ring current creates a disturbance in the magnetic field strength (diamagnetic effect) that is detected at the Earth's surface. The disturbance is relatively small, and in most cases, it is less than 300 nT, or less than 1% of the Earth's magnetic field at the surface. However, the disturbance field has a characteristic time dependence. It covers ground-based magnetometer stations in a wide range of longitudes, lasting a few hours and then

6 Only EUV radiation can ionize primary thermospheric species such as O, O_2 and N_2 (ionization energy thresholds are ~91, 103 and 80 nm respectively). An important thermospheric compound NO could be ionized by FUV Lyman-alpha radiation, contributing to the major ionization source at ~ 50–90 km, so-called D-layer of the ionosphere.

7 The current sources of International sunspot Number and F10.7 are the sunspot Index Data Centre at Royal Observatory of Belgium for ISN (https://www.sidc.be/silso/datafiles) and National Research Council Canada/Natural Resources Canada (https://www.spaceweather.gc.ca/forecast-prevision/solar-solaire/solarflux/sx-en.php).

recovering more slowly over another half day to days. The ring current disturbance is measured at low geomagnetic latitudes (20°–30°) where the effect of auroral currents and equatorial electrojet are negligible and is called the disturbance storm-time or Dst index (now a 1-min SYM-H index is available). When the Dst index falls below -50 nT, it signals the occurrence of a geomagnetic storms (Gonzalez et al., 1994; Lakhina and Tsurutani, 2018).

The increase in magnetospheric convection during geomagnetic storms causes the increase of energy input into the coupled thermosphere - ionosphere system. Ionospheric plasma interacts with neutrals through collisions, therefore ionospheric currents generate a substantial amount of Joule heating that directly affects thermospheric (neutrals) uplift and satellite drag effects (See Section 5.5). Enhanced convection is also responsible for the transport of electrons and ions from the magnetospheric tail region toward the Earth and the inner magnetosphere, as well as electron and ion energization by conservation of the adiabatic invariants, and creation of the ring current. The interval of large southward B_z component, when the strong ring current builds up, is the "storm main phase". It is often followed by an interval of northward B_z, making the magnetosphere more closed with minimal energy input. The stored ring current particles are lost through several different mechanisms (Kozyra et al., 1997). The storm phase in which the Dst index recovers is called 'storm recovery phase' and normally lasts a half day or more. The entire process could take a day or two.

In summary, a geomagnetic storm is a period of increased geomagnetic activity characterized by enhanced energy input into the coupled magnetosphere-ionosphere-thermosphere system, controlled to large extent by extended intervals of the southward component of the interplanetary magnetic field, high solar wind velocities (Gonzalez and Tsurutani, 1987; Echer et al., 2008; Lakhina and Tsurutani, 2016), and by other parameters as solar wind density and temperature (Newell et al., 2007).

The main features of magnetospheric dynamics and plasma circulation during geomagnetic storms can be reproduced by modern global 3D codes. These codes combine multiple modules to represent different regions and different physical processes of the coupled magnetosphere-ionosphere system (De Zeeuw et al., 2004; Toth et al., 2005; Fok et al., 2006; Moore et al., 2008).

Figure 4 shows an example output of a global code and demonstrates the structure of all major current systems in the magnetosphere, including the ring current. The global codes are mature enough to be used for forecasting geomagnetic activity at NOAA's Space Weather Prediction Center (SWPC) (https://www.swpc.noaa.gov/products/geospace-magnetosphere-movies). Despite the progress in development, many important issues remain unresolved. To name a few, it is not understood how to properly describe the ionospheric source(s) of plasma in global codes (so-called ionospheric outflow), how to describe and include multiple wave-particle interactions, and what is the role of kinetic effects on the global structure and dynamics of the magnetosphere.

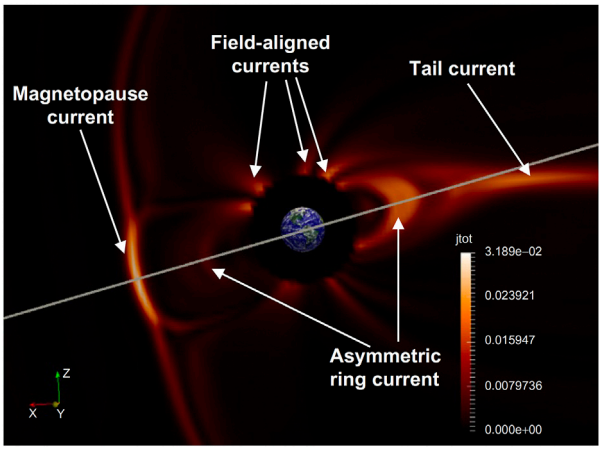


FIGURE 4
A meridional cut through the global 3D simulation domain showing the results for the total current density for 23 July 2015 geomagnetic storm. The main magnetospheric current systems are presented: the asymmetric ring current, tail current, magnetopause current, and field-aligned currents (more visible in the regions of converging magnetic field). The sun is from the left. The gray line is the projection of a plane through the approximate geomagnetic equator in the inner magnetosphere. Current density is shown in μA/m^2. From Buzulukova et al., 2018b.

4.2 Geomagnetic substorms and HILDCAA intervals

Another type of geomagnetic activity is a *geomagnetic substorm*. There can be many substorms within a magnetic storm or they can occur independently as isolated events. There are a variety of different models for the cause of substorms. In fact, all may be correct if applied for different events. We will describe below one commonly used model of isolated substorms. For an isolated substorm, the energy from the solar wind is first transported from the dayside to the nightside magnetosphere or magnetotail through the dayside magnetic reconnection process. In the tail, the energy is stored in the form of magnetic energy with a very stretched tail configuration and large values of magnetic field B (substorm growth phase). Since the process of energy accumulation is not infinitive, at some point the magnetosphere loses its stability and the energy is released in non-stationary tail reconnection (substorm onset). After the substorm onset, the stored magnetic energy is transformed into kinetic energy of plasma (note the analogy with CME release, Section 2), appearance of auroras and intense ionospheric currents (substorm expansion phase). Substorms cause additional particle transport and energization, that is,

particle substorm injections into the ring current plasma. The last step is the return of the system to the initial state (substorm recovery phase).

High-Intensity, Long-Duration, Continuous AE Activity (HILDCAA: Tsurutani and Gonzalez, 1987) intervals are by definition intervals of enhanced auroral activity and are a combination of cyclical substorms and DP2 (Nishida, 1968a) events (Tsurutani et al., 1995; Tsurutani et al., 2004). The most intense substorms occur during the solar cycle declining phase when there are large coronal holes at the Sun that emit HSSs of solar wind plasma. HILDCAA activity is produced by magnetic reconnection associated with embedded Alfvén waves within the HSSs. HILDCAAs and enhanced electromagnetic chorus waves produce enhanced fluxes of relativistic MeV "killer" electrons in the magnetosphere (Hajra et al., 2015; Hajra and Tsurutani, 2018).

It is important to understand that substorms are sometimes parts of geomagnetic storms and sometimes not. Substorms are related to the buildup/release of magnetic energy, and it is advantageous to consider them independently from the buildup of kinetic energy in the ring current during geomagnetic storm. The buildup of kinetic energy and induced changes of B-field at nightside may alter stability properties of the tail. It has been reported that storm-time substorms and isolated substorms have different characteristics (e.g., Pulkkinen et al., 2007; Partamies et al., 2013). There are other modes of magnetospheric behavior besides cyclical substorms, such as steady magnetospheric convection events (convection bays: Sergeev et al., 1996) and recurrent internally driven substorms (Keiling et al., 2022).

Examples of isolated substorms can be found in Tsurutani and Meng (1972) and in HILDCAAs (Tsurutani and Gonzalez, 1987). For both cases, where there is insufficient ring current activity to indicate the presence of a magnetic storm (Dst/SYM-H > -50 nT). On the other hand, it has been shown that in some MC-induced storms, there are no substorms (Tsurutani et al., 2004c). This has been speculated as being caused by the very smooth rotations of the IMF Bz. Are these the Sergeev et al. (1996) convection bays but giant in size? It is not clear at the present time.

The magnetosphere often transitions between different modes, and possibly exhibits a memory, where the present mode is defined by the prehistory of both the solar wind input and magnetospheric/magnetotail dynamics. At present, it is difficult for scientists to accurately predict what type of magnetospheric activity will be induced by a given solar wind/interplanetary magnetic field input. This is currently an active area of research.

5 Major space weather effects and impacts

We have briefly reviewed some of the Space Weather processes shown in Figure 2. The system is very complex, and many Space Weather effects and impacts could be linked to multiple solar origins. Among the most important Space Weather effects/impacts are: 1) Geomagnetically Induced Currents (GICs); 2) radiation and plasma effects/impacts related to surface charging and arcing, deep dielectric charging, avionics and s/c electronics; 3) effects/impacts of radiation related to human health, 4) ionospheric effects/impacts related to navigation and communication; 5) thermosphere effects and satellite drag. The above divisions are not unique. Since Space Weather science has many facets, there are multiple ways to organize these classifications. The detailed description of each Space Weather impact is a separate branch of Space Weather science; therefore, below we only briefly present the most important Space Weather effects/impacts.

5.1 Geomagnetically induced currents (GICs)

As mentioned above, the first encounters of GIC effects on human technologies took place in the middle of the 19th century with the widespread development of telegraph systems. Now, GICs are recognized as a one of the main Space Weather threats. The underlying physics could be explained by Faraday's law: changes in the large currents systems (e.g., electrojet) cause a changing magnetic field that induce electric fields (Lakhina et al., 2020). The electric fields drive strong currents on the ground if the conductivity is large enough. These induced currents flow near the Earth's surface. Sometimes the currents can get into transmission cables and power lines and cause strong GIC effects. Any technological system that has good electrical conductors as an essential design element and also by design electrically connected to the Earth is potentially vulnerable to the effect of GICs. The best-known examples are power grid outages and transformers failures, with more exotic cases including railway system failures, pipeline monitoring system failures and impacts on undersea telecommunication traffic cables (Lanzerotti, 1992; Boteler and Jansen van Beek, 1999; Viljanen et al., 1999; Wik et al., 2009; Eroshenko et al., 2010; Hapgood, 2018; Ngwira and Pulkkinen, 2018).

Variations of the geomagnetic field and resultant GICs are due to the reconfiguration of the giant current loops circulating in the near-Earth environment. These currents create a magnetic field and shape the magnetosphere. For example, the magnetic field of the magnetopause current (Chapman-Ferraro current) defines the shape of the magnetopause, and the cross-tail current (and current loops) defines the shape of the geomagnetic tail. When the global current system reconfigures itself as a result of dynamic changes in the magnetic field/plasma conditions of the solar wind (e.g., geomagnetic storms) and/or changes in the balance

of stored energy inside the magnetosphere (e. g. magnetospheric substorms), the ionospheric part of the current loop also changes, producing strong and fast variations of the magnetic field at the Earth's surface. The main drivers of GICs are thought to be:

i) Substorm ionospheric currents and variations in the auroral electrojet (Pulkkinen et al., 2003; Viljanen et al., 2006; Tsurutani and Hajra, 2021). The ionospheric currents can be as large as 10^6 A and located only ~100 km above the Earth's surface;
ii) Rapid reconfiguration of dayside Chapman-Ferraro currents related to shock wave/sheath impact and possible interactions with the equatorial electrojet (Carter et al., 2015). Storm commencement due to shock impact is believed to be the main driver of low-latitude GICs (Gaunt and Coetzee, 2007).
iii) Other sources of strong geomagnetic variations include geomagnetic field pulsations, possibly associated Kelvin-Helmholtz instabilities, and possibly related ring current reconfigurations (Buzulukova et al., 2018).

Recent studies show that intense GICs are correlated with the occurrence of intense substorms (Tsurutani and Hajra, 2021). However, it is very difficult at present to predict the occurrence of an intense substorm and related intense GICs at a particular location because GICs could be very localized (Ngwira et al., 2015; Pulkkinen et al., 2015; Tsurutani and Hajra, 2021), and likely related to transient structures, bursty bulk flows or depolarization flux bundles or both, sub-units of geomagnetic substorms (Kepko et al., 2015). Global models of the magnetosphere are improving each year, but at the current state of knowledge, it is difficult to reproduce intense GIC structures as observed in the data. A detailed summary of the challenges of predicting and understanding GICs is presented in Thomson et al. (2010), Pulkkinen et al. (2017), Lakhina et al. (2020).

5.2 Particle radiation and plasma effects on satellites

Exposure to radiation from a variety of sources including energetic electrons, protons, heavy ions and neutrons degrades satellite systems. When properly designed, electronic and avionic components of s/c systems should meet certain aerospace electronics standards to be resilient to radiation exposure in a particular orbit, e.g., LEO, Medium Earth Orbit (MEO), or Geosynchronous orbit (GEO). High-energy particles passing through microelectronic devices undergo a chain of complex interactions with the material of the device, depositing the energy and creating charge. There are a few main classes of radiation effects that are taken into account when designing satellite systems: *single event effects (SEEs), total ionizing dose, and displacement damage dose*. *Surface charging and arcing*, as well as *deep dielectric charging* are usually considered separately (Baker et al., 2018).

SEEs are broad class of anomalies related to a single passage of an energetic particle, the production of an electron-hole pair[8] in a silicon substrate being probably the most common case (Baumann, 2005). These can include pulses in logic or bitflips in memory cells. They are mostly non-destructive and transient, but could potentially become an overload during a radiation storm. Some SEEs may require a power reset, and some are considered potentially destructive. *Total ionizing dose effects* are related to the amount of energy deposited by energetic particles, and *displacement damage dose* refers to the creation of defects in semiconductor lattices (displacement of atoms from their lattice sites caused by passage of the energetic particle). Total ionizing doses and displacement doses are called cumulative radiation effects, and should be estimated for a given orbit. More information on radiation effects could be found in the NASA GSFC Radiation Effects and Analysis web-site on https://radhome.gsfc.nasa.gov/top.htm, NASA Applied Space Environments web-site on https://www.nasa.gov/nase, and ESA's SPace ENVironment Information System (SPENVIS) on https://www.spenvis.oma.be/. See also Barnaby and Marinella (2013) and Tsurutani et al. (2000). The latter was a NASA white paper written by an international group of particle scientists and engineers for the protection of the Parker Solar Probe mission.

There are multiple Space Weather processes that cause radiation damage. The main sources are: GCRs, SEPs, and trapped radiation from the electron and proton radiation belts. The GCR background is well-known. The high-intensity proton inner radiation belt fluxes and spectra are relatively well known and vary slowly with solar cycle except SEPs effects (Selesnick and Albert, 2019; Lozinski et al., 2021). Therefore, the most important sources of radiation for Space Weather effects are solar energetic protons and ions (SEPs/radiation storms), and MeV electrons of the inner/outer radiation belts (substorms/CIR storms/HILDCAA events). A review of radiation belt processes can be found in Baker et al. (2018).

Surface charging/arcing effects and *deep dielectric charging* remain serious Space Weather threats to satellites. Surface charging/arcing effects refer to an accumulation of electric charge on parts of s/c due to the interaction with the ambient

8 In silicon, one electron-hole pair is produced for every 3.6 eV of energy lost by the ion. Therefore, a single 1 MeV ion could produce up to ~278,000 electron-hole pairs. See Baumann, 2005 for the details.

energetic plasma. An enhanced level of electron fluxes in the energy range 100 eV-50 keV is known to be a hazardous environment. A significant potential difference between different (not electrically connected) parts of s/c[9] (up to a few kV) could build up over a relatively short period of time (minutes). This could cause s/c anomalies and loss of satellite control. A potential difference of a few kV can cause arcing effects, solar arrays being one of the most vulnerable parts. LEO satellites passing through the auroral zone, and satellites in GEO, are believed to be the most vulnerable to surface charging effects. The main Space Weather processes responsible for surface charging are thought to be geomagnetic substorms or substorm-like transients (Section 3.2), which cause enhanced fluxes of 1–50 keV electrons in the GEO region, or enhanced fluxes of precipitating electrons in the auroral zone (DeForest, 1972; Spence et al., 1993; Ganushkina et al., 2021). More details about surface charging can be found in Ferguson (2018).

Deep dielectric charging refers to the accumulation of electric charge and intense electric field build-up, caused by more energetic particles, usually electrons with energies up to MeVs, but also by ions of MeV energies, because the latter could amplify charging (Lai et al., 2018). If the electric field reaches a critical value 10^6–10^8 V/m depending on the material, there is a high probability for dielectric discharges to occur. The discharges can cause s/c anomalies, with potential damage to some important s/c systems, e. g. printed circuit boards. Since the high energy electron flux is far less than that of 1–50 keV electrons, the accumulation of sufficient charge for a discharge could take days or even weeks, therefore showing a delay time from the onset of the charging. Statistically, most of the internal charging/discharging effects have been shown to occur during the declining/minimum phases of the solar cycle (Wrenn et al., 2002; Lohmeyer and Cahoy, 2013). These results are consistent with findings that outer MeV electron radiation belt tends to intensify during the declining phase (Miyoshi and Kataoka, 2011). The main Space Weather processes responsible for the deep dielectric charging effects are geomagnetic substorms, or substorm-like events (HILDCAAs) that cause outer MeV radiation belt fluxes to intensify (Hajra et al., 2015).

More information about s/c charging could be found in the NASA Technical Handbooks NASA-HDBK-4002A w/ CHANGE 1, NASA-HDBK-4006 as well as in Ferguson and Hillard, 2003; Garrett and Whittlesey, 2012; Ferguson, 2018; Zheng et al., 2019.

9 To reduce the risk of electrical discharges between different surface parts, a s/c needs to be designed and built so that all surface parts are electrically interconnected. Although efforts have been made to make the s/c surfaces as conductive as possible, the spacecraft area still contains insulating materials, for example solar cell cover glass (Puthanveettil et al., 2014).

5.3 Radiation effects related to human health

Studies of Space Weather radiation effects on human health could be divided into two broad groups: 1) radiation field in space and assessment of the health risks for astronauts aboard the ISS, or crewed missions to Mars and Moon; 2) radiation field at aviation altitudes and assessment of health risks for aircrew, high-altitude pilots and frequent travelers.

Two major sources of radiation risks for humans in space are: major SEP events and GCRs. GCR ions are more energetic than SEP ions and electrons with a broad spectrum of GeV energies and higher. GCR ions have an isotropic distribution whose flux variations during the solar cycle are well-known. GCR particle fluences are therefore predictable and could be calculated for a given orbit; however, due to the very high energies of GCRs it is almost impossible to construct an efficient shielding outside of the Earth's magnetic field and atmosphere.

SEPs typically have a spectrum with energies of up to hundreds of MeV. The spectrum is difficult to predict due to the underlying complexity of the formation mechanisms (see Tsurutani et al., 2020). The fluxes for major SEPs are much higher than compared to background GCRs. The threat for astronauts from SEPs is not hypothetical. Estimations show that the 4 August 1972 event that occurred a few months before NASA's Apollo 17 landed would have been lethal for astronauts walking on the lunar surface (Lockwood and Hapgood, 2007). However, for long-duration missions (e.g., to Mars) chronic exposure to GCRs will likely be more challenging (Schwadron et al., 2014; Dobynde et al., 2021). Due to a very different radiation profile, chronic exposure for GCRs and acute burns for SEPs, mitigation approaches are likely to be very different (Narici et al., 2018; Hapgood, 2019).

Radiation dose and its effects on avionics at aviation altitudes has been studied with environment monitors onboard supersonic Concorde flights (Dyer et al., 1989). Awareness of natural radiation at aviation altitudes as a health hazard began to increase in the early 1990s (Band et al., 1990; Barish, 1990). In 1994, the US Federal Aviation Administration (FAA) formally recognized the associated health risks (Friedberg and Copeland, 2003). This triggered the need for quantification and characterization of the global radiation environment at aviation altitudes (Friedberg and Copeland, 2011; Tobiska et al., 2016). This is an area of ongoing research activity, both for modeling and measurements.

The radiation field of concern at aviation altitudes mainly originates from SEPs, GCRs and relativistic magnetospheric electrons. SEPs and GCRs with energies more than 400 MeV are a hazard for low and middle magnetic latitudes. Even for quiet Space Weather conditions, the radiation field at aviation

altitudes is more than an order of magnitude higher than at the ground level. After the great Halloween geomagnetic storms of 2003, the Space Weather community realized that the radiation index designed for geostationary orbit (particle flux with energy 10 MeV and above) was insufficient for aviation since a major part of the 10–100 MeV radiation is shielded by the Earth's atmosphere. A special index for the aviation radiation environment has been introduced, the D-index (Meier and Matthia, 2014). Extreme environments are still difficult to monitor during real-time flights, which is why the radiation field estimations are made from historical GLE observations on the ground. Dyer et al., 2018 provided a detailed analysis based on the 23 February 1956 extreme GLE. Appropriate mitigation measures, such as lowering aircraft altitude with minimal additional fuel consumption, have been shown to reduce radiation exposure during major events by up to 40% (Matthia et al., 2015).

Magnetospheric relativistic 1–10 MeV electrons, such as GCRs and SEPs, do not reach airplane altitudes. However, they form a "shower" by cascading into other particles like muons, gamma ray photons, electrons and positrons (Tsurutani et al., 2016). This radiation has just recently been realized to be a hazard to airplane pilots and personnel (Tobiska et al., 2018).

In November 2019, a new service for global civil aviation has been launched under the auspices of the International Civil Aviation Organization (ICAO), to provide real-time and worldwide space weather updates for commercial and general aviation. This service aims to generate and share space weather advisories using data collected from dedicated space weather centers established by 17 countries (https://www.icao.int/Newsroom/Pages/New-global-aviation-space-weather-network-launched.aspx).

5.4 Ionospheric Space Weather effects related to navigation and communication

It has been just 125 years since the invention of the first radio wave transmitter and receiver, but it is hard to imagine our life without radio communication devices. What if an extreme Space Weather event disrupted all of these systems?

The range of radiocommunication frequencies extends from a few 10s of kHz for ground-wave propagation to a 10s of GHz for satellite communication frequencies[10]. The choice of frequencies depends on the systems and their implementations. However, there are common underlying physical processes that control the propagation of radio waves in the ionosphere, and therefore control the effects.

1) The ionosphere has a layered structure, with a few layers of increased ionization. Each layer (D, E, F) is created by a particular ionizing source of photons from the Sun (UV, EUV, X-ray), and SEPs and auroral precipitation, thus naturally forming the dependence on Space Weather processes. Radio waves can only propagate at a frequency higher than the plasma frequency, which depends on the square root of plasma number density (electron number density). The plasma frequency for the Earth's ionosphere generally varies in the range of 2–20 MHz, defining the frequency range for HF communications. Therefore, any change in electron number density in the ionosphere caused by changes in the ionizing sources will have a significant effect on wave propagation in the radio HF range.
2) In the context of GNSS signals, the important quantity is an ionospheric delay, the error in estimating the Euclidian distance between the receiver and the s/c due to the group speed/phase delay of the wave packet in the ionosphere. More simply stated, the ionospheric delay occurs because the ionosphere affects the propagation path and speed of radio waves, making the former typically longer (due to refraction) and the latter less than in a vacuum. The ionospheric delay is related to the plasma density between the receiver and the s/c. The total electron column density is called the slant Total Electron Content (TEC), which in turn could be recalculated into vertical TEC, or VTEC. Global VTEC maps are needed to calculate ionospheric delay and GNSS signal propagation errors for users with single frequency receivers. Dual-frequency GNSS receiver data are used to calculate and remove the contribution of ionospheric delay using two of the broadcast frequencies from the GNSS satellites. When dual-frequency reception is not available, single-frequency GNSS receivers rely on ionospheric models and data from the dual-frequency ground reference station network to estimate VTEC values. Any sudden and localized changes that are not reflected in the model or in the VTEC measurements will cause errors for users of single frequency receivers. An example of a system that currently supports single-frequency user receivers is the US GPS-based Wide Area Augmentation System (WAAS), designed to assist civil aviation users (Sparks et al., 2022). This type of Space Weather hazard has potential to decrease in the future, when users are able to use dual-frequency receivers more widely, with the advent of a GPS constellation that transmits a second frequency.
3) Turbulence in the ionosphere can cause inhomogeneities in ionospheric plasma density that refract/diffract the incoming radio signals and produce interference patterns (Ghobadi et al., 2020). Inhomogeneities at scale sizes of the order of

10 Frequencies ~10 GHz and up are likely not very sensitive to space weather. See also https://www.alsa.mil/News/Article/2532178/true-impacts-of-space-weather-on-a-ground-force/for some examples.

a Fresnel zone (10s–100 s m) are effective[11]. In the context of GNSS signals, these patterns introduce rapid fluctuations in the wave phase and the amplitude, known as ionospheric scintillation. If the scintillations are large enough, they prevent the receiver from maintaining lock on an incoming signal, i.e., disrupting tracking of the signal by the receiver. Since scintillations are caused by small-scale fluctuations in ionospheric densities, it remains challenging to model and predict the effects of this phenomenon.

4) The coupling between electrons and neutral plasma at low altitudes is responsible for radio wave absorption in the D-layer (60–90 km) of the ionosphere. If the plasma density suddenly increases in this region where the electron-neutral collision frequency is higher than the wave frequency, then the waves propagating through these regions will be absorbed. This effect is called a 'sudden ionospheric disturbance' or SID, and is responsible for radio blackouts within the HF range caused by X-ray flares.
5) Solar radio bursts (SRBs) have the potential to directly interfere with radar systems[12] in a wide range of frequencies (Hey, 1946; Knipp et al., 2016; Marqué et al., 2018) and GNSS signals in the frequency range that falls within the range of the GNSS transmitters, 1–2 GHz (Klobuchar et al., 1999; Cerruti et al., 2008). This effect is less common than ionospheric scintillations, but could be profound for particularly large events, thus defining SRBs as an important process in Space Weather (Bala et al., 2002; Gary and Bastian, 2021). Solar radio bursts can last from a few hours to a few days. It is important to define what is the radio flux threshold in the GNSS frequency band that will introduce significant errors and disturbances for a given system.[13] Recent studies show that this threshold may be lower than originally thought (Yue et al., 2018).

Some of the most important Space Weather processes that control ionospheric variability are the UV/FUV/EUV/soft X-ray part of the solar radiation that comes from both background processes (active regions) and eruptive processes (energetic solar flares), particles of SEPs, and geomagnetic storms and substorms. Studies of energetic flares demonstrate that the combined effect of UV/FUV/EUV/soft Xray radiation could increase the ionospheric photoionization rate within minutes, thus introducing rapid ionospheric variations (Tsurutani et al., 2005).

On the nightside ionosphere, magnetospheric precipitation of electrons with energies 0.1–100 keV (Newell et al., 2009; Miyoshi et al., 2015) is the main source of ionospheric ionization and it is closely related to storm/substorm activity. Geomagnetic storms/substorms inject energy into the magnetosphere and ionosphere. It creates plasma instabilities and particle precipitation in the coupled thermosphere-ionosphere system, and irregularities in the ionospheric density. Complexity of the coupled system therefore defines the response over a wide range of spatial and temporal scales.

5.5 Thermosphere effects and satellite drag

The ionosphere is closely coupled with the upper Earth atmosphere, or thermosphere (~90–900 km) *via* collisions with neutrals. One major Space Weather effect of concern is the expansion of the upper atmosphere during magnetic storms where there is atmospheric heating in the auroral and subauroral zones. Collisions with neutrals produce an atmospheric drag effect for s/c in LEO. The result is a gradual reduction of s/c altitude, and, in extreme cases such as during intense magnetic storms, a loss of s/c track. An example could be the geomagnetic storm of March 1989, when after 5 days–1,000 satellites became "lost" for a few days (Fig. 7 from Hapgood, 2018).

The coupling between the thermosphere and ionosphere allows an exchange of momentum between the plasma and the neutral atmospheric species. As a result, global motion of the neutral component (e.g., winds) could drive plasma flows, and *vice versa*, strong plasma flows could drive neutral flows. This makes the variations of neutrals in the upper atmosphere closely coupled with Space Weather processes (Crowley and Azeem, 2018; Fuller-Rowell et al., 2018; Hapgood, 2018; Mannucci and Tsurutani, 2018). At the present time computer codes can model neutral-ion coupling, but there does not exist an adequate code for ion-neutral coupling for extreme storms like the Carrington event at present (Deng et al., 2018).

One important parameter that controls atmospheric expansion is ionospheric Joule heating (Bates, 1974). Joule heating is the main channel for magnetospheric energy dissipation in the ionosphere: energy deposition into ionosphere due to magnetospheric energetic particle precipitation is small in comparison with energy deposition due to Joule heating (Wilson et al., 2006). But precipitation controls ionospheric conductance (Robinson et al., 1987), and hence the Joule heating pattern. To calculate Joule heating it is necessary to know global distributions of ionospheric conductance and electric fields. This problem is still not solved completely, and uncertainty for estimations of Joule heating rate translates into uncertainty for satellite tracking.

Another important uplifting mechanism for neutrals that has recently gained attention is the "equatorial plasma super-fountain"

11 The Fresnel zone defines the distance scale over which destructive interference can occur.

12 The effects of solar radio bursts on radars were first presented in secret reports during World War II documenting widespread disturbances of British radar systems operating at 55–85 MHz on February 26–28, 1942 (Hey, 1946).

13 Specification of the threshold for radio burst interference is receiver dependent.

(Tsurutani et al., 2004b; Tsurutani et al., 2008; Mannucci et al., 2005). The global convection electric field during storms is known to penetrate to low ionospheric latitudes in the equatorial region (Nishida, 1968a; Nishida, 1968b). This is called the Prompt Penetration Electric Field (PPEF). It is thought that PPEF arises from insufficient shielding of the global convection electric field by the ring current *via* Region II Birkeland currents flowing in/out of the ionosphere (Toffoletto et al., 2003). Under quiet conditions, the PPEF is weak but intensifies during the storm main phase. Estimates show that during strong geomagnetic storms the PPEF will cause the dayside upper atmosphere to be uplifted through ion-neutral drag in the near-equatorial regions (Lakhina and Tsurutani, 2017).

There is an important Space Weather process affecting neutral density variations that is not related in general to geomagnetic storms. Increased solar UV, FUV, and EUV radiation directly heats the thermosphere, therefore variabilities in these quantities will also cause atmospheric expansion and uplift. Increased levels of solar radiation are linked to the group of sunspots, active regions, rotating with the Sun and therefore lasting a few weeks. Estimates show that a moderate increase in EUV/FUV level could increase the thermospheric neutral density at 400 km and 850 km by 100% and 200%, correspondingly (Fuller-Rowell et al., 2018).

Space Weather effects on satellite tracking is one of risks that evolve in time and currently is increasing. This is related to a more general problem of collisions with space debris. It is currently a real challenge since there are thousands of satellites being launched into LEO (e.g., Starlink). A concern is that an extreme Space Weather event similar to what happened in the past would increase satellite drag unpredictably. This in turn would increase the risk of collisions with space debris/other satellites and start the Kessler effect or 'domino effect' when each subsequent collision increases chances for the next one (Kessler and Cour-Palais, 1978).

To summarize, the factors controlling satellite drag effects and influencing satellite orbit prediction and collision prevention are: 1) neutral winds as they affect s/c orbital velocity and also affect the coupling between neutrals and plasma in general; 2) neutral composition as it modulates the expansion by changing scale heights; 3) global plasma flows horizontally and vertically, because they translate into Joule heating rates and drive motions of neutrals (through ion-neutral collisions); 4) modulation of global heating rate by radiative nitric oxide (NO) cooling (e.g., Knipp et al., 2017); 5) variations in solar UV, FUV and EUV radiation related to Sun's active regions due to direct heating of the neutrals; 6) the Kessler effect in the case of a hypothetical extreme Space Weather event.

6 Discussion: Major challenges and the worst-case scenario

Given the growing importance of the Space Weather effects for a technologically advanced society, it is important to understand and anticipate the associated risks and dangers. Space Weather regularly affects satellite communications and avionics, for example with partial degradation of GNSS services or planned gradual radiation damage to satellites. Society is developing mitigation techniques and in parallel a better understanding of the underlying physical processes. However, there is one very important area where we still have significant fundamental knowledge gaps. This area of concern relates to the prediction of extreme solar eruptive events and their interactions with Earth.

Solar eruptive phenomena result from the constantly changing solar magnetic field, plasma instabilities and finally from magnetic reconnection in the solar corona. However, at the present it is unclear how reconnection controls the ejection of CMEs (e.g., Gou et al., 2019; Zhu et al., 2020). As a result, it is very difficult, if not impossible, to predict whether a group of sunspots will cause an event on Earth of similar intensity to the Carrington event. One principal difficulty is predicting the direction of the CME magnetic field and if it remains in the same direction as it propagates from the Sun to the Earth (Tsurutani et al., 2020). To add complexity, it should be noted that the Carrington event was not accompanied by an intense radiation storm (Wolff et al., 2012). Solar flares, radiation storms and geomagnetic storms are related Space Weather phenomena, but can occur together or independently.

With growing data sets available, it seems possible that soon we will have statistical and/or AI-based models accurate enough to predict that an energetic flare will occur with some lead time so society can implement mitigation measures. Then the next step is to predict the occurrence of a radiation storm on Earth, International Space Station, the Moon or Mars, if the area of concern also includes human space flight missions. This is not easily obtainable in general, since there is not a simple cause-effect relationship between an intense flare and radiation storms. Currently it is difficult for physics-based models of the heliosphere to accurately predict the propagation of a CME through the interplanetary medium or make an accurate prognosis for SEPs (again see Tsurutani et al., 2020 for detailed discussions).

We note that accurate prediction of CME trajectory, CME arrival time and CME properties (e.g., profiles of magnetic field near the Earth) is widely recognized as a strategic goal for the international space weather community. As noted by Riley et al. (2018), progress is slow. However, results from recent works that combine physics-based models, data assimilation and machine learning methods are cautiously optimistic (Amerstorfer et al., 2021; Alobaid et al., 2022; Kay et al., 2022).

McIntosh et al. (2015) showed examples where bursts of extreme solar activity occur as a changing pattern of solar magnetism with characteristic period 6–18 months. These bursts have been attributed to the existence of solar Rossby waves that are formed in the. Solar 'tachocline', a region of solar interior that separates the differentially rotating convection zone from solidly rotating radiative zone (see the review by

Dikpati and McIntosh, 2020). Recent advances in modeling and understanding of solar Rossby waves, coupled with advances in data assimilation techniques, promises to deliver space weather forecasting on a time scale from a few weeks to a few years (Dikpati and McIntosh, 2020). In addition to the Rossby waves hypothesis, there are other recent developments to understand solar cycle variability and quantify the likelihood of severe space weather events during different phases of solar cycle (Chapman et al., 2020; Leamon et al., 2022).

What would be the worst-case scenario? How strong would the greatest geomagnetic storm be? Radiation storm? Geomagnetically induced currents? How would this affect the avionics/satellite electronics? Telecommunications? GNSS services? Power grid infrastructure? These questions are becoming important as different countries begin to realize the threats and dangers associated with extreme Space Weather conditions. As stated in recently submitted report from the UN Expert Group on Space Weather (UN-EGSW)[14], 'improved international cooperation and coordination can lead to improved global resilience and preparedness in response to the adverse impacts of space weather.' Estimations made by the Royal Academy of Engineering in the UK show that an extreme space weather event will be difficult to deal with, but not catastrophic (Cannon et al., 2013; Hapgood et al., 2021). But we note that our current knowledge is based on limited statistics of extreme events that have happened in the past.

Could magnetic storms more extreme than the Carrington event occur in the future? The answer is yes, but a following question is how much larger? Using the maximum observed CME speed of 3,000 km/s, Tsurutani and Lakhina (2014) showed that it is theoretically possible to have a magnetic storm that is twice the intensity of the Carrington event. However, the authors mentioned a caveat, "if the magnetosphere does not saturate at β = 1.0" (β is the ratio of plasma thermal pressure to magnetic pressure).

In August 1972 a solar active region launched several CMEs in the direction of the Earth. Although the CME that impacted the Earth was faster than the Carrington CME, the 1972 CME had northward IMFs so the resultant geomagnetic storm was small (Tsurutani et al., 1992). However, at the same time the AT&T L4 telecommunications cable line was down for an hour and the radiation flux was one of the highest during the space age (Lanzerotti, 1992). This was a case of extreme Space Weather effects (CME speed, radiation and GIC effects) with a low magnetic storm intensity. All of these effects were related but obviously not strongly correlated in intensities.

On 23 July 2012 the Sun released an extremely fast CME with a transit time to one au of 19 h, slightly slower than the 17 h 40 min of the Carrington event. Baker et al. (2013) estimated that this could have produced a magnetic storm of peak intensity Dst = -1,182 nT. Luckily, the 2012 CME missed the Earth!

What can we learn from other stellar systems? Data from solar-type stars could help to accumulate necessary statistics on extreme events and estimate worst-case scenarios for our Sun. Maehara et al. (2012) studied "superflares" at solar-type stars. These superflares are 10 to 1,000 times more intense than the most intense flares recorded on our Sun. Extending these previous works, Okamoto et al. (2021) analyzed all of the Kepler primary mission data, applying the method of Notsu et al. (2019), and suggested that the Sun can cause superflares with energies of ~7 × 10^{33} erg (~X700-class flares) and ~1 × 10^{34} erg (~X1000-class flares) once every ~3,000 and ~6,000 years, respectively.

Takahashi et al. (2016) showed that the upper limit of the CME speed associated with X1,000 and X100 superflares could be 9,100 km/s and 6,200 km/s, respectively (see Fig. 5 of Takahashi et al., 2016). Takahashi and Shibata (2017) further showed that these superflares are expected to produce massive CMEs (M ~ 6 × 10^{17} g for X1,000 flares) so that the deceleration between the Sun and the Earth would be small. That means that giant sunspots at our Sun could possibly produce a superflare and a CME with potentially disastrous Space Weather consequences (Shibata et al., 2013).

Miyake et al. (2013) and Usoskin et al. (2013) have studied tree ring and ice core data and have concluded that enormous radiation events or a series of events have occurred in 774–775 AD and 992–993 AD. If a similar event happens today, what would be the consequences? We know some of the answers, but they are incomplete. It is very important to identify the realistic worst case, along with the probability of them happening, and the potential impacts assessed. Then we need to make informed decisions on what level of risk we should take as a society, or as a business, and at what cost. It is our responsibility to understand and determine the variability and extremes of the natural environment and calculate their consequences.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

For NB this work has been partially supported by NASA grant 80NSSC19K0085. Portions of this research was performed at the Jet Propulsion Laboratory, California Institute of Technology under contract with NASA.

14 https://www.unoosa.org/oosa/oosadoc/data/documents/2022/aac.105c.1l/aac.105c.1l.401_0.html See also Mann et al., 2018.

Acknowledgments

NB thanks ISSI in Bern Switzerland for support of the International Team #523 "Imaging the Invisible: Unveiling the Global Structure of Earth's Dynamic Magnetosphere". BT thanks ISSI in Bern Switzerland for support of the International Team #455 on "Complex Systems Perspectives Pertaining to the Research of the Near-Earth Electromagnetic Environment". The authors thank P. F. Chen, R. Hajra, R. B. Horne, G. S. Lakhina, A. J. Mannucci, and K. Shibata for their reviews of the draft version of this paper. What was particularly useful was their expertise in specific detailed areas of Space Weather. Their comments were invaluable.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akasofu, S.-I. (1964). The development of the auroral substorm. *Plan. Spa. Phys.* 12, 273–282. doi:10.1016/0032-0633(64)90151-5

Alobaid, K. A., Abduallah, Y., Wang, J. T. L., Wang, H., Jiang, H., Xu, Y., et al. (2022). Predicting CME arrival time through data integration and ensemble learning. *Front. Astron. Space Sci.* 9, 1013345. doi:10.3389/fspas.2022.1013345

Amerstorfer, T., Hinterreiter, J., Reiss, M. A., Möstl, C., Davies, J. A., Bailey, R. L., et al. (2021). Evaluation of CME arrival prediction using ensemble modeling based on heliospheric imaging observations. *Space weather.* 19, e2020SW002553. doi:10.1029/2020SW002553

Anderson, K. A., and Milton, D. W. (1964). Balloon observations of X rays in the auroral zone: 3. High time resolution studies. *J. Geophys. Res.* 69, 4457–4479. doi:10.1029/jz069i021p04457

Axford, W. I., and Hines, C. O. (1961). A unifying theory of high-latitude geophysical phenomena and geomagnetic storms. *Can. J. Phys.* 39, 1433–1464. doi:10.1139/p61-172

Baker, D. N., Li, X., Pulkkinen, A., Ngwira, C. M., Mays, M. L., Galvin, A. B., et al. (2013). A major solar eruptive event in July 2012: Defining extreme space weather scenarios. *Space weather.* 11, 585–591. doi:10.1002/swe.20097

Baker, D. N., Erickson, P. J., Fennell, J. F., Foster, J. C., Jaynes, A. N., and Verronen, P. T. (2018). Space weather effects in the Earth's radiation belts. *Space Sci. Rev.* 214 (17), 17. doi:10.1007/s11214-017-0452-7

Bala, B., Lanzerotti, L. J., Gary, D. E., and Thomson, D. J. (2002). Noise in wireless systems produced by solar radio bursts. *Radio Sci.* 37 (2), 2-1–2-7. doi:10.1029/2001RS002481

Band, P. R., Spinelli, J. J., Ng, V. T. Y., Math, M., Moody, J., and Gallagher, R. P. (1990). Mortality and cancer incidence in a cohort of commercial airline pilots. *Aviat. Space Environ. Med.* 61 (4), 299–302.

Barish, R. J. (1990). Health physics concerns in commercial aviation. *Health Phys.* 59, 199–204. doi:10.1097/00004032-199008000-00006

Barlow, W. H. (1849). On the spontaneous electrical currents observed in the wires of the electric telegraph. *Philosophical Trans. R. Soc. Lond.* 139, 61–72.

Barnaby, H., and Marinella, M. (2013). *Total ionizing dose and displacement damage effects in embedded memory technologies (tutorial presentation).* US: National Nuclear Security Administration. Available at: https://www.osti.gov/servlets/purl/1115539.

Bates, H. F. (1974). Atmospheric expansion from Joule heating. *Planet. Space Sci.* 22, 925–937. ISSN 0032-0633. doi:10.1016/0032-0633(74)90162-7

Baumann, R. C. (2005). Radiation-induced soft errors in advanced semiconductor technologies. *IEEE Trans. Device Mater. Reliab.* 5, 305–316. doi:10.1109/TDMR.2005.853449

Birkeland, K. (1908). "The polar elementary storms," in *The Norwegian aurora Polaris expedition 1902–1903, vol. 1* (Christiania (now Oslo), Norway: H. Aschehoug & Co), 84–143.

Boteler, D. H. (2006). The super storms of August/September 1859 and their effects on the telegraph system. *", Adv. Space Res.* 38 (2), 159–172. doi:10.1016/j.asr.2006.01.013

Boteler, D. H., and Jansen van Beek, G. (1999). August 4, 1972 revisited: A new look at the geomagnetic disturbance that caused the L4 cable system outage. *Geophys. Res. Lett.* 26 (5), 577–580. doi:10.1029/1999gl900035

Bothmer, V., and Daglis, I. A. (2007). *'Space weather – physics and effects.* New York, NY USA: Published by Praxis Publishing, Chichester, UK, and Springer Science+Business Media, LLC. ISBN 978-3-540-23907-9. doi:10.1007/978-3-540-34578-7

Brown, R. R., Barcus, J. R., and Parsons, N. R. (1965). Balloon observations of auroral zone X-rays in conjugate regions, 2. Microbursts and pulsations. *J. Geophys. Res.* 70, 2599–2612. doi:10.1029/jz070i011p02599

Burlaga, L. F., Sittler, E., Mariani, F., and Schwenn, R. (1981). Magnetic loop behind an interplanetary shock: Voyager, Helios, and IMP 8 observations. *J. Geophys. Res.* 86, 6673–6684. doi:10.1029/ja086ia08p06673

Burlaga, L. F., Klein, L., Sheeley, N. R., Michels, D. J., Howard, R. A., Koomen, M. J., et al. (1982). A magnetic cloud and a coronal mass ejection. *Geophys. Res. Let.* 9, 1317–1320. doi:10.1029/gl009i012p01317

N. Buzulukova (Editor) (2018). *Extreme events in Geospace, origins, predictability and consequences* (Wash. DC: Elsevier).

Buzulukova, N., Fok, M.-C., Glocer, A., Komar, C., Kang, S.-B., Martin, S., et al. (2018) 'Geomagnetic storms: First-principles models for extreme Geospace environment', Editor(s): N. Buzulukova, *Extreme events in Geospace*, Elsevier, 231–258. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00010-8

Byrne, J., Maloney, S., McAteer, R., Refojo, J. M., and Gallagher, P. T. (2010). Propagation of an Earth-directed coronal mass ejection in three dimensions. *Nat. Commun.* 1, 74. doi:10.1038/ncomms1077

Cannon, P., Angling, M., Barclay, L., Curry, C., Dyer, C., Edwards, R., et al. (2013). *Extreme space weather: Impacts on engineered systems and infrastructure.* London: Royal Academy of Engineering. ISBN 1-903496-95-0. Available at: https://www.raeng.org.uk/publications/reports/space-weather-full-report.

Carrington, R. C. (1859). Description of a singular appearance seen in the Sun. *Monthly Notices of the Royal Astronomical Society* 20 (1), 13–15. doi:10.1093/mnras/20.1.13

Carter, B. A., Yizengaw, E., Pradipta, R., Halford, A. J., Norman, R., and Zhang, K. (2015). Interplanetary shocks and the resulting geomagnetically induced currents at the equator. *Geophys. Res. Lett.* 42, 6554–6559. doi:10.1002/2015GL065060

Cerruti, A. P., Kintner, P. M., Gary, D. E., Mannucci, A. J., Meyer, R. F., Doherty, P., et al. (2008). Effect of intense December 2006 solar radio bursts on GPS receivers. *Space weather.* 6, S10D07. doi:10.1029/2007SW000375

Chapman, S., and Bartels, J. (1940). *Geomagnetism*, I. New York: Oxford Univ. Press, 328–337.

Chapman, S. C., McIntosh, S. W., Leamon, R. J., and Watkins, N. W. (2020). Quantifying the solar cycle modulation of extreme space weather. *Geophys. Res. Lett.* 47, e2020GL087795. doi:10.1029/2020GL087795

Chen, P. F. (2011). Coronal mass ejections: Models and their observational basis. *Living Rev. Sol. Phys.* 8, 1. doi:10.12942/lrsp-2011-1

Chree, C. (1913). Some phenomena of sunspots and of terrestrial magnetism at kew observatory', philosophical. *Trans. R. Soc. Lond. Ser. A* 212, 75–116. doi:10.1098/rsta.1913.0003

Clauer, C. R., and Siscoe, G. (2006). The great historical geomagnetic storm of 1859: A modern look. *Adv. Space Res.* 38 (2), 117–118. ISSN 0273-1177. doi:10.1016/j.asr.2006.09.001

Coster, A. J., Erickson, P. J., Lanzerotti, L. J., Zhang, Y., and Paxton, L. J. (2021). *Space weather effects and applications'*. Washington D. C.: AGU Geophysical Monograph Series, 262. doi:10.1002/9781119815570

Cranmer, S. R. (2009). Coronal holes. *Living Rev. Sol. Phys.* 6, 3. doi:10.12942/lrsp-2009-3

Crowley, G., and Azeem, I. (2018) 'Extreme ionospheric storms and their effects on GPS systems', Editor(s): N. Buzulukova, *Extreme events in Geospace*, Elsevier, Pages 555–586. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00023-6

Daglis, I. A. (2001). The storm-time ring current. *Space Sci. Rev.* 98, 343–363. doi:10.1023/a:1013873329054

Dang, T., Li, X., Luo, B., Li, R., Zhang, B., Pham, K., et al. (2022). Unveiling the space weather during the Starlink satellites destruction event on 4 February 2022. *Spa. Weath.* 20, e2022SW003152. doi:10.1029/2022SW003152

De Zeeuw, D. L., Sazykin, S., Wolf, R. A., Gombosi, T. I., Ridley, A. J., and To´th, G. (2004). Coupling of a global MHD code and an inner magnetospheric model: Initial results. *J. Geophys. Res. (Space Phys.)* 109, A12219. doi:10.1029/2003JA010366

DeForest, S. E. (1972). Spacecraft charging at synchronous orbit. *J. Geophys. Res.* 77 (4), 651–659. doi:10.1029/JA077i004p00651

Deng, Y., Sheng, C., Tsurutani, B. T., and Mannucci, A. J. (2018). 'Possible influence of extreme magnetic storms on the thermosphere in the high latitudes. *Space weather.* 16, 802–813. doi:10.1029/2018SW001847

Desai, M., and Giacalone, J. (2016). Large gradual solar energetic particle events. *Living Rev. Sol. Phys.* 13, 3. doi:10.1007/s41116-016-0002-5

Dessler, A. J., and Parker, E. N. (1959). Hydromagnetic theory of geomagnetic storms. *J. Geophys. Res.* 64 (12), 2239–2252. doi:10.1029/JZ064i012p02239

Dikpati, M., and McIntosh, S. W. (2020). Space weather challenge and forecasting implications of Rossby waves. *Space weather.* 18, e2018SW002109. doi:10.1029/2018SW002109

Dobynde, M. I., Shprits, Y. Y., Drozdov, A. Y., Hoffman, J., and Li, J. (2021). 'Beating 1 sievert: Optimal radiation shielding of astronauts on a mission to Mars. *Space weather.* 19, e2021SW002749. doi:10.1029/2021SW002749

Du, A. M., Tsurutani, B. T., and Sun, W. (2008). Anomalous geomagnetic storm of 21-22 january 2005: A storm main phase during northward IMFs. *J. Geophys. Res.* 113, 284. doi:10.1029/2008Ja013284

Dungey, J. W. (1961). Interplanetary magnetic field and the auroral zones. *Phys. Rev. Lett.* 6, 47–48. doi:10.1103/physrevlett.6.47

Dyer, C. S., Sims, A. J., Farren, J., and Stephen, J. (1989). Measurements of the SEU environment in the upper atmosphere. *IEEE Trans. Nucl. Sci.* 36 (6), 2275–2280. doi:10.1109/23.45435

Dyer, C., Hands, A., Ryden, K., and Lei, F. (2018). Extreme atmospheric radiation environments and single event effects. *IEEE Trans. Nucl. Sci.* 65 (1), 432–438. doi:10.1109/TNS.2017.2761258

Echer, E., Gonzalez, W. D., and Tsurutani, B. T. (2008). Interplanetary conditions leading to superintense geomagnetic storms (Dst < - 250 nT) during solar cycle 23. *Geophys. Res. Lett.* 35, L06S03. doi:10.1029/2007GL031755

Egeland, A., and Burke, W. J. (2010). Kristian Birkeland's pioneering investigations of geomagnetic disturbances. *Hist. Geo Space. Sci.* 1, 13–24. doi:10.5194/hgss-1-13-2010

Eroshenko, E. A., Belov, A. V., Boteler, D., Gaidash, S. P., Lobkov, S. L., Pirjola, R., et al. (2010). Effects of strong geomagnetic storms on Northern railways in Russia. *Adv. Space Res.* 46, 1102–1110. doi:10.1016/j.asr.2010.05.017

Fang, T.-W., Kubaryk, A., Goldstein, D., Fuller-Rowell, T., Millward, G., Singer, H. J., et al. (2022). Space weather environment during the SpaceX Starlink satellite loss in February 2022. *Space weather.* 20, e2022SW003193. doi:10.1029/2022SW003193

Ferguson, D. C. (2018) 'Extreme space weather spacecraft surface charging and arcing effects', Editor(s): N. Buzulukova, *Extreme events in Geospace*, Elsevier, 401–418. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00015-7

Ferguson, D. C., and Hillard, G. B. (2003). *Low Earth orbit spacecraft charging design guidelines*. NASA/TP—2003-212287.

Floyd, L., Newmark, J., Cook, J., Herring, L., and McMullin, D. (2005). Solar EUV and UV spectral irradiances and solar indices. *J. Atmos. Solar-Terrestrial Phys.* 67, 3–15. doi:10.1016/j.jastp.2004.07.013

Fok, M. C., Moore, T. E., Brandt, P. C., Delcourt, D. C., Slinker, S. P., and Fedder, J. A. (2006). Impulsive enhancements of oxygen ions during substorms. *J. Geophys. Res. (Space Phys.)* 111, A10222. doi:10.1029/2006ja011839

Forbes, T. (2000). A review on the Genesis of coronal mass ejections. *J. Geophys. Res.* 105, 23153–23165. doi:10.1029/2000JA000005

Frey, H. U., Mende, S. B., Angelopoulos, V., and Donovan, E. F. (2004). Substorm onset observations by IMAGE-FUV. *J. Geophys. Res.* 109, A10304. doi:10.1029/2004JA010607

Friedberg, W., and Copeland, K. (2003). *What aircrews should know about their occupational exposure to ionizing radiation ionizing*. Oklahoma City, OK: Federal Aviation Administration, Civil Aerospace Medical Institute. DOT Report No. DOT/FAA/AM-03/16.

Friedberg, W., and Copeland, K. (2011). *Ionizing radiation in earth's atmosphere and in space near earth*. Oklahoma City, OK: Federal Aviation Administration, Civil Aerospace Medical Institute. DOT Report No. DOT/FAA/AM-11/9.

Fuller-Rowell, T., Emmert, J., Fedrizzi, M., Weimer, D., Codrescu, M. V., Pilinski, M., et al. (2018) 'How might the thermosphere and ionosphere react to an extreme space weather event?', Editor(s): N. Buzulukova, *Extreme events in Geospace*, Elsevier, 513–539. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00021-2

Ganushkina, N. Y., Swiger, B., Dubyagin, S., Matéo-Vélez, J.-C., Liemohn, M. W., Sicard, A., et al. (2021). Worst-case severe environments for surface charging observed at LANL satellites as dependent on solar wind and geomagnetic conditions. *Space weather.* 19, e2021SW002732. doi:10.1029/2021SW002732

Garrett, H. B., and Whittlesey, A. C. (2012). *Guide to mitigating spacecraft charging effects*. Hoboken, NJ: John Wiley & Sons. doi:10.1002/9781118241400

Gary, D. E., and Bastian, T. S. (2021). "Solar radio burst effects on radio- and radar-based systems," in *Space weather effects and applications*. Editors A. J. Coster, P. J. Erickson, L. J. Lanzerotti, Y. Zhang, and L. J. Paxton. doi:10.1002/9781119815570.ch6

Gaunt, C., and Coetzee, G. (2007). "Transformer failures in regions incorrectly considered to have low GIC-risk," in Proceeding of the 2007 IEEE Lausanne Power Tech, Switzerland, July 2007 (IEEE), 807–812. doi:10.1109/PCT.2007.4538419

Ghobadi, H., Spogli, L., Alfonsi, L., Cesaroni, C., Cicone, A., Linty, N., et al. (2020). Disentangling ionospheric refraction and diffraction effects in GNSS raw phase through fast iterative filtering technique. *GPS Solut.* 24, 85. doi:10.1007/s10291-020-01001-1

Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., et al. (1994). What is a geomagnetic storm? *J. Geophys. Res.* 99, 5771–5792. doi:10.1029/93JA02867

Gonzalez, W. D., and Tsurutani, B. T. (1987). Criteria of interplanetary parameters causing intense magnetic storms (Dst < -100 nT). *Planet. Space Sci.* 35, 1101–1109. doi:10.1016/0032-0633(87)90015-8

Gou, T., Liu, R., Kliem, B., Wang, Y., and Veronig, A. M. (2019). The birth of A coronal mass ejection. *Sci. Adv.* 5, 7004. doi:10.1126/sciadv.aau7004

Hajra, R., Echer, E., Tsurutani, B. T., and Gonzalez, W. D. (2013). Solar cycle dependence of High-Intensity Long-Duration Continuous AE Activity (HILDCAA) events, relativistic electron predictions? *J. Geophys. Res.* 118, 5626–5638. doi:10.1002/jgra.50530

Hajra, R., Tsurutani, B. T., Echer, E., Gonzalez, W. D., and Santolik, O. (2015). Relativistic (E >0.6, >2.0, and >4.0 MeV) electron acceleration at geosynchronous orbit during high-intensity longduration continuous AE activity (HILDCAA) events. *Astrophysical J.* 799, 39. doi:10.1088/0004-637X/799/1/39

Hajra, R., and Tsurutani, B. T. (2018). "Magnetospheric 'killer' relativistic electron dropouts (REDs) and repopulation: A cyclical process," in *Extreme events in Geospace: Origins, predictability, and consequences*. Editor N. Buzulukova (Elsevier), 373–400. doi:10.1016/B978-0-12-812700-1.00014-5

Hanslmeier, A. (2007). *The sun and space weather*. edition 2. Springer Netherlands. doi:10.1007/978-1-4020-5604-8

Hapgood, M., Angling, M. J., Attrill, G., Bisi, M., Cannon, P. S., Dyer, C., et al. (2021). 'Development of space weather reasonable worst-case scenarios for the UK National Risk Assessment. *Space weather.* 19, e2020SW002593. doi:10.1029/2020SW002593

Hapgood, M. (2018) 'Linking space weather science to impacts - the view from the Earth', Editor(s): N. Buzulukova, *Extreme events in Geospace*, Elsevier, 3–34. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00001-7

Hapgood, M. (2019). "The impact of space weather on human missions to Mars: The need for good engineering and good forecasts," in *The human factor in a mission to Mars* (Cham: Springer), 69–91. doi:10.1007/978-3-030-02059-0_4

Harvey, J. W., and Sheeley, N. R., Jr (1979). Coronal holes and solar magnetic fields. *Space Sci. Rev.* 23, 139–158. doi:10.1007/bf00173808

Hathaway, D. H. (2015). The solar cycle. *Living Rev. Sol. Phys.* 12, 4. doi:10.1007/lrsp-2015-4

Hayakawa, H., Tamazawa, H., Kawamura, A. D., and Isobe, H. (2015). Records of sunspot and aurora during CE 960–1279 in the Chinese chronicle of the Sòng dynasty. *Earth Planet Sp.* 67, 82. doi:10.1186/s40623-015-0250-y

Hey, J. (1946). Solar radiations in the 4–6 metre radio wave-length band. *Nature* 157, 47–48. doi:10.1038/157047b0

Hodgson, R. (1859). On a curious appearance seen in the Sun. *Mon. Notices R. Astronomical Soc.* 20, 15–16. doi:10.1093/mnras/20.1.15a

Jiao, Z., Sun, H., Wang, X., Manchester, W., Gombosi, T., Hero, A., et al. (2020). 'Solar flare intensity prediction with machine learning models. *Space weather.* 18, e2020SW002440. doi:10.1029/2020SW002440

Kamide, Y., and Kusano, K. (2015). No major solar flares but the largest geomagnetic storm in the present solar cycle. *Space weather.* 13, 365–367. doi:10.1002/2015SW001213

Kay, C., Mays, M. L., and Collado-Vega, Y. M. (2022). Osprei: A coupled approach to modeling CME-driven space weather with automatically generated, user-friendly outputs. *Space weather.* 20, e2021SW002914. doi:10.1029/2021SW002914

Keiling, A., Ramos, C., Vu, N., Angelopoulos, V., and Nosé, M. (2022). Statistical properties and proposed source mechanism of recurrent substorm activity with one-hour periodicity. *J. Geophys. Res. Space Phys.* 127, e2021JA030064. doi:10.1029/2021JA030064

Kennel, C. F., Scarf, F. L., Coroniti, F. V., Russell, C. T., Wenzel, K. P., Sanderson, T. R., et al. (1984). Plasma and energetic particle structure upstream of a quasi-parallel interplanetary shock. *J. Geophys. Res.* 89, 5419–5435. doi:10.1029/JA089iA07p05419

Kennel, C. F., Edmiston, J. P., and Hada, T. (1985). "A quarter century of collisionless shock research," in *Collisionless shocks in the heliosphere: A tutorial review*. Editors R. G. Stone and B. T. Tsurutani. doi:10.1029/GM034p0001

Kepko, L., McPherron, R. L., Amm, O., Apatenkov, S., Baumjohann, W., Birn, J., et al. (2015). Substorm current wedge revisited. *Space Sci. Rev.* 190, 1–46. doi:10.1007/s11214-014-0124-9

Kessler, D. J., and Cour-Palais, B. G. (1978). Collision frequency of artificial satellites: The creation of a debris belt. *J. Geophys. Res.* 83, 2637–2646. doi:10.1029/JA083iA06p02637

Kimball, D. S. (1960). A study of the aurora of 1859. *Sci. Rept.* 6, UAGR109. Univ. of Alaska, Fairbanks, Alaska, 1960. Available at: https://scholarworks.alaska.edu/handle/11122/3607.

Klein, L. W., and Burlaga, L. F. (1982). Interplanetary magnetic clouds at 1 AU. *J. Geophys. Res.* 87, 613. doi:10.1029/ja087ia02p00613

Klobuchar, J., Kunches, J., and VanDierendonck, A. (1999). Eye on the ionosphere: Potential solar radio burst effects on GPS signal to noise. *GPS Solutions* 3, 69–71. doi:10.1007/PL00012794

Knipp, D. J., Ramsay, A. C., Beard, E. D., Boright, A. L., Cade, W. B., Hewins, I. M., et al. (2016). The May 1967 great storm and radio disruption event: Extreme space weather and extraordinary responses. *Space weather.* 14, 614–633. doi:10.1002/2016SW001423

Knipp, D. J., Pette, D. V., Kilcommons, L. M., Isaacs, T. L., Cruz, A. A., Mlynczak, M. G., et al. (2017). Thermospheric nitric oxide response to shock-led storms. *Space weather.* 15, 325–342. doi:10.1002/2016SW001567

Koskinen, H. (2011). *Physics of space storms: From the solar surface to the earth*. Edition 1. Berlin: Springer-Verlag. doi:10.1007/978-3-642-00319-6

Kozyra, J. U., Jordanova, V. K., Horne, R. B., and Thorne, R. M. (1997). "Modeling of the contribution of electromagnetic ion cyclotron (EMIC) waves to stormtime ring current erosion,". *Magnetic storms, geophys. Monogr. Ser.* Editor B. T. Tsurutani, et al. (Washington, D. C: AGU), 98, 187.

Lai, S. T., Cahoy, K., Lohmeyer, W., Carlton, A., Aniceto, R., and Minow, J. (2018). "'Deep dielectric charging and spacecraft anomalies," in *Extreme events in Geospace*. Editor N. Buzulukova (Elsevier), 419–432. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00016-9

Lakhina, G. S., Alex, S., Tsurutani, B. T., and Gonzalez, W. D. (2005). "Research on historical records of geomagnetic storms," in *Coronal and stellar mass ejections, proceedings of the 226th symposium of the international astronomical union held in beijing, China, september 13–17, 2004*. Editors K. P. Dere, J. Wang, and Y. Yan (Cambridge, U. K: Cambridge Univ. Press), 3–15.

Lakhina, G. S., Alex, S., Tsurutani, B. T., and Gonzalez, W. D. (2012). "Super magnetic storms: Hazard to society,". *Extreme events and natural hazards: The complexity perspective, geophys. Mon. Ser.* Editors A. S. Sharma, A. Bunde, V. P. Dimri, and D. N. Baker (Washington: AGU), 196, 267. doi:10.1029/2011GM001073

Lakhina, G. S., Hajra, R., and Tsurutani, B. T. (2020). "Geomagnetically induced currents," in *Encyclopedia of solid Earth geophysics, encyclopedia of Earth sciences series*. Editor H. K. Gupta (Springer Nature Switzerland AG). doi:10.1007/978-3-030-10475-7_245-1

Lakhina, G. S., and Tsurutani, B. T. (2016). Geomagnetic storms: Historical perspective to modern view. *Geosci. Lett.* 3, 5. doi:10.1186/s40562-016-0037-4

Lakhina, G. S., and Tsurutani, B. T. (2017). Satellite drag effects due to uplifted oxygen neutrals during super magnetic storms. *Nonlin. Process. Geophys.* 24, 745–750. doi:10.5194/npg-24-745-2017

Lakhina, G. S., and Tsurutani, B. T. (2018). "Super-geomagnetic storms: Past, present, and future," in *Extreme events in Geospace*. Editor N. Buzulukova (Elsevier), 157–185. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00007-8

Lanzerotti, L. J. (1992). Comment on 'Great magnetic storms' by Tsurutani et al. *Geophys. Res. Lett.* 19, 1991–1992. doi:10.1029/92GL02238

Lanzerotti, L. J. (2017). Space weather: Historical and contemporary Perspectives. *Space Sci. Rev.* 212, 1253–1270. doi:10.1007/s11214-017-0408-y

Leamon, R. J., McIntosh, S. W., and Title, A. M. (2022). Deciphering solar magnetic activity: The solar cycle clock. *Front. Astron. Space Sci.* 9, 886670. doi:10.3389/fspas.2022.886670

Leka, K. D., and Barnes, G. (2018). "Solar flare forecasting: Present methods and challenges," in *Extreme events in Geospace*. Editor N. Buzulukova (Elsevier), 65–98. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00003-0

Lockwood, M., and Hapgood, M. (2007). The rough guide to the Moon and Mars. *Astronomy Geophys.* 48 (6), 6.11–6.17. ISSN 13668781. doi:10.1111/j.14684004.2007.48611.x

Lohmeyer, W. Q., and Cahoy, K. (2013). Space weather radiation effects on geostationary satellite solid-state power amplifiers. *Space weather.* 11, 476–488. doi:10.1002/swe.20071

Loomis, E. (1861). On the great auroral exhibition of Aug. 28th to Sept. 4th, 1859 and on auroras generally; 8th article. *Am. J. Sci.* 32 (96), 318–335. doi:10.2475/ajs.s2-32.96.318

Love, J. J. (2014). Auroral Omens of the American civil war. *Weatherwise* 67 (5), 34–41. doi:10.1080/00431672.2014.939912

Lozinski, A. R., Horne, R. B., Glauert, S. A., Del Zanna, G., and Claudepierre, S. G. (2021). Modeling inner proton belt variability at energies 1 to 10 MeV using BAS-PRO. *J. Geophys. Res. Space Phys.* 126, e2021JA029777. doi:10.1029/2021JA029777

Lukianova, R., and Mursula, K. (2011). Changed relation between sunspot numbers, solar UV/EUV radiation and TSI during the declining phase of solar cycle 23. *J. Atmos. Solar-Terrestrial Phys.* 73, 235–240. ISSN 1364-6826. doi:10.1016/j.jastp.2010.04.002

Maden, N. (2020). Historical aurora borealis catalog for Anatolia and Constantinople (hABcAC) during the Eastern Roman Empire period: implications for past solar activity. *Ann. Geophys.* 38, 889–899. doi:10.5194/angeo-38-889-2020

Maehara, H., Shibayama, T., Notsu, S., Notsu, Y., Nagao, T., Kusaba, S., et al. (2012). Superflares on solar-type stars. *Nature* 485, 478–481. doi:10.1038/nature11063

Mann, I. R., Di Pippo, S., Opgenoorth, H. J., Kuznetsova, M., and Kendall, D. J. (2018). International collaboration within the united nations committee on the peaceful uses of outer space: Framework for international space weather services (2018–2030). *Space weather.* 16, 428–433. doi:10.1029/2018SW001815

Mannucci, A. J., and Tsurutani, B. T. (2018). "'Ionosphere and thermosphere responses to extreme geomagnetic storms," in *Extreme events in Geospace*. Editor N. Buzulukova (Elsevier), 493–511. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00020-0

Mannucci, A. J., Tsurutani, B. T., Iijima, B. A., Komjathy, A., Saito, A., Gonzalez, W. D., et al. (2005). Dayside global ionospheric response to the major interplanetary events of October 29-30, 2003 "Halloween Storms". *Geophys. Res. Lett.* 32, L12S02. doi:10.1029/2004GL021467

Marqué, C., Klein, K. L., Monstein, C., Opgenoorth, H., Pulkkinen, A., Buchert, S., et al. (2018). Solar radio emission as a disturbance of aeronautical radionavigation. *J. Space Weather Space Clim.* 8, A42. doi:10.1051/swsc/2018029

Matthia, D., Schaefer, M., and Meier, M. M. (2015). Economic impact and effectiveness of radiation protection measures in aviation during a ground level enhancement. *J. Space Weather Space Clim.* 5, A17. doi:10.1051/swsc/2015014

Maunder, E. W. (1905). The solar origin of the terrestrial magnetic disturbances. *Pop. Astron.* 13, 59–64.

Maunder, E. W. (1906). The solar origin of terrestrial magnetic disturbances. *Pop. Astron.* 14, 228–238.

McIntosh, S. W., Leamon, R. J., Krista, L. D., Title, A. M., Hudson, H. S., Riley, P., et al. (2015). The solar magnetic activity band interaction and instabilities that shape quasi-periodic variability. *Nat. Commun.* 6 (1), 6491. doi:10.1038/ncomms7491

Meier, M. M., and Matthiä, D. (2014). A space weather index for the radiation field at aviation altitudes. *J. Space Weather Space Clim.* 4, A13. doi:10.1051/swsc/2014010

Meier, R. R. (1991). Ultraviolet spectroscopy and remote sensing of the upper atmosphere. *Sp. Sci. Rev.* 58, 1–185. doi:10.1007/BF01206000

Meng, X., Tsurutani, B. T., and Mannucci, A. J. (2019). The solar and interplanetary causes of superstorms (minimum $Dst \leq -250$ nT) during the space age. *J. Geophys. Res. Space Phys.* 124, 3926–3948. doi:10.1029/2018JA026425

Millan, R. M. (2011). Understanding relativistic electron losses with BARREL. *J. Atmos. Sol.-Terr. Phys.* 73, 1425–1434. doi:10.1016/j.jastp.2011.01.006

Miyake, F., Masuda, K., and Nakamura, T. (2013). Another rapid event in the carbon-14 content of tree rings. *Nat. Commun.* 4, 1748. doi:10.1038/ncomms2783

Miyoshi, Y., and Kataoka, R. (2011). Solar cycle variations of outer radiation belt and its relationship to solar wind structure dependences. *J. Atmos. Solar-Terrestrial Phys.* 73, 77–87. ISSN 1364-6826. doi:10.1016/j.jastp.2010.09.031

Miyoshi, Y., Oyama, S., Saito, S., Kurita, S., Fujiwara, H., Kataoka, R., et al. (2015). Energetic electron precipitation associated with pulsating aurora: EISCAT and van allen Probe observations. *J. Geophys. Res. (Space Phys.)* 120, 2754–2766. doi:10.1002/2014JA020690

Moore, T. E., Fok, M. C., Delcourt, D. C., Slinker, S. P., and Fedder, J. A. (2008). Plasma plume circulation and impact in an MHD substorm. *J. Geophys. Res. (Space Phys.)* 113. doi:10.1029/2008ja013050

Narici, L., Rizzo, A., Berrilli, F., and Del Moro, D. (2018). "'Solar particle events and human deep space exploration: Measurements and considerations," in *Extreme events in Geospace*. Editor N. Buzulukova (Elsevier), 433–451. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00017-0

Newell, P. T., Sotirelis, T., Liou, K., Meng, C.-I., and Rich, F. J. (2007). A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables. *J. Geophys. Res.* 112, A01206. doi:10.1029/2006JA012015

Newell, P. T., Sotirelis, T., and Wing, S. (2009). Diffuse, monoenergetic, and broadband aurora: The global precipitation budget. *J. Geophys. Res.* 114, A09207. doi:10.1029/2009JA014326

Ngwira, C. M., and Pulkkinen, A. A. (2018). "'An overview of science challenges pertaining to our understanding of extreme geomagnetically induced currents," in *Extreme events in Geospace*. Editor N. Buzulukova (Elsevier), 187–208. ISBN 9780128127001. doi:10.1016/B978-0-12-812700-1.00008-X

Ngwira, C. M., Pulkkinen, A. A., Bernabeu, E., Eichner, J., Viljanen, A., and Crowley, G. (2015). Characteristics of extreme geoelectric fields and their possible causes: Localized peak enhancements. *Geophys. Res. Lett.* 42, 6916–6921. doi:10.1002/2015GL065061

Nishida, A. (1968a). GeomagneticD_p2 fluctuations and associated magnetospheric phenomena. *J. Geophys. Res. Space. Phys.)* 73, 1795–1803. doi:10.1029/ja073i005p01795

Nishida, A. (1968b). Coherence of geomagnetic DP 2 fluctuations with interplanetary magnetic variations. *J. Geophys. Res. Space. Phys.)* 71 (17), 5549–5559. doi:10.1029/ja073i017p05549

Notsu, Y., Maehara, H., Honda, S., Hawley, S. L., Davenport, J. R. A., Namekata, K., et al. (2019). Do kepler superflare stars really include slowly rotating sun-like stars? Results using APO 3.5 m telescope spectroscopic observations and gaia-DR2 data. *Astrophysical J.* 876, 58. doi:10.3847/1538-4357/ab14e6

Okamoto, S., Notsu, Y., Hiroyuki, M., Namekata, M., Honda, S., Ikuta, K., et al. (2021). Statistical properties of superflares on solar-type stars: Results using all of the kepler primary mission data. *Astrophysical J.* 906, 72. N2. doi:10.3847/1538-4357/abc8f5

Oughton, E. J., Skelton, A., Horne, R. B., Thomson, A. W. P., and Gaunt, C. T. (2017). Quantifying the daily economic impact of extreme space weather due to failure in electricity transmission infrastructure. *Space weather.* 15, 65–83. doi:10.1002/2016SW001491

Partamies, N., Juusola, L., Tanskanen, E., and Kauristie, K. (2013). Statistical properties of substorms during different storm and solar cycle phases. *Ann. Geophys.* 31, 349–358. doi:10.5194/angeo-31-349-2013

Paschmann, G., Haaland, S., and Treumann, R. (2003). "'Auroral plasma physics,". *International Space Science Institute, Bern Switzerland and Max-Planck-Institut Fur Extraterrestrische Physik, Garching, Germany, Reprinted from Space Science Reviews*. Editor G. Paschmann (Dordrecht: Kluwer Academic Publishers), Vol. 103. Nos. 1-4, 2002.

Petrinec, S. M., Chenette, D. L., Mobilia, J., Rinaldi, M. A., and Imhof, W. L. (1999). Statistical X ray auroral emissions - PIXIE observations. *Geophys. Res. Lett.* 26, 1565–1568. doi:10.1029/1999GL900295

Phillips, J. L., Bame, S. J., Barnes, A., Barraclough, B. L., Feldman, W. C., Goldstein, B. E., et al. (1995). Ulysses solar wind plasma observations from pole to pole. *Geophys. Res. Lett.* 22, 3301–3304. N 23. doi:10.1029/95GL03094

Pulkkinen, A., Thomson, A., Clarke, E., and McKay, A. (2003). April 2000 geomagnetic storm: Ionospheric drivers of large geomagnetically induced currents. *Ann. Geophys.* 21, 709–717. doi:10.5194/angeo-21-709-2003

Pulkkinen, T. I., Partamies, N., McPherron, R. L., Henderson, M., Reeves, G. D., Thomsen, M. F., et al. (2007). Comparative statistical analysis of storm time activations and sawtooth events. *J. Geophys. Res.* 112, A01205. doi:10.1029/2006JA012024

Pulkkinen, A., Bernabeu, E., Eichner, J., Viljanen, A., and Ngwira, C. (2015). Regional-scale high-latitude extreme geoelectric fields pertaining to geomagnetically induced currents. *Earth Planet Sp.* 67, 93. doi:10.1186/s40623-015-0255-6

Pulkkinen, A., Bernabeu, E., Thomson, A., Viljanen, A., Pirjola, R., Boteler, D., et al. (2017). Geomagnetically induced currents: Science, engineering, and applications readiness. *Space weather.* 15, 828–856. doi:10.1002/2016SW001501

Puthanveettil, S. E., Sankaran, M., and Gupta, S. B. (2014). "Mitigation techniques for arcing on space solar panels-results from ISRO," in Proceeding of the 2014 International Symposium on Discharges and Electrical Insulation in Vacuum (ISDEIV), Mumbai, India, November 2014 (IEEE), 701–703. doi:10.1109/DEIV.2014.6961779

Reames, D. V. (1999). Particle acceleration at the Sun and in the heliosphere. *Space Sci. Rev.* 90, 413–491. doi:10.1023/A:1005105831781

Riley, P., Mays, L., Andries, J., Amerstorfer, T., Biesecker, D., Delouille, V., et al. (2018). Forecasting the arrival time of coronal mass ejections: Analysis of the CCMC CME scoreboard. *Space weather.* 16, 1245–1260. doi:10.1029/2018SW001962

Robinson, R. M., Vondrak, R. R., Miller, K., Dabbs, T., and Hardy, D. (1987). On calculating ionospheric conductances from the flux and energy of precipitating electrons. *J. Geophys. Res. Space. Phys.* 92, 2565. doi:10.1029/ja092ia03p02565

K. Scherer, H. Fichtner, B. Heber, and U. Mall (Editors) (2005). *Space weather: The physics behind a slogan* (Springer). ISBN 978-3-540-22907-0.

Schwadron, N. A., Blake, J. B., Case, A. W., Joyce, C. J., Kasper, J., Mazur, J., et al. (2014). Does the worsening galactic cosmic radiation environment observed by CRaTER preclude future manned deep space exploration? *Space weather.* 12, 622–632. doi:10.1002/2014SW001084

Selesnick, R. S., and Albert, J. M. (2019). Variability of the proton radiation belt. *J. Geophys. Res. Space Phys.* 124, 5516–5527. doi:10.1029/2019JA026754

Sergeev, V. A., Pellinen, R. J., and Pulkkinen, T. I. (1996). Steady magnetospheric convection: A review of recent results. *Space Sci. Rev.* 75, 551–604. doi:10.1007/BF00833344

S. Sharma, Y. Kamide, and G. S. Lakhina (Editors) (2004). *Storm-substorm relationship* (Washington, DC: Amer. Geophys. Un. Press), 142.

Shea, M. A., and Smart, D. F. (2006). Compendium of the eight articles on the Carrington event attributed to or written by elias loomis in the American journal of science, 1859 1861. *Adv. Sp. Res.* 38, 313–385. doi:10.1016/j.asr.2006.07.005

Shelley, E. G. (1995). The auroral acceleration region: The world of beams, conies, cavitons, and other plasma exotica. *Rev. Geophys.* 33, 709–714. doi:10.1029/95RG00253

Shibata, K., Isobe, H., Hillier, A., Choudhuri, A. R., Maehara, H., Ishii, T. T., et al. (2013). Can superflares occur on our sun?. *Publ. Astron. Soc. Japan* 65 (3), 49. doi:10.1093/pasj/65.3.49

Shibata, K., and Magara, T. (2011). Solar flares: Magnetohydrodynamic processes. *Living Rev. Sol. Phys.* 8, 6. doi:10.12942/lrsp-2011-6

Smith, E. J., and Wolfe, J. H. (1976). Observations of interaction regions and corotating shocks between one and five AU: Pioneers 10 and 11. *Geophys. Res. Lett.* 3, 137–140. doi:10.1029/GL003i003p00137

Sparks, L., Altshuler, E., Pandya, N., Blanch, J., and Walter, T. (2022). WAAS and the ionosphere- a historical perspective: Monitoring storms. *Navigation* 69 (1), navi.503. navi.503. doi:10.33012/navi.503

Spence, H. E., Blake, J. B., and Fennell, J. F. (1993). Surface charging analysis of high-inclination, high-altitude spacecraft: Identification and physics of the plasma source region. *IEEE Trans. Nucl. Sci.* 40, 1521–1524. doi:10.1109/23.273510

Stone, R. G., and Tsurutani, B. T. (1985). *Collisionless shocks in the heliosphere: A tutorial review*. Washington D. C.: AGU monograph, 33.

Suess, S. T. (1979). Models of coronal hole flows. *Space Sci. Rev.* 23, 159–200. doi:10.1007/bf00173809

S. T. Suess and B. T. Tsurutani (Editors) (1998). *From the sun: Auroras, magnetic storms, solar flares, cosmic rays* (Washington D. C.: AGU monograph), 50.

Takahashi, T., Mizuno, Y., and Shibata, K. (2016). Scaling relations in coronal mass ejections and energetic proton events associated with solar superflares. *Astrophysical J.* 833, L8. doi:10.3847/2041-8205/833/1/L8

Takahashi, T., and Shibata, K. (2017). Sheath-accumulating propagation of interplanetary coronal mass ejection. *Astrophysical J.* 837, L17. doi:10.3847/2041-8213/aa624c

Tang, F., Tsurutani, B. T., Gonzalez, W. D., Akasofu, S. I., and Smith, E. J. (1989). Solar sources of interplanetary southward$B_z$events responsible for major magnetic storms (1978-1979). *J. Geophys. Res.* 94, 3535–3541. doi:10.1029/ja094ia04p03535

Tang, F., and Tsurutani, B. T. (1990). Reply [to "Low-latitude coronal hole as the only possible explanation for the November 25, 1978, geomagnetic storm: Comment on ';Solar sources of interplanetary southward Bz events responsible for major magnetic storms (1978–1979)' F. Tang et al."]. *J. Geophys. Res.* 95, 10721. doi:10.1029/ja095ia07p10721

Thomson, A. W. P., Gaunt, C. T., Cilliers, P., Wild, J. A., Opperman, B., McKinnell, L.-A., et al. (2010). Present day challenges in understanding the geomagnetic hazard to national power grids. *Adv. Space Res.* 45, 1182–1190. doi:10.1016/j.asr.2009.11.023

Tobiska, W. K., Bouwer, D., Smart, D., Shea, M., Bailey, J., Didkovsky, L., et al. (2016). Global real-time dose measurements using the automated radiation measurements for aerospace safety (ARMAS) system. *Space weather.* 14, 1053–1080. doi:10.1002/2016sw001419

Tobiska, W. K., Didkovsky, L., Judge, K., Weiman, S., Bouwer, D., Bailey, J., et al. (2018). 'Analytical representations for characterizing the global aviation radiation environment based on model and measurement databases. *Space weather.* 16, e2018SW001843. doi:10.1029/2018SW001843

Toffoletto, F., Sazykin, S., Spiro, R., and Wolf, R. (2003). Inner magnetospheric modeling with the rice convection model. *Sp. Sci. Rev.* 107, 175–196. doi:10.1023/A:1025532008047

Toriumi, S., and Wang, H. (2019). Flare-productive active regions. *Living Rev. Sol. Phys.* 16, 3. doi:10.1007/s41116-019-0019-7

Torr, M. R., Torr, D. G., Zukic, M., Johnson, R. B., Ajello, J., Banks, P., et al. (1995). A far ultraviolet imager for the international solar-terrestrial physics mission. *Space Sci. Rev.* 71, 329–383. doi:10.1007/bf00751335

To´th, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., De Zeeuw, D. L., et al. (2005). Space weather modeling framework: A new tool for the space science community. *J. Geophys. Res. (Space Phys.)* 110, A12226. doi:10.1029/2005JA011126

Tsurutani, B. T., Gonzalez, W. D., Tang, F., Akasofu, S. I., and Smith, E. J. (1988). Origin of interplanetary southward magnetic fields responsible for major magnetic storms near solar maximum (1978–1979). *J. Geophys. Res.* 93, 8519–8531. doi:10.1029/ja093ia08p08519

Tsurutani, B. T., Gonzalez, W. D., Tang, F., Lee, Y. T., Okada, M., and Park, D. (1992). Reply to L. J. Lanzerotti: Solar wind RAM pressure corrections and an estimation of the efficiency of viscous interaction. *Geophys. Res. Lett.* 19, 1993–1994. doi:10.1029/92GL02239

Tsurutani, B. T., Gonzalez, W. D., Gonzalez, A. L. C., Tang, F., Arballo, J. K., and Okada, M. (1995). Interplanetary origin of geomagnetic activity in the declining phase of the solar cycle. *J. Geophys. Res.* 100, 21717–21733. doi:10.1029/95JA01476

B. T. Tsurutani, W. D. Gonzalez, Y. Kamide, and J. K. Arballo (Editors) (1997). *Magnetic storms* (Washington D. C.: AGU Monograph), 98.

Tsurutani, B. T., Kallenrode, M. B., Kemski, R. P., Klein, J. W., Lin, R. P., Ling, J. C., et al. (2000). *Radiation environment near the sun: Solar Probe*. NASA white paper, JPL/NASA internal report.

Tsurutani, B. T., Gonzalez, W. D., Lakhina, G. S., and Alex, S. (2003). The extreme magnetic storm of 1–2 September 1859. *J. Geophys. Res.* 108, 1268. doi:10.1029/2002JA009504

Tsurutani, B. T., Gonzalez, W. D., Guarnieri, F., Kamide, Y., Zhou, X. Y., and Arballo, J. K. (2004a). Are high-intensity long-duration continuous AE activity (HILDCAA) events substorm expansion events? *J. Atmos. Sol.-Terr. Phys.* 55, 167–176. doi:10.1016/j.jastp.2003.08.015

Tsurutani, B. T., Mannucci, A. J., Iijima, B., Abdu, M. A., Sobral, J. H. A., Gonzalez, W. D., et al. (2004b). Global dayside ionospheric uplift and enhancement associated with interplanetary electric fields. *J. Geophys. Res.* 109 (A), A08302. doi:10.1029/2003JA010342

Tsurutani, B. T., Zhou, X. Y., and Gonzalez, W. D. (2004c). "A lack of substorm expansion phases during magnetic storms induced by magnetic clouds," in *Storm-substorm relationship*. Editors S. Sharma, Y. Kamide, and G. Lakhina (Washington, DC: Amer. Geophys. Un. Press), 142, 23.

Tsurutani, B. T., Judge, D. L., Guarnieri, F. L., Gangopadhyay, P., Jones, A. R., Nuttall, J., et al. (2005). The October 28, 2003 extreme EUV solar flare and resultant extreme ionospheric effects: Comparison to other Halloween events and the Bastille Day event. *Geophys. Res. Lett.* 32, L03S09. doi:10.1029/2004GL021475

B. T. Tsurutani, R. L. McPherron, W. D. Gonzalez, G. Lu, J. H. A. Sobral, and N. Gopalswamy (Editors) (2006a). *Recurrent magnetic storms: Corotating solar wind streams* (Washington D. C.: AGU Monograph), 167.

Tsurutani, B. T., Gonzalez, W. D., Gonzalez, A. L. C., Guarnieri, F. L., Gopalswamy, N., Grande, M., et al. (2006b). Corotating solar wind streams and recurrent geomagnetic activity: A review. *J. Geophys. Res.* 111, A07S01. doi:10.1029/2005JA011273

Tsurutani, B. T., Verkhoglyadova, O. P., Mannucci, A. J., Saito, A., Araki, T., Yumoto, K., et al. (2008). Prompt penetration electric fields (PPEFs) and their ionospheric effects during the great magnetic storm of 30-31 October 2003. *J. Geophys. Res.* 113, A05311. doi:10.1029/2007JA012879

Tsurutani, B. T., Verkhoglyadova, O. P., Mannucci, A. J., Lakhina, G. S., Li, G., and Zank, G. P. (2009). A brief review of "solar flare effects" on the ionosphere. *Radio Sci.* 44, RS0A17. doi:10.1029/2008RS004029

Tsurutani, B. T., Lakhina, G. S., Verkhoglyadova, O. P., Gonzalez, W. D., Echer, E., and Guarnieri, F. L. (2011). A review of interplanetary discontinuities and their geomagnetic effects. *J. Atmos.Sol-Terr. Phys.* 73, 5–19. doi:10.1016/j.jastp.2010.04.001

Tsurutani, B. T., Hajra, R., Tanimori, T., Takeda, A., Bhanu, R., Mannucci, A. J., et al. (2016). Heliospheric plasma sheet (HPS) impingement onto the magnetosphere as a cause of relativistic electron dropouts (REDs) via coherent EMIC wave scattering with possible consequences for climate change mechanisms. *J. Geophys. Res. Space. Phys.)* 121. doi:10.1002/2016JA022499

Tsurutani, B. T., Lakhina, G. S., and Hajra, R. (2020). The physics of space weather/solar terrestrial physics (STP): What we know now and what the current and future challenges are. *Nonl. Proc. Geophys.* 27, 75–119.

Tsurutani, B. T., Zank, G. P., Sterken, V. J., Shibata, K., Nagai, T., Mannucci, A., et al. (2022a). 'Space plasma physics: A review IEE. *Trans. Plasma Sci*, 1–61. doi:10.1109/TPS.2022.3208906

Tsurutani, B. T., Green, J., and Hajra, R. (2022b). The possible cause of the 40 SpaceX Starlink satellite losses in February 2022: Prompt penetrating electric fields and the dayside equatorial and midlatitude ionospheric convective uplift. https://arxiv.org/abs/2210.07902.

Tsurutani, B. T., and Hajra, R. (2021). The interplanetary and magnetospheric causes of geomagnetically induced currents (GICs) > 10 A in the mäntsälä Finland pipeline: 1999 through 2019. *J. Space Weather Space Clim.* 11, 23. doi:10.1051/swsc/2021001

Tsurutani, B. T., and Lakhina, G. S. (2014). An extreme coronal mass ejection and consequences for the magnetosphere and Earth. *Geophys. Res. Lett.* 41, 287–292. doi:10.1002/2013GL058825

Tsurutani, B. T., and Gonzalez, W. D. (1987). The cause of high-intensity long-duration continuous AE activity (HILDCAAs): Interplanetary Alfven wave trains. *Planet. spa. Sci.* 35 (4), 405–412. doi:10.1016/0032-0633(87)90097-3

Tsurutani, B. T., and Lin, R. P. (1985). Acceleration of >47 keV Ions and >2 keV electrons by interplanetary shocks at 1 AU. *J. Geophys. Res.* 90, 1–11. doi:10.1029/JA090iA01p00001

Tsurutani, B. T., and Meng, C.-I. (1972). Interplanetary magnetic field variations and substorm activity. *J. Geophys. Res.* 77, 2964–2970. doi:10.1029/ja077i016p02964

Tsurutani, B. T., and Stone, R. G. (1985). *Collisionless shocks in the heliosphere: Reviews of current research*. Washington D. C.: AGU Monograph, 34., v.

Usoskin, I. G., Kromer, B., Ludlow, F., Beer, J., Friedrich, M., Kovaltsov, G. A., et al. (2013). The AD775 cosmic event revisited: The sun is to blame. *Astronomy Astrophysics* 552, L3. doi:10.1051/0004-6361/201321080

Van Driel-Gesztelyi, L., and Green, L. M. (2015). Evolution of active regions. *Living Rev. Sol. Phys.* 12, 1. doi:10.1007/lrsp-2015-1

Viljanen, A., Amm, O., and Pirjola, R. (1999). Modeling geomagnetically induced currents during different ionospheric situations. *J. Geophys. Res.* 104, 28059–28071. doi:10.1029/1999ja900337

Viljanen, A., Tanskanen, E. I., and Pulkkinen, A. (2006). Relation between substorm characteristics and rapid temporal variations of the ground magnetic field. *Ann. Geophys.* 24, 725–733. doi:10.5194/angeo-24-725-2006

Von Humboldt, A. (1808). Die Vollstandigste aller bisherigen Beobachtungen uber den Einfluss des Nordlichts auf die Magnetnadel angestellt. *Ann. Phys.* 29, 425–429. doi:10.1002/andp.18080290806

Wik, M., Pirjola, R., Lundstedt, H., Viljanen, A., Wintoft, P., and Pulkkinen, A. (2009). Space Weather events in July 1982 and October 2003 and the effects of geomagnetically induced currents on Swedish technical systems. *Ann. Geophys* 27, 1775–1787. doi:10.5194/angeo-27-1775-2009

Wilson, G. R., Weimer, D. R., Wise, J. O., and Marcos, F. A. (2006). Response of the thermosphere to Joule heating and particle precipitation. *J. Geophys. Res.* 111, A10314. doi:10.1029/2005JA011274

Winckler, J. R., Peterson, L., Hoffman, R., and Arnoldy, R. (1959). Auroral X-rays, cosmic rays, and related phenomena during the storm of February10–11, 1958. *J. Geophys. Res.* 64, 597–610. doi:10.1029/jz064i006p00597

Wolff, E. W., Bigler, M., Curran, M. A. J., Dibb, J. E., Frey, M. M., Legrand, M., et al. (2012). The Carrington event not observed in most ice core nitrate records. *Geophys. Res. Lett.* 39, L08503. doi:10.1029/2012GL051603

Woods, T. N., and Rottman, G. J. (2002). *Solar ultraviolet variability over time periods of aeronomic interest*. Washington D.C.: AGU Monograph, 130.

Wrenn, G. L., Rodgers, D. J., and Ryden, K. A. (2002). A solar cycle of spacecraft anomalies due to internal charging. *Ann. Geophys.* 20, 953–956. doi:10.5194/angeo-20-953-2002

Yashiro, S., Gopalswamy, N., Akiyama, S., Michalek, G., and Howard, R. A. (2005). Visibility of coronal mass ejections as a function of flare location and intensity. *J. Geophys. Res.* 110, A12S05. doi:10.1029/2005JA011151

Yue, X., Wan, W., Yan, L., Sun, W., Hu, L., and Schreiner, W. (2018). "Chapter 22 - the effect of solar radio bursts on GNSS signals," in *Extreme events Geospace orig. Predict. Cons.* Editor N. Buzulukova (Amsterdam: Elsevier), 541–554. doi:10.1016/B978-0-12-812700-1.00022-4

Zheng, Y., Ganushkina, N. Y., Jiggens, P., Jun, I., Meier, M., Minow, J. I., et al. (2019). Space radiation and plasma effects on satellites and aviation: Quantities and metrics for tracking performance of space weather environment models. *Space weather.* 17, 1384–1403. doi:10.1029/2018SW002042

Zhu, C., Qiu, J., Liewer, P., Vourlidas, A., Spiegel, M., and Hu, Q. (2020). How does magnetic reconnection drive the early-stage evolution of coronal mass ejections? *Astrophysical J.* 893, 141. doi:10.3847/1538-4357/ab838a

J. B. Zirker (Editor) (1977). "Coronal holes and high-speed wind streams," *A monograph from skylab solar workshop I* (Boulder: Colorado Associated University Press).

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 05 January 2023
DOI 10.3389/fspas.2022.1095701

Check for updates

OPEN ACCESS

EDITED BY
Barbara Cavalazzi,
University of Bologna, Italy

REVIEWED BY
Akos Kereszturi,
Hungarian Academy of Sciences (MTA), Hungary
Terence Phillip Kee,
University of Leeds, United Kingdom

*CORRESPONDENCE
Frances Westall,
frances.westall@cnrs-orleans.fr

SPECIALTY SECTION
This article was submitted to Astrobiology, a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 11 November 2022
ACCEPTED 16 December 2022
PUBLISHED 05 January 2023

CITATION
Westall F, Brack A, Fairén AG and Schulte MD (2023), Setting the geological scene for the origin of life and continuing open questions about its emergence.
Front. Astron. Space Sci. 9:1095701.
doi: 10.3389/fspas.2022.1095701

COPYRIGHT
© 2023 Westall, Brack, Fairén and Schulte. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Setting the geological scene for the origin of life and continuing open questions about its emergence

Frances Westall[1]*, André Brack[1], Alberto G. Fairén[2,3] and Mitchell D. Schulte[4]

[1]Centre de Biophysique Moléculaire, CNRS, Orléans, France, [2]Centro de Astrobiología (CAB, CSIC-INTA), Madrid, Spain, [3]Cornell University, Ithaca, NY, United States, [4]NASA Headquarters, Washington, DC, United States

The origin of life is one of the most fundamental questions of humanity. It has been and is still being addressed by a wide range of researchers from different fields, with different approaches and ideas as to how it came about. What is still incomplete is constrained information about the environment and the conditions reigning on the Hadean Earth, particularly on the inorganic ingredients available, and the stability and longevity of the various environments suggested as locations for the emergence of life, as well as on the kinetics and rates of the prebiotic steps leading to life. This contribution reviews our current understanding of the geological scene in which life originated on Earth, zooming in specifically on details regarding the environments and timescales available for prebiotic reactions, with the aim of providing experimenters with more specific constraints. Having set the scene, we evoke the still open questions about the origin of life: did life start organically or in mineralogical form? If organically, what was the origin of the organic constituents of life? What came first, metabolism or replication? What was the time-scale for the emergence of life? We conclude that the way forward for prebiotic chemistry is an approach merging geology and chemistry, i.e., far-from-equilibrium, wet-dry cycling (either subaerial exposure or dehydration through chelation to mineral surfaces) of organic reactions occurring repeatedly and iteratively at mineral surfaces under hydrothermal-like conditions.

KEYWORDS

origin of life, early earth, prebiotic environments, volcanic rocks, stochastic chemistry, hydrothermal environments

1 Introduction

Studies related to the emergence of life, whether from the point of view of prebiotic chemistry, or from molecular biology, take place on the timescales of laboratory experiments, the lifetime of funded projects and funded students and postdocs. Each carefully crafted experiment aims at maximising the results obtainable on these

timescales, each experiment adds to our accumulated knowledge and to the advancement of the field. However, no experiment has yet been run end to end, *i.e.* from the initial ingredients (organic molecules, liquid water, energy and essential elements, such as H, N, O, P, S, and transition metals) to the emergence of a protocell. Apart from the necessity of using stochastic chemistry, an acknowledged concept (Dass et al., 2016) but one that is difficult to put into practice, realistic timescales may not be conducive to "reaching the goal".

This contribution does not aim to produce a simple formula to account for the emergence of life; rather, it seeks to describe the Hadean environment, when life emerged, in as realistic a way as possible to help guide future experiments. In addition to providing information on the prevailing Hadean rock lithologies, their mineralogy and chemical composition, we seek in particular to emphasise aspects of the environment, such as stability and longevity of particular conditions, which will constrain reaction rates for prebiotic chemistry. If a certain environment is envisaged as a possible prime locale for the emergence of life, how long was it stable, how long did it last? (Begging the question, how long did it take to go from simple molecular ingredients to the emergence of the first cell?). These concepts go over and beyond the very localised conditions in which stability would be death to prebiotic chemistry and where gradients and instability need to be the norm.

We take as outer boundary conditions for this study the consolidation of the Earth (4.54, Dalrymple, 2001, or ~4.53 Ga according to Kleine et al., 2009; Connelly et al., 2012) and the oldest traces of life (an eternally controversial concept but, in our view, which will be explained below, certainly by 3.75 Ga and very probably before, e.g., 3.95 Ga). Note that the oldest preserved traces of life do not represent the first life forms. Occurring in Eo-Palaeoarchaean rocks (4.0–3.33 Ga) that formed after the likely emergence of life in the Hadean (4.53–4.0 Ga), these traces reflect an evolutionary stage that had already comprised chemotrophy and phototrophy by at least 3.45 Ga and possibly before (Westall et al., 2011a; Noffke et al., 2013; Hickman-Lewis et al., 2018). Within the bounds of these two critical dates is the establishment of habitable conditions on the Hadean Earth, by which we mean the appearance of water on the surface at temperatures conducive to prebiotic chemistry (e.g., Sleep et al., 2014; Zahnle et al., 2015).

Obviously establishing the dates of these critical time limits relies heavily on modelling and comparative planetology, as well as inherited geochemical signatures of former Hadean crust, inherited Hadean zircon crystals, and the few portions of Hadean (4.56–4.0 Ga), Eoarchaean (4.0–3.5), and Palaeoarchaean (3.5–3.2 Ga) crust that have been preserved. As a consequence, dating the emergence of habitable conditions is like a movable feast, it depends on the estimations of the models. These general considerations are perhaps less important than the relative stability and longevity of the location(s) where life may have emerged and that would necessarily have existed on much shorter time scales.

Our approach is from the global to the local scale in terms of the emergence of life. It is complementary to the approach taken by Saha et al. (2022) that regards the prebiotic processes leading to the emergence of life on the microenvironmental scale. In their recent review, Saha et al. note that, critical at these scales are the physicochemical properties of the substrates and different prebiotic reaction microenvironments on early Earth. These microenvironments comprise various kinds of liquid and mineral or aqueous environments. Examples of liquid microenvironments include aqueous ones (bulk aqueous solution, sea spray, gels, ice); non-aqueous liquids (solvents); e.g., organic acids; deep eutectic solvents, e.g., urea, glycerol, and acetamicide; high presuure super-critical fluids, e.g., CO2, H20; tar; the interiors of structures, such as lipid bilayers; condensed droplet microenvironments. Examples of the solid microenvironments include mineral surfaces, e.g., clay minerals, sulphides, phosphorous-bearing minerals, as well as the early mantle conditions, primarily its oxidation state, temperature and pressure.

2 Planetary formation and the condensation of water

Planetary formation is generally considered to have occurred about 4.54 Ga (Dalrymple, 2001) or ~4.53 Ga according to Kleine et al. (2009) and Connelly et al. (2012), with our planet accumulating from pebble-sized materials rather than planetesimals and protoplanets, according to the latest models (Morbidelli et al., 2012; Johansen et al., 2021; Raymond, 2021). Nevertheless, continued accretion occurred and collision with another planet the size of Mars (Halliday et al., 1996) that occurred approximately 4.51 Ga (Barboni et al., 2017) and led to the formation of the Moon. Latest models suggest that this major collision was a double collision that took place at a relatively slow pace (otherwise Theia would have been drawn towards Venus) and at an angle of ~45° (Emsenhuber et al., 2017; Asphaug et al., 2021). A slightly younger date for the collision, between 4.426–4.417 Ga, has been proposed by Connelly and Bizzaro (2016).

The consolidation of the Earth's initial magma ocean, core formation and crustal differentiation is calculated by using various isotopic proxies (reviewed by Brasser et al., 2021) shown in Table 1. Core formation occurred between 4.45–4.53 Ga (see references in Table 1). Fractionation processes in the mantle/crust were perturbed by the Moon-forming impact, and the subsequent cooling of the planet would have been retarded by the huge, opaque CO_2/H_20 degassed atmosphere that would also have contained additional volatiles from the molten crust/magma (Zahlnle et al., 2015). Zahnle et al. (2015) estimate that the post-impact atmosphere could

TABLE 1 Collection of published ages related to core formation, differentiation and crust formation. Ages reported here are both model/intercept ages and absolute ages, but not models of ages (after Brasser et al., 2021).

Earth	Age (Ma)	References
U-Pb age of silicate differentation	4,450–4,500	Manhes et al. (1979), Albarède and Juteau (1984), Allègre et al. (2008)
I-Pu-Xe atmosphere retention age	4,450–4,530	Staudacher and Allègre (1982), Ozima and Podosek (1999), Mukhopadhyay (2012), Avice & Marty (2014), Caracausi et al. (2016)
Hf-W age of core formation	4,450–4,530	Halliday et al. (1996), Yin et al. (2002), Kleine et al. (2002, 2009)
Oldest-known terrestrial zircon	4,380	Valley et al. (2014)
Sm-Nd silicate differentiation age	4,460–4,530	Caro et al. (2003), Boyet & Carlson (2005)
Lu-Hf intercept age of crust formation	4,450–4,500	Harrison et al. (2005, 2008)
Moon		
Lu-Hf intercept age of crust formation	4,270–4,510	Taylor et al. (2009), Barboni et al. (2017), Maurice et al. (2020)
Hf-W age of lunar core formation	4,505–4,530	Jacobsen (2005), Kleine et al. (2005), Touboul et al. (2009), Thiemens et al. (2019)

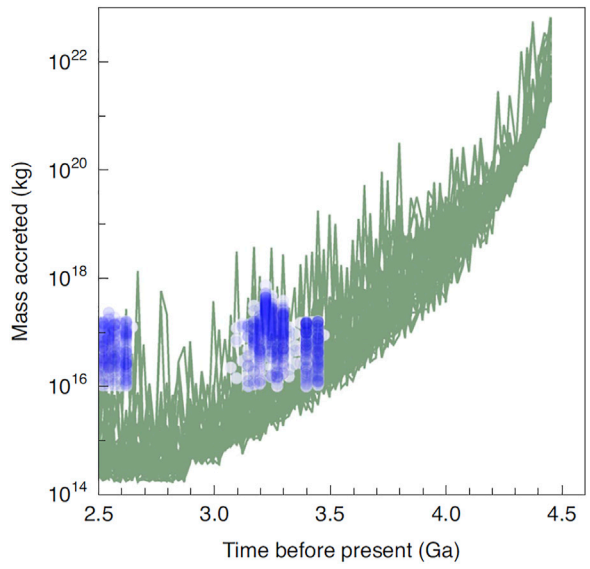


FIGURE 1
Earth's collisional history based on modelling with the dates of Archaean impact spherule layers added (blue) (after Marchi et al., 2021).

have contained between 100–1,000 bars of H_2O and CO_2 that was augmented by smaller amounts of CO, H_2, N_2, various sulphur-containing gases, and other geochemical volatiles evaporated from the magma (Schaefer and Fegley, 2010; Fegley and Schaefer, 2013). In addition to the effects of a thick atmosphere, tidal heating caused by the closer Moon would have slowed down cooling of the mantle (Zahnle et al., 2015). Once internal heating decreased below the critical runaway greenhouse threshold, Sleep et al. (2014) estimate that the ocean could have condensed onto Earth's surface under a CO_2 atmosphere of about 100 bar. Initial temperatures of ~200°C could have been too warm for prebiotic chemistry, however. If the upper temperature limit of life is taken as a benchmark, Sleep et al. (2014) calculate that surface temperatures reached about 122°C when the CO_2 pressure decreased to approximately 25 bar. Critical to the establishment of habitable conditions (i.e. temperatures at which critical prebiotic chemistry could occur) at the surface of the Earth was therefore the removal of CO_2 from the atmosphere, which occurred mainly by tectonic processes, modelled to have taken place on a time scale of about 100 My or more after the Moon-forming impact (Sleep, 2016).

Another factor of importance to take into account were the continued impacts, in particular those over 500 m diameter that could potentially have been planet-sterilising, while impactors over 300 m diameter would have increased surface water temperatures over about 100°C (Marchi et al., 2021). Their models suggest that the latest planet-sterilising impact occurred about 4.27 Ga. Figure 1 illustrates the modelled impact curve with known Palaeo-Neoarchaean impact spherule layers overlain (after Marchi et al., 2021).

2.1 Evidence for water and habitable conditions

There are a number of proxies for calculating whether or not there was liquid water at the surface of the Earth: 1) fractionated felsic crust, i.e. mafic crust that has been altered in the presence of water and fractionated to produce more silica-rich (felsic) crust; 2) zircon crystals of Hadean age that have ?^{18}O signature indicative of liquid water (i.e. formed in felsic melts); 3) sediments formed under aqueous conditions or specific mafic

lava structures, such as pillow lavas, indicative of extrusion under water.

We will deal with felsic crust formation later because, owing to the absence of directly-preserved early crustal materials from the Hadean, we rely on inherited information from long-lived components, e.g., zircon crystals, that may have been formed either through fractionation of mafic mantle or during the production of aqueously-mediated felsic crust. These robust crystals generally have complicated histories often including overgrowths formed under different geological ages over the original crystal. Studies of the oxygen isotopic signatures preserved in Hadean-age zircon crystals (4.4–4.3 Ga, Wilde et al., 2001; Mojzsis et al., 2001; Valley et al., 2014) and combined oxygen and silicon isotope measurements (Trail et al., 2018) suggest that they formed in the presence of hydrothermal processing of the crust, implying the presence of water recycled into the crust from the surface of the Earth by 4.4 Ga. A number of more recent studies have questioned the age dating of the older zircons because of the fact that the crystals are long lived and exhibit multiple mantles of growth (Whitehouse et al., 2017), as well as the resetting of their U/Pb ages by later metamorphism or impact events (Griffin et al., 2014). What is important here is that the oldest, *bona fide*, oxygen isotope signature indicative of aqueous interaction with a fractionating felsic melt is revealed in a zircon aged 4.15 Ga (Muehlenbachs and Clayton, 1976; Whitehouse et al., 2017). This then is the baseline for liquid water at the surface, although, following the models described above, it is likely that liquid water at temperatures below ~120°C existed previously.

Dating the oldest sediments known on Earth, i.e. detrital or chemical deposits formed in aqueous conditions, is highly controversial, as different dating techniques provide different ages. For example, the Nuvvuagittuq terrane in northern Quebec is purported to have a formation age of between 3.7 Ga according to U-Pb methods (Cates and Mojzsis, 2007) and 4.321 Ga according to the Sm-Ne method (O'Neill et al., 2012). This could be crucial because, if the terrane is Hadean in age, it holds the oldest known sediments. If not, it falls in the age range of the sediments from the 3.7–3.8 Ga Isua terrane in West Greenland (Appel et al., 1998), also containing metamorphosed sediments. In both cases, the sediments in question comprise banded iron formations, i.e. chemically (and possibly biologically) precipitated alternations of Fe-rich and Si-rich layers, as well as possible subaerial conglomerates. Volcanic sediments deposited under water are common in the younger, better preserved crustal remnants of the ~3.5 Ga Barberton and Pilbara Greenstone Belts. Note that the Isua terrane contains the oldest pillow lava structures, i.e. evidence of direct lava extrusion under water. The rims of the pillow lavas are enriched in Fe_2O_3, MgO, MnO, K_2O, Rb, Ba, Ga, Y, and transition metals compared to the cores (higher concentrations of SiO_2, Na_2O, P_2O_5, Sr, Pb, U, Nb, and the light rare earth elements (REEs) than the rims, Polat et al., 2003). Given the recent successful experiments to convert ribonucleoside triphosphates to polyribonucleic acid when incubated with the glassy rims of Hadean-analogue basalts (Jerome et al., 2022), these transition metal-enriched, vitreous rock surfaces are certainly very interesting for prebiotic chemistry.

2.2 Oldest potential and definitive evidence for life

The oldest morphological signs of life are purported to be hosted at the 3.7–4.3 Ga Nuvvuagittuq terrane. Dodd et al. (2017) and Papineau et al. (2022) report hematite filaments and tubes in the jasper-carbonate banded iron formation. The hematite filaments are centimetre-size, pectinate-branching, parallel-aligned, undulated, and contain Fe^{2+}-oxides. Morphologically, they resemble modern Fe oxidising filamentous microorganisms. Papineau et al. (2022) invoke metabolic pathways including Fe-oxidation and S-disproportionation, as well as anoxygenic photosynthesis. However, given the great age and metamorphic condition of the purported traces of life it may be difficult to definitively conclude their biogenicity, and as pointed out by Papineau et al. themselves, the structures could be simply abiotic in origin. Indeed, these structures have been alternatively described as chemical precipitates because of their hydrothermal affinities (McMahon, 2019). Other studies have also demonstarted their non-biogenicity (Greer et al., 2020; Lan et al., 2022), describing the structures as metamorphic phenomena. Nevertheless, the fact that the sediment was clearly a chemical deposit indicates precipitation in water.

On the other hand, the oldest potential chemical signs of life during the Hadean comes from analysis of the carbon isotope signature of graphite in a 4.1 Ga zircon crystal of—24‰^{13}C, consistent with biological fractionation (Bell et al., 2015). Similarly, the sediments in the 3.8–3.7 Ga Isua terrane hold potential evidence of life, also in the form of carbon isotope signatures (Rosing, 1999) as well as, more importantly, remnant organic molecules with compositions and structures suggestive of a biogenic origin (Hassenkam et al., 2017). The latter *in situ* FTIR investigation of metastable carbonaceous inclusions in a garnet crystal within the metamorphosed sediments documents structural binding of nitrogen, oxygen and possibly phosphorous to organic molecules, strong signatures that the material derived from living organisms. On the other hand, purported stromatolites from the same Isua Greenstone Belt (Nutman et al., 2016) are clearly metamorphic artefacts (Allwood et al., 2018; Zawaski et al., 2020).

Definitive evidence of widespread and varied life forms is well preserved in the ~3.5 Ga Pilbara and Barberton Greenstone Belts. Both chemotrophic (Westall et al., 2006a; Westall et al., 2011a) and photrophic life forms (Hofmann et al., 1999; Allwood et al., 2006; Noffke et al., 2013) have been interpreted from the fossil

remains. In the case of the Pilbara and Barberton Greenstone Belts, their excellent preservation has conserved morphological remains of microorganisms, biofilms, and stromatolites, as well as various isotopic and organo-geochemical evidence. The widespread distribution and degree of evolution of early life by 3.5 Ga strongly suggests its appearance much earlier (for a review of evidence of life and its implications and evolution in Southern Africa, see Hickman-Lewis and Westall, 2021).

Thus, by 4.15 Ga there is evidence of water at the surface of the Earth, as documented by zircon crystals of that age formed in a mantle environment influenced by altered, supracrustal material of felsic origin (Cavoisie et al., 2005). Apparently, a mechanism for fractionating carbon that was similar to that used by life today was in operation by then or soon after (Bell et al., 2015). By Isua times (3.8–3.7 Ga) life was very likely present on Earth, and by about 3.5–3.45 Ga it was widespread.

3 Rocky ingredients: Protocontinents, volcanic rocks (including sediments), hydrothermal silica

From the above, we conclude that the surface of the Earth was habitable before 4.15 Ga but, because of the uncertainties regarding timing of critical events, such as the Moon-forming impact, cooling of the planet, and removal of a large fraction of the CO_2 from the atmosphere, that can only be addressed through modelling, it is not possible to be more precise in dating the initiation of habitable conditions.

In this section, we document what we know or think we understand about the Hadean/Eo-Palaeoarchaean crust in terms of protocontinents, and the evidence for exposed landmasses and fluvial input into the shallow water basins surrounding the exposed volcanoes. We document the composition of the volcanic rocks on the early Earth and detail the different kinds of environments that existed (for which we have evidence) or may have existed (for which we have no preserved geological record, e.g., deep ocean basins). We also emphasise the importance on the processes conducive to the origin of life of hydrothermal activity in the Hadean-Palaeoarchaean period, and of seawaters saturated in dissolved silica that precipitated out at different rates depending upon the environment, consequently exerting significant influence on the rates of prebiotic reactions.

3.1 Protocontinents and volcanic rocks

An important question regarding some of the possible scenarios for the origins of life is the availability of exposed landmasses. Models suggest that there was up to 40% more water on the surface (Sim et al., 2016), which was eventually slowly removed through plate tectonic recycling. It had originally been

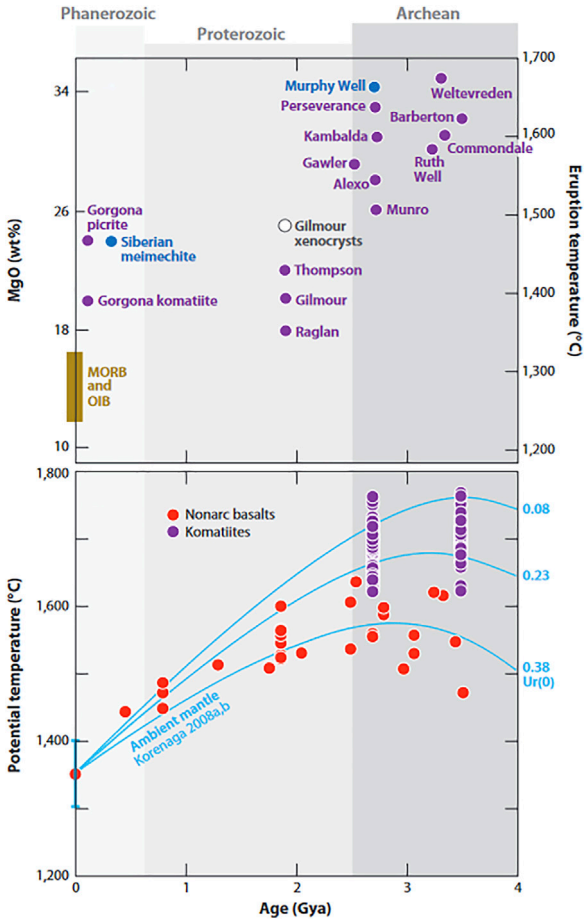


FIGURE 2
Eruption temperatures and mantle temperatures of komatiite lavas and melts through time highlighting the higher mantle temperatures during the Archaean and Hadean (after Arndt and Nisbet, 2012).

thought that the abundance of Hadean zircons indicated widespread production of felsic crust (and exposed landmass) through modern-style plate tectonics. This is not the place for a detailed discussion on the timing and origin of plate tectonics (a recent review can be found in Westall et al., 2022), but we will provide a brief overview here.

During the Hadean, when life emerged, the planet was dominated by higher mantle temperatures (Franck, 1998; Schubert et al., 2001; O'Neill et al., 2007; Perchuk et al., 2020), possibly up to six times higher than present day mantle temperatures. Indeed, ultramafic lavas, such as Mg and Fe-rich komatiites, common during the Archaean, are testimony to higher early mantle temperatures (Figure 2) (Arndt et al., 2008). The tectonic regime is uncertain but the Hadean Earth was likely dominated by a stagnant lid with likely squishy lid/plume-lid and plate tectonics (Sizova et al., 2015; Rozel et al., 2017; Lourenco et al., 2018), as well as some proto continents.

TABLE 2 Elemental compositions of Archaean basalts (tholeites, komatiites, cumulates and evolved tholeites from the Kromberg Formation, Barberton Greenstone Belt, South Africa) from Vennemann and Smith (1999) compared with Mid-Ocean Ridge Basalts from Gale et al. (2013).

	Kromberg Fm. Tholeitic basalt	Komatiites	Cumulates	Evolved tholeites	Gale et al. (2013)
					MORB all
SiO_2	50,73	48	52,18	66,13	50.17
TiO_2	1,47	5	0,36	1,45	1.68
Al_2O_3	14,75	6,54	3,37	13,63	14.7
Fe_2O_3	15,5	12,79	13,29	8,99	
FeO					10.43
MnO	0,2	14	0,27	0,09	0.184
MgO	6,1	22,4	15,09	3,73	7.58
CaO	7.38	9,47	15,02	1,34	11.39
Na_2O	2,14	0,01	0,36	3,61	2.79
K_2O	0,8	0	0,01	0,73	0.16
P_2O_5	0,17	0,06	0,05	0,3	0.184

Structural and geochemical evidence indicates that plate tectonics initiated between 4.0 and 3.0 Ga (Griffin et al., 2014; Lammer et al., 2018; Dehant et al., 2019), with some contending that modern style plate tectonics did not occur until only about 1 Ga (Korenaga, 2018; Hawkesworth and Cawood, 2020). The process would have been gradual and likely took place in different locations around the globe at different times.

Was there emerged continental crust during the Hadean? As mentioned above, the apparent relative abundance of Hadean age zircon crystals has been interpreted to suggest the formation of a significant amount of fractionated (felsic) Hadean crust (cf. Arndt and Nisbet, 2012). However, on the basis of cathodoluminescence investigations of zircon crystals to determine their age and history, Whitehouse et al. (2017) interprets many of the Hadean-age crystals as being younger, and the majority of them formed in the mantle rather than the crust. These authors noted four pulses of zircon formation centred around 4.37, 4.15, 4.1, and 4.02 Ga, but whether they indicate pulses in felsic crustal formation, enclaves of fractionated crust in a mainly mafic protocrust, or melts formed by impacts or other processes, is uncertain. Based on modelling and geochemical analysis (U–Pb ages, and $^{177}Hf/^{176}Hf$ ratios) of the ancient zircons, as well as Re–Os model ages on sulphides and alloys in mantle-derived rocks and crystals, Griffin et al. (2014) also conclude that the Hadean/Eoarchaean Earth was highly volcanic with rocks of predominantly mafic composition. Recycling of the crust would have been provoked by impacts, with fractionation occurring due to massive melting associated with huge impacts, similar to the Sudbury impact. Indeed, they also noted episodic clumping of zircon ages recording possible peaks in felsic crustal formation at 4.5 Ga (?); 4.2–4.3 Ga; 3.8–3.9 Ga; ?3.3–3.4 Ga, possibly due to mantle overturns or major plume episodes, each followed by 150–300 Ma of quiescence.

Hadean protocontinents were not like modern continents. Modern continents are characterised by highly felsic cores, high elevations, and a thick, stabilising underplating "keel". They are difficult to subduct. This was not the case during the Hadean. Since only highly metamorphosed remnants of Hadean protocrust survive, we will base our description on well-preserved crustal remnants dating from the Palaeoarchaean, the ~3.5 Ga Barberton and Pilbara Greenstone Belts, that can be used, to a certain extent, as proxies of older crust. However, these ancient terranes record only remnants of the upper parts of protocontinents but no "deep sea crust" *per se*. They document abundant mafic and ultramafic effusions (massive and pillow basalts) and intrusions. Most of the volcanic lithologies are tholeiitic, calk-alkaline or komatiite-type basalts. The latter particularly are characterised by higher Mg contents than present day volcanics (by several orders of magnitude) (Table 2), and were formed from hotter melts (Arndt et al., 2008). The ultra/mafic volcanics may also be accompanied by more fractionated lavas or intrusions, including andesites, dacites and rhyolites. Table 3 documents the characteristics of Archaean basalts arand Arndt et al., 2008).

The Eoarchaean-Palaeoarchean continental crust was characterised by thick layers of mainly ultramafic to mafic lavas and intrusions and sediments derived from these materials, interspersed with rarer, more fractionated, felsic lavas and intrusions. Formation of the early granites (tonalite-trondjemite-granitoids, TTGs, less rich in Si and K compared to modern granites) intruding into this crust appears to have been

TABLE 3 Characteristics of Archaean basalts (Barnes and Arndt, 2019).

Basalt type	Geological setting	Mineralogy	Major elements	Trace elements
Tholeitic	Monotonoous basaltic sequences	Plagioclase and pyroxene	Moderate MgO, TiO_2, Al_2O_3, FeO; Fe enrichment during fractionation	Flat REE, no Nb anomalies
Calc-alkaline	Associated with andesite, dacite, rhyolite	Plagioclase, clinopyroxene+/-oxides	High SiO_2 and Al_2O_3, moderate MgO, low TiO_2 and FeO; SiO_2-CaO-alkali enrichment during differentiation	Enriched LREE, negative Nb anomalies
Komatiitic	Associated with komatiite	Olivine and chromite on liquidus, spinifex textures	High MgO, low TiO_2, Al_2O_3, FeO	Al/Ti ratios match associated komatiites

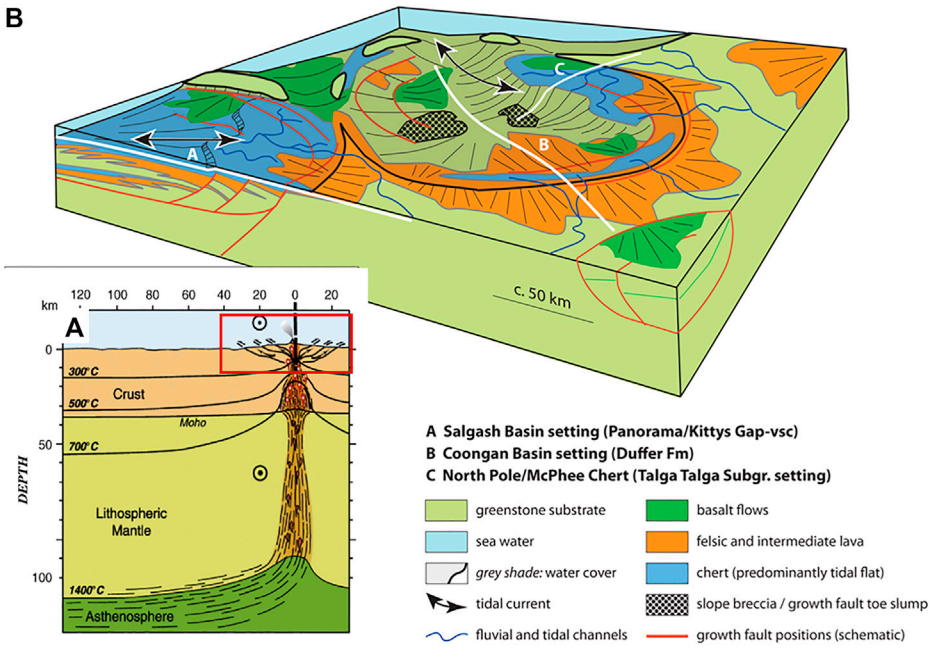


FIGURE 3
Schematic illustration of tectonic fracturing and basin formation on the early Earth. **(A)** A cross section of the early Earth's crust and mantle highlighting crustal fracturing around the intersection between the mantle plume and the top of the crust (outlined by a red box) (after Vauchez et al., 2012). **(B)** Block diagram model of the upper crust, equivalent to the red box in **(A)**, showing the formation of collapse basins bordered by faults and fractures in the softer crust of the Hadean and Palaeoarchaean resulting in the formation of shallow water basins on top of the oceanic plateaux/protocontinents (after Nijman et al., 2017).

related to lower crust melting rather than to plate tectonic subduction (Smithies et al., 2007). Indeed, structural and geochemical evidence of granites formed through plate tectonic subduction is not found in the Pilbara until after 3.2 Ga (Smithies et al., 2007).

Vanderhaeghe et al. (2019) suggest that the hotter Archaean crust was at least 40 km thick. Despite this, there is no evidence for widespread erosion of landmasses and exposure of the granitic cores of the early protocontinents, these were subdued landscapes with much of the continental crust submerged (Arndt and Nisbet, 2012). This means that the delivery of detrital material eroded from continents was lower than it is today. Thus, the early sediments comprised volcaniclastic detritus, either very locally derived from locally eroded rocks (the particles are generally euhedral to subhedral in shape, indicating lack of long transport by water, i.e. rivers), or ashfall into water bodies. These sedimentary materials were deposited into mainly shallow water basins and, although there is evidence of rare subaerial/deltaic fans, such evidence is rarely preserved, and one of the few examples is the 3.45 Ga Hoogenoeg Formation in Barberton (Lowe, 1999). Possible subaerial sediments have also been identified on the 3.75 Ga

Isua and 3.8 Ga Nuvviagittuq terranes (Fedo, 2000; Bolhar et al., 2005; Cates and Mojzis, 2007; O'Neill et al., 2011).

In a sedimentological-structural investigation of the basinal structures in the Pilbara Craton, Nijman and De Vries (2004) and Nijman et al. (2017) describe these features as collapse basins (Figure 3), similar to the coronae structures on Venus or early Mars (although not all researchers agree with this interpretation). These collapse basins are roughly circular in shape and of the order of 50 to several hundreds of kilometres in diameter. These basins were generally not very deep, since most sedimentary horizons were deposited in shallow settings (offshore to onshore), although some sedimentary sequences were deposited below wave depth (less than about 100 m). The basins formed on top of mantle plumes in the weaker (because hotter) Palaeoarchaean crust. Figure 3 illustrates collapse basins in the Pilbara Craton where the continental crust is mostly submerged, and only low relief volcanic edifices and surrounding areas emerged from the sea. In the example from the Pilbara Greenstone Belt shown in this figure, the emergent land masses are felsic volcanoes.

3.2 Hadean/Palaeoarchaean volcanic rock compositions

As representative examples of the mineralogical composition of Palaeoarchaean mafic lavas, tholeite basalts from the Kromberg Formation in the Barberton Greenstone Belt, South Africa, are comprised mainly of plagioclase (Na-Ca feldspar) and clinopyroxene (a Ca,Mg,Fe, Ti alumina silicate) with minor olivine and chromite (Vennemann and Smith, 1999) (Tables 2, 3). Titano-magnetites occur in the more evolved tholeites, while cumulates comprise only pyroxene and olivine. Komatiitic lavas (even more mafic than tholeites) are composed primarily of olivine and chromite (chrome spinels). They are characterised by anomalously high Mg (and Fe) contents owing to their very high temperatures of formation and eruption, up to >300°C hotter than basalts (Arndt et al., 2008). (Note that the original mineralogy of these Palaeaoarchaean basalts has been affected by seafloor alteration and subsequent metamorphism to actinolite-tremolite, chlorite, albite quartz, and rare, relict pyroxene.) Other Palaeoarchean basalt types include calk-alkaline basalts (enriched in magnesium and calcium oxides) comprise plagioclase, clinopyroxene and various metal oxides, such as magnetite.

In the Pilbara Greenstone Belt, there is much felsic volcanic material interlayered with the ultramafic and mafic lavas (Smithies et al., 2007). Here, more evolved lavas include andesite to dacitic rocks with sodic ratios of K_2O/Na_2O between .05 and .45 (van Kranendonk et al., 2007), as well as more enriched felsic rocks (low K_2O < 1.0 wt%, high Fe, HREE and Y concentratinos positively correlated with SiO_2 and La/Yb ratios) derived from fractionated tholeiitic lavas. Other felsic rocks have higher sodic ratios and have been derived from tonalite-trondgehemite-granodiorite (TTG) granites, while some highly potassic rhyolites may have been formed by remelting of pre-existing continental crust (van Kranendonk et al., 2007).

Examples of the compositions of andesitic (fine-grained lavas with phenocrysts of plagioclase and the ferromagnesian minerals, pyroxene and amphibole) to dacitic lavas from the Duffer Formation of the Pilbara Greenstone Belt (fractionated from basalts) are shown in Table 4. They include plagioclase porphyry (i.e. large plagioclase crystals (phenocrysts) in a fine-grained groundmass. Even more fractionated, porphyritic rhyolitic lavas occur in the Palaeoarchaean terranes, comprising more silica rich minerals, such as quartz and feldspars.

Typical of continental crust are granitic intrusions formed through fractionation of hydrated ultramafic to mafic crust, and sometimes remelting of previous granitic crust. In fact, the most common early crustal remnants are comprised predominantly of the early granitic cores of the protocontinents. However, as with the ultramafic lavas, the Hadean/Palaeoarchaean granites (TTGs) differed from their modern counterparts in their compositions, being less rich in potassium feldspathic minerals than modern granites (Moyen and Martin, 2013).

3.3 Volcanic sediments

While the early detrital sediments were mainly volcaniclastic in origin (including volcanic glass, protoliths, spherules, pseudomorphed feldspars and pyroxenes, as well as accessory minerals, such as chromites, zircon, rutile, quartz, carbonates, Fe oxides, and barite, see Lowe and Byerly, 2003), many of the sediments comprise varying amounts of carbonate minerals that formed diagenetically. All these lithologies have been significantly altered during early diagenesis to mostly Fe carbonate, sericite, chlorite, muscovite, and biotite followed by more or less complete replacement by silica, now forming the microcrystalline quartzitic rock, chert. With respect to the carbonates, given the diagenetic and metamorphic overprint, it is difficult to determine their original mineralogy, but Lowe and Knauth (1977) suggest local primary Fe-rich dolomite and ankerite for the Kromberg Formation in Barberton, now replaced by dolomite, siderite, ankerite and calcite. The Strelley Pool carbonates were Fe-dolomites (Shields, 2007).

Carbonates have also been described from the Isua Greenstone Belt in Greenland, but the calcites, Fe-dolomites and siderites are generally interpreted to be metasomatic deposits, although small amounts of carbonates associated with banded iron formations and metacherts may be original (van Zuilen et al., 2003).

3.4 Volcanic particle alteration

Mineral surfaces, such as clays, basaltic glasses, pyrite, or other reactive materials, are an important substrate for prebiotic

TABLE 4 Compositions of felsic lavas from the 3.46 Ga Duffer Formation and Panorama Formation at Kitty's Creek (Smithies et al., 2007), compared to recent (36 Ma) rhyolites from Texas (Elliott, 2018). The Palaeoarchaean rhyolites are enriched in Fe and Mg and are poor in K, compared to the younger rhyolites.

	Duffer formation, ~3,46 Ga	Panorama formation Kitty's creek, ~3.46 Ga	Round top rhyolite, Texas, 36 Ma
sample no.	179,726	180,223	262–01
SiO_2	70,77	74,64	75,5
TiO_2	0,31	33	0,02
Al_2O_3	12,9	15,93	13,31
Fe_2O_3	3,73	1,08	1,05
FeO			0,55
MnO	0,06	0	0,07
MgO	0,53	0,83	0,07
CaO	2,12	0,04	0,1
Na_2O	4,74	0,04	4,82
K_2O	2,31	5,38	4,28
P_2O_5	0,05	0,02	<0,01

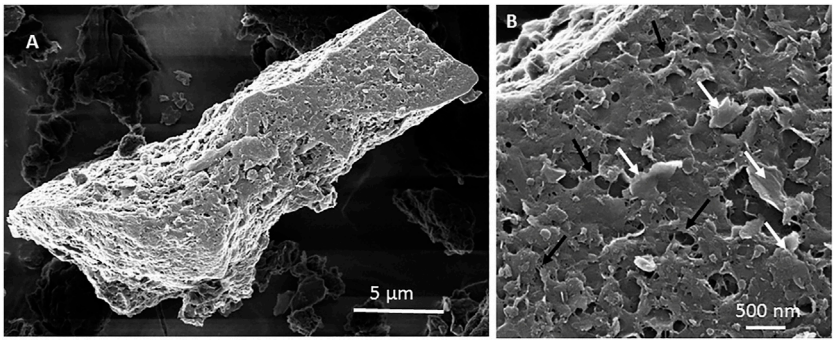


FIGURE 4
Alteration of the surface of a plagioclase fragment from crushed basalt by a Hadean seawater simulant at 73°C **(A)** showing rapid coating of the whole surface by minute clay particles **(B)**, white arrows. Note also that the mineral particle is coated with a "conditioning layer" of organic matter automatically precipitated from the ambient water (black arrows).

reactions, as demonstrated by the recent conversion of ribonucleoside triphosphates to polyribonucleic acid when incubated with basaltic glass (Jerome et al., 2022). In the hot, generally acidic waters of the Hadean oceans, inorganic surfaces of volcanic materials or meteorites, would have been rapidly altered. For example, in the preserved Palaeoarchaean rocks, most of the volcanic detrital particles were altered to phyllosilicate (probably some kind of smectite, now mostly sericite, chlorite or muscovite after metamorphism) and to anatase before being silicified (Lowe, 1999; Foucher et al., 2010; Westall et al., 2015b). Moreover, an experiment to alter basalt and obsidian surfaces in artificial Hadean seawater at 73°C showed that the surfaces were rapidly coated with a mixture of clay particles and organic matter, the former apparently forming in association with the organic film (Figure 4) (unpublished data, Westall). In this experiment, the mineral surfaces were coated by a "conditioning film" of organic molecules as soon as the material was exposed to seawater. This is a known, automatic chemical reaction caused by hydrostatic interaction between the hydrophobic organic macromolecules and the hydrophilic surface of materials immersed into water, and it takes place within minutes (Costeron et al., 1995; Westall et al., 2000). All natural water contains dissolved organic matter (DOM), all biologically derived, in the reduced form. Indeed, Catala et al. (2021) note that DOM in marine waters has one of the most diverse molecular compositions known, consisting of millions of

individual, low mass compounds, including compounds that are alicyclic, organic acids with amphiphilic properties (Hertkorn et al., 2006; Dittmar et al., 2008; Zark et al., 2017). While today much of the DOM is rapid recycled by life, some molecules have life times of >1,000 years. The latter situation would have been the case before the advent of cellular life in the prebiotic Hadean oceans. Today, it is estimated that there is more than 1 Eg (exagram, 1,018) of DOM on the planet (662 ± 32 Pg (petagram, 1,015) carbon; Hansell et al., 2009). DOM in the Hadean oceans would have been sourced from the mantle, hydrothermal crustal processes, and particularly from extraterrestrial meteorites and micrometeorites (Maurette, 2006). DOM from hydrothermal fluids would have been more common in the Hadean than today, given the higher crustal temperatures and more vigorous hydrothermal recycling. DOM in such fluids has been thermally altered (Rossel et al., 2017) and comprises more aromatic molecules and with less carboxyl-rich alicyclic species.

The close association of clay minerals with organic matter for prebiotic reactions has long been investigated (Cairns-Smith, 1966; Hartman, 1975; Ferris et al., 1979; Ferris et al., 1996). Interactions between the inorganic and organic phases are facilitated by adsorption, intercalation and cation exchange. Furthermore, negatively charged organic ions bind to clay minerals by positive edge charges or by the exchange of structural OH groups (Lagaly, 1984). Kloprogge and Hartman (2022) review present understanding with respect to the role of clay minerals, in particular mixed layer, Fe-clays, such as smectites, in the origin and development of metabolism, noting 1) the formation of amino acids on the surface of clay minerals on carbonaceous chondrites from simpler molecules, e.g., CO_2, NH_3, and HCN; and 2) the catalytic role of small organic molecules, such as dicarboxylic acids and amino acids found on carbonaceous chondrites, in the formation of Fe-clays themselves. Importantly, the evolution of metabolism can be replicated and catalysed by clays that can synthesise monomers, such as amino acids and nucleotides, that will subsequently polymerise, an important reaction for RNA-peptide worlds (op.cit.).

3.5 Hydrothermal activity and silica

One important component of these early sediments is chemically-precipitated silica (Dass et al., 2018; Westall et al., 2018; Ledevin et al., 2019). Hadean-Palaeoarchaean ocean waters were supersaturated in silica owing to global, widespread hydrothermal silica input (Hofmann and Harris, 2008), surficial weathering, and devitrification of the volcanic protoliths in water. Since chemical silica was deposited in rhythmic layers interspersed with other sediments, Ledevin et al. (2019) suggest that this was controlled by seasons, with silica precipitation occurring during winter months. The fact is that we have no idea what the yearly climate was like during the Palaeoarchaean era. Nevertheless, the rhythmic alternation of chemical silica layers with other sediment does indicate some kind of cyclical control, similar perhaps to that functioning for the banded iron formations.

One of the consequences of the early silica-saturated seawater was the pervasive silicification of all lithologies in contact with water, both sediments and volcanic rock surfaces. Silicification was extremely rapid in the vicinity of hydothermal activity in these facies (Westall et al., 2015b; Westall et al., 2018), and slower away from the direct influence of hydrothermal vents, where there is evidence of more advanced diagenesis. The degree of silicification and its rapidity has important consequences for prebiotic chemistry, as discussed below.

White smoker-type hydrothermal vents, with their associated carbonate deposits, are not recorded in the crustal record of Eo-Palaeaoarchaean cratons. They were either rare, did not exist, or were not preserved. They may have existed in the Hadean deep oceans. What do exist in the geological record are silica-rich vents that expulsed silica-rich fluids and, in the process, thoroughly silicified the surrounding lithologies (Hofmann and Harris, 2008; Westall et al., 2015a), forming pods of sediment around the conduits that are now >99.9% SiO_2 (Figure 5). Silicification implies permeation of all matter, particulate, solid rock or biogenic, by silica and replacement of the pre-existing feature by silica. Replacement may be more or less complete, leading to ghost structures; or partial, depending upon the degree of saturation of the seawater by silica. Other factors, such as temperature and pH, will have an effect on the rate of silicification.

One way to estimate the rate of silicification is the degree of diagenesis of the non-organic detrital and biogenic facies. Close to hydrothermal effluents, biogenic remains were so rapidly silicified that there was no time for diagenetic alteration that would lead to the precipitation of framboidal pyrite (sometimes associated with biogenic activity, such as degradation of pre-existing organic matter, e.g., dead cellular materials, by heterotrophic microbes, such as Sulphur reducing bacteria (SRBs), e.g., Duverger et al., 2020; Duverger et al., 2021). At further distance from hydrothermal activity, framboidal pyrite associated with degraded organic matter was common (F. Westall, personal observations). In an experiment to document the formation of pyrite in a pure microbial culture of sulphate reducing bacteria, Duverger et al. (2020) noted that spherules of pyrite a few hundreds of nm in size precipitated after 1 month of incubation. Indeed, pyrite framboids have not been described in association with well-preserved, cellular fossils from the Palaeaoarchaean sediments, confirming their rapid silicification (Westall et al., 2001; Westall et al., 2006a; Westall et al., 2006b; Westall et al., 2011a; Westall et al., 2011b). Superb cellular preservation by silicification is another means of estimating the time scales of silicification. The Palaeoarchaean cells mentioned above document preservation of intact cells

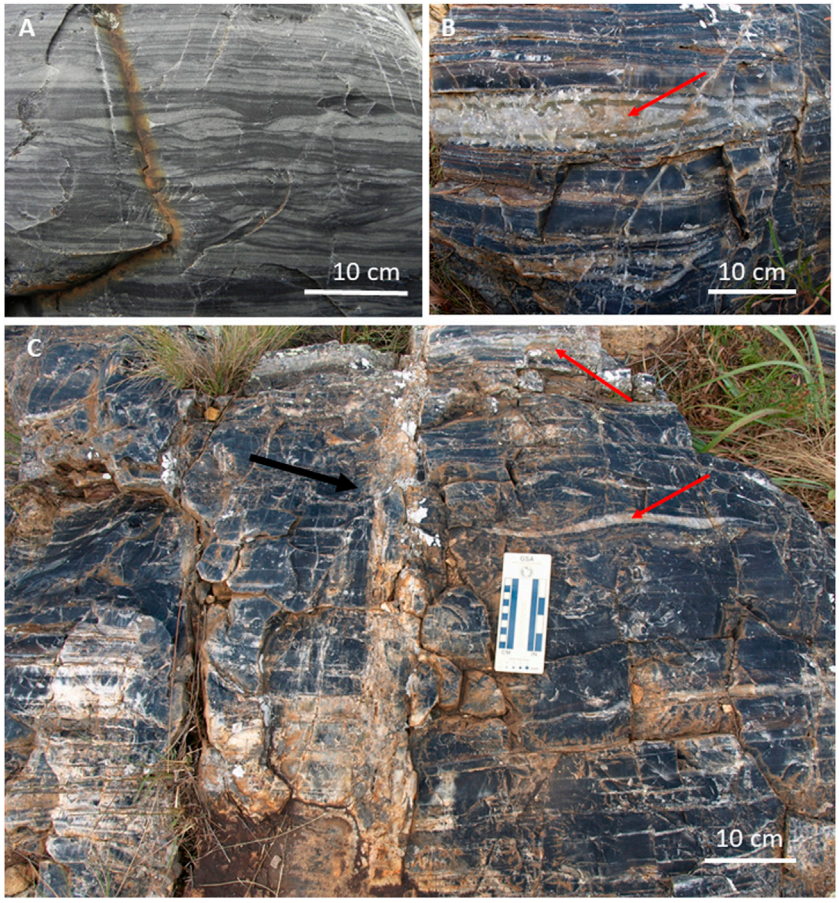


FIGURE 5
The effects of silicification on volcanic sediments, examples from the 3.33 Ga Josefsdal Chert, Barberton Greenstone Belt, South Africa. **(A)** Layers of volcanic ashfall deposited in a littoral setting and away from direct hydrothermal influence with a total silica concentration about 96% (@J.Bréhéret). **(B)** Similar sediments deposited close to hydrothermal activity with a total silica concentration >99.9%. Note the bedding parallel hydrothermal infiltration and possible vadose crack indicating subaerial exposure (red arrow). **(C)** Hydrothermal vein traversing the horizontally laminated shallow water sediments (black arrow) and lateral infiltrations of hydrothermal fluids into already slightly consolidated sedimentary layers (red arrows).

undergoing cell division, as well as cells and colonies in varying stages of degradation.

Experimental silicification of the kinds of thermophilic Archaea that could have existed on the early Earth, at temperatures of 60°C documents initiation of fossilisation already within 24 h (Orange et al., 2009). Note that the Palaeoarchaean shallow seas were very warm, ~50°C–76°C, water temperatures being estimated from oxygen and silicon isotopes, e.g., van den Boorn et al. (2007), Tartèse et al. (2017), and Lowe et al. (2014) In these cases, the organic substrates functioned as substrates for the chelation of silica, either directl.y onto the organic matter, or through cation bridging, e.g., with Fe. This was rapidly followed by direct chemical precipitation of silica in an acidic environment that formed a matrix cement around the organic substrate. Initially the silica would have presented a gel-like consistency that permitted permeation of fluids (Iler, 1979) and prebiotic reactions. As it polymerised and dehydrated, this permeable capacity would be lost. The rate of polymerisation of silica into silica gel depends on the saturation of silica, pH and temperature, as well as the presence of dissolved salts (Iler, 1979). We can conclude that, close to hydrothermal activity in the salty (Knauth, 2005, estimates early salinity to be 1.5–2.0 times present level because of the relative paucity of continents and enclosed water bodies in which halite could precipitate, consistent with estimates for salinity also on early Mars in Fairén et al., 2009), warm to hot environment of the early Earth, initial precipitation of silica was on the order of hours to days, while further polymerisation and dehydration leading to the cessation of prebiotic chemical reactions taking place within the gel, would have occurred on longer time scales. Further away

from direct hydrothermal influence, the rates of precipitation, polymerisation and dehydration would have been slower (time scales of one to a few months, cf. Duverger et al. (2020)), as documented by the more prolonged diagenetic alteration of the organic matter leading to framboidal pyrite precipitation.

The silica formed initially by polymerisation of monomeric silica in solution, nucleating to form ever larger particles depending upon the availability of the silica. The supersaturated Palaeoarchaean seawater thus precipitated silica, both within the sediments as well as on top of them forming layers of chemically precipitated silica gel (cf. Ledevin et al., 2019). Evidence of this chemically precipitated gel formation is seen in where colonies of microbes grew in three-dimensional spicular colonies within a mixture of fine volcanic dust and silica gel (Westall et al., 2015a; Hickman-Lewis et al., 2020a). Again, in these cases, the rarity of pyrite associated with the degraded, silicified organic matter is testimony to the rapidity of silicification.

The fact that silica gel was pervasive in the Palaeoarchaean oceans has important consequences for prebiotic reactions. Silica gel is a porous medium. Incorporating reactive volcanic detritus (including primary minerals, such as (rare) olivine, pyroxene, plagioclase feldspar, oxides e.g., chromite, and pyrite, and secondary minerals (alteration phases), such as clays, carbonates, anatase), the silica gel could act as myriads of microscopic reactors, especially when influenced by hydrothermal effluent (Dass et al., 2018; Westall et al., 2018). Here, gradients in pH, ionic concentrations and temperature from the vents through the immediately surrounding sediments would have been important factors for prebiotic reactions taking place in these sediments. Indeed, Pollack (2001), Trevors and Pollack (2005) and Dass et al. (2018) note that gels represent an excellent mechanism to maintain concentration gradients, alter the structure of water, and influence ion-macromolecule interactions. Gradients of protons and cations are conferred because the ability of gels to exclude solutes maintains the overall system out of equilibrium by creating an imbalance of solutes inside and outside the gel.

The rate of silica gel formation and its subsequent dehydration and lithification would have been an additional control on rates of prebiotic reactions. For example, using an artificially prepared solution of silica (sodium silicate) in artificial Hadean seawater, Dass et al. (2018); Dass et al. (2018) showed that the gel can form within hours upon addition of glacial acetic acid (pH 2.5). Of course, in the warmer Hadean waters, where pH would have been variable at the surfaces of the volcanic rocks and sediments (initially a relatively high pH 8, then decreasing after a couple of days to pH 6, Westall et al., 2018), the rate of silica gel formation would also have been influenced by the saturation of silica in the ocean waters, which, as we noted above, appears to have been controlled also by volcanism and direct hydrothermal input. Rapid silica gel formation would have occurred during the Hadean-Palaeoarchaean over time spans of hours to possibly days. With pore sizes of the order of 100 s nm to a couple of micrometres, these pervasive gels were a critical physical and chemical component of the early rock/sediment interfaces.

While the silica gel was in a porous state that allowed the transport of fluids, small molecules and dissolved components, prebiotic reactions could take place within the precipitate. However, dehydration and solidification of the gel that would put an end to these kind of reactions would depend on factors, such as temperature and pH. For example, Gallo and Klein (1988) note that silica gel dehydration is constrained by the diffusion of hydrogen-bonded species at low temperatures (<600°C), as would have been the situation in the Hadean oceans. Gels formed under basic conditions are structurally different to those formed under acidic ones (Orcel et al., 1986), the former being characterised by a macroporous structure with surface silanols and the latter by a microporous one with internal silanols. Acid-formed gels have a smaller scale silica network, larger surface area, smaller pores and larger pore interconnection. However, dehydration of silica gels formed under basic conditions is more rapid than those formed under acidic conditions. Sleep et al. (2001) and Sleep (2010) consider that the Hadean oceans were probably as neutral as they are today because of buffering of the seawater chemistry due to reactions with basalt. However (Ueda et al., 2021), suggest that fluids close to hydrothermal vents would have been probably more acidic at low temperatures owing to high water/rock ratios during the hydrothermal reactions (although pH would have been more alkaline at higher temperatures). Also, the predominantly CO2 atmosphere would have resulted in slightly acidic waters, as equally proposed for early Mars (Fairén et al., 2004). As noted above, while more alkaline alteration of the surfaces of Palaeoarchaean subaqueous basalts occurred at high temperatures (e.g., Robins et al., 2010), the pervasive silicification of the sedimentary horizons and the tops of the basaltic sequences is testimony to a lower temperature, silica-rich, acidic medium.

3.6 Fluvial input

In the Barberton craton, the situation is similar (Nijman and de Vries, 2004). As noted above, most of the sedimentary detritus derived from exposed volcanic lavas and volcanic ashfall. Evidence for riverine input into the shallow basins comes from geochemical data. Hickman-Lewis et al. (2020b) made *in situ* LA ICPMS measurements of microbial mat layers (i.e. relatively short-lived, ephemeral phenomena) that document flat, light REE-enriched REE + Y patterns and chondritic Y/Ho ratios indicating major contributions from terrigenous, riverine fluids, i.e. continental weathering, on a short temporal scale. This signature is superimposed on a long-term, background signature of predominantly hydrothermal origin (Figure 6).

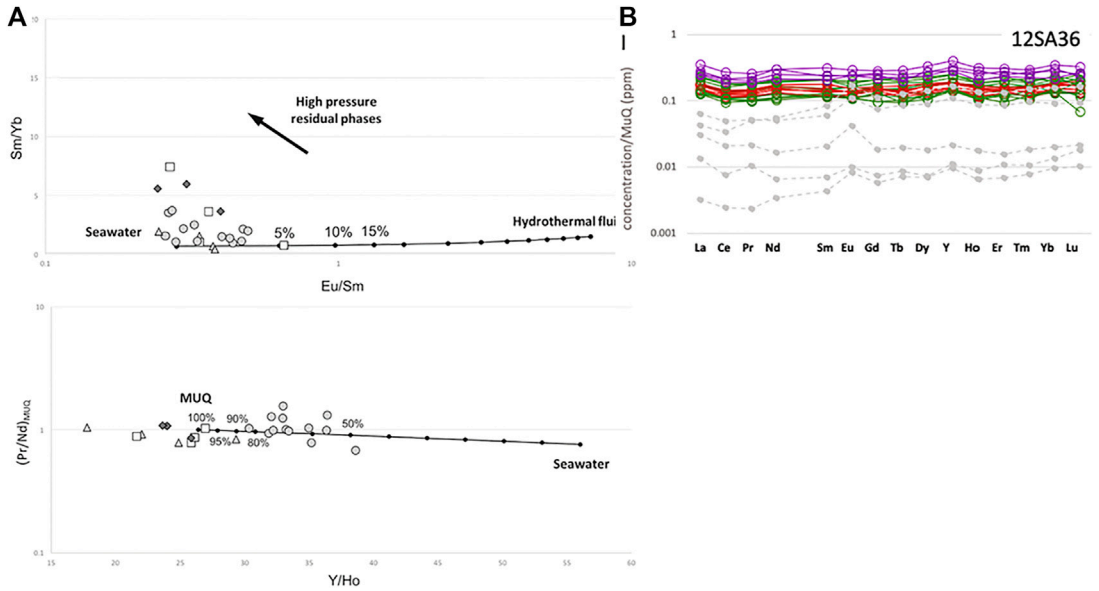


FIGURE 6
Geochemical evidence for seawater, hydrothermal and riverine signatures in the Palaeoarchaean shallow seas (the Josefsdal Chert, Barberton greenstone Belt). **(A)** The Sm/Yb and Eu/Sm ratios show a distinct hydrothermal influence, while the Pr/Nd and Y/Ho ratios document a distinct terrigenous influence **(B)** While bulk measurements of the Josefsdal cherts show the La, Eu and Y anomalies typical of hydrothermal fluids, *in situ* measurements document the flat REE distribution indicative of terrigenous, i.e., riverine input (After Hickman-Lewis et al., 2020b).

3.7 Solar irradiation

Numerous different types of locations have been suggested for the emergence of life (see review in Westall et al., 2018). The most commonly considered are hydrothermal environments but locations including associated with impact craters (Osinski et al., 2020; Sasselov et al., 2020), pumice rafts (Brasier et al., 2011; Brasier et al., 2013), deep seated fault zones (Schreiber et al., 2012), and radioactive placer sands (Adam et al., 2018) have also been invoked. For many of these environments, solar irradiance at the surface of the Hadean Earth, both on exposed landmasses and in water bodies, could have played an important role in the emergence of life. Pascal et al. (2013) hypothesise on the possible importance that UV radiation may have had as a free energy requirement for the origin of life (Figure 7), although not all prebiotic chemists adhere to this theory. Modelling suggests that solar irradiation at the surface of the early earth was much higher than today.

Solar X-ray and UV radiation covers wavelengths from .1 to 320 nm. The luminosity of the Hadean/Palaeoarchaean Sun was about 74%–77% of its present value (Bahcall et al., 2001) but the solar X-ray and UV (EUV and FUV) luminosity reaching the top of the early Earth's atmosphere was higher (Micela et al., 1985; Micela et al., 1988; Micela, 2002; Cnossen et al., 2007). Radiation at the surface of the early Earth would have been controlled by the composition of the atmosphere. Cnossen et al. (2007) modelled the spectral radiance of the young Sun on the basis of evolution of solar-like stars, taking into account factors, such as atmospheric composition, solar flares and activity cycles, as well as aerosol concentrations and cloud cover. They showed that shorter wavelength radiation (<200 nm) would have been attenuated by the atmosphere, while the flux of longer wavelengths (>200 nm) reaching the surface of the Earth would have been ~10^{-5} Wm^{-2} nm^{-1} at 200 nm, to ~.4 W m^{-2} nm^{-1} at 320 nm for small concentrations of CO_2 in the atmosphere (.024 bar). Higher CO_2 atmospheric concentrations (e.g., 1.20 bar) would have significantly attenuated the surface flux of radiation (~10–16 W m^{-2} nm^{-1} at 200 nm, to .02 W m^{-2} nm^{-1} at 320 nm. Cnossen et al. (2007) calculated the irradiance at the top of the atmosphere and the surface irradiance (180–320 nm because the shorter wavelengths do not reach the surface) for the early Earth (Archaean in this modelled case, i.e. 4.0–2.45 Ga) at different atmospheric CO_2 concentrations, compared with the DNA weighted spectra for different CO_2 concentrations (Cnossen et al., 2007). also demonstrated that small amounts of other gases in the atmosphere, such as N_2, CO_2, CH_4, H_2O, O_3, O_2, and SO_2 would have little effect on the attenuation of solar irradiation. They conclude that the amount of radiation reaching the early Earth's surface would have been damaging for the DNA of early life. From the point of view of Pascal et al. (2013), this high radiation is essential for providing the necessary initializing energy for prebiotic reactions in order for the "multiplying entities (to be) associated with the dissipation of free energy" (Figure 7).

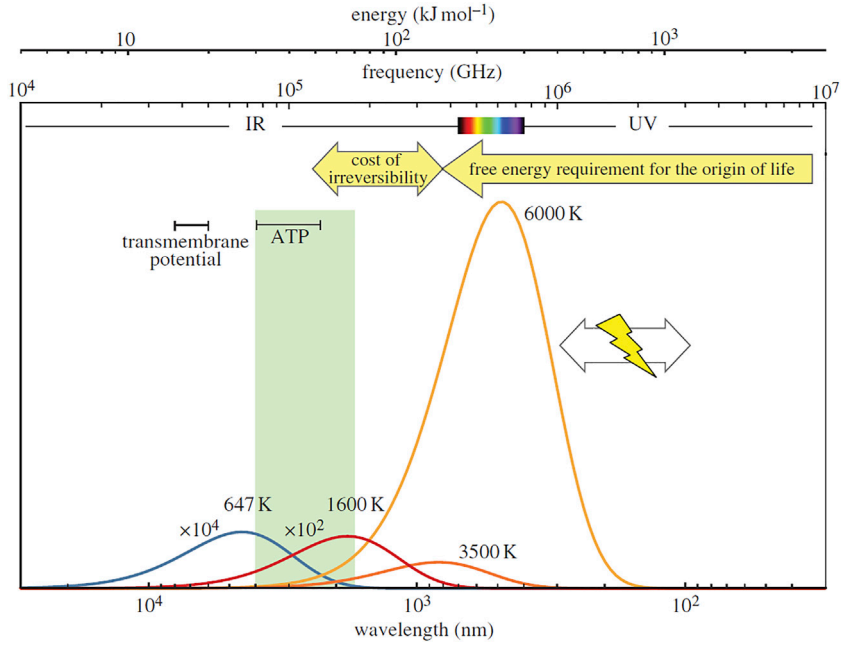


FIGURE 7
Free energy source requirements in living systems (after Pascal et al., 2013). The energy available at the critical point of water (at 647 K) is shown by the blue line; typical Hadean magma temperatures (~1600 K) are in red; the surface temperatures of examples of M-stars (3500 K) or G-stars (e.g. Sun, 6000 K) in dark and light orange lines; and lightning (T ≥ 104 K). Pascal et al. (2013) note that a much higher potential [ca 150 kJ mol−1 (42,43)] than the free energy potential of common biochemicals (green rectangle 30–70 kJ mol−1, including ATP) was required to trigger the self-organization of life after taking into account the cost of irreversibility (yellow arrows), namely visible and UV light, as well as lightning.

Solar radiation is, however, generally rapidly attenuated by water, reaching at maximum depths of a few metres, although note that blue-green light (UVA at 360 nm) can penetrate up to depths of 50–70 m in oceanic waters when the Sun is at its zenith (lee et al., 2013). Attenuation of radiation at all wavelengths will be increased depending upon water opacity. On the volcanically-active early Earth, water bodies would have carried much suspended and sedimenting particulate matter (volcanic debris, e.g., ashfall and detrital sediments) during eruptive phases, which would have significantly decreased solar radiation penetration depths. The fluvial contribution to the solid particulate load being deposited in shallow water basins adjacent to the exposed volcanoes would also have been higher during and after eruptions owing to the increased topography and erosion of the volcanic edifices. However, the intervening quiescent periods would have lower detrital input and hence suspended sediment load.

4 Characteristics of hypothesised locations for the emergence of life

Numerous different types of locations have been suggested for the emergence of life (see review in Westall et al., 2018). The most commonly considered are hydrothermal environments but locations including associated with impact craters (Osinski et al., 2020; Sasselov et al., 2020), pumice rafts (Brasier et al., 2011; Brasier et al., 2013), deep seated fault zones (Schreiber et al., 2012), and radioactive placer sands (Adam et al., 2018) have also been invoked but in this section we will concentrate on hydrothermal environments.

Of importance for any of the hydrothermal environments described below are gradients present, some of which, such as redox gradients, can fuel prebiotic reactions (Villafañe-Barajas and Colín-García, 2021). The various gradients include, temperature, solute concentrations, density, pH, as well as redox gradients. Reduced carbon compounds (e.g., HCOOH and CH_3OH) have been produced *via* thermal gradients and reversible reactions between dissolved gases (i.e. CO_2, CO and H_2) (Seewald et al., 2006). Moreover, it has been suggested that temperature gradients and oxidation–reduction reactions can contribute to the synthesis of organic molecules from gases, such as CO_2 and H_2 (Shock, 1993; Shock and Schulte, 1998; McDermott et al., 2015). Note, however, that for some metals, such as Fe and Mn, redox reactions can be kinetically slow (McCollom, 2000). Nevertheless, a recent experiment has shown that dipeptides may be formed by interaction of amino acids with minerals, such as olivine and orthopyroxene (Takahagi et al., 2019).

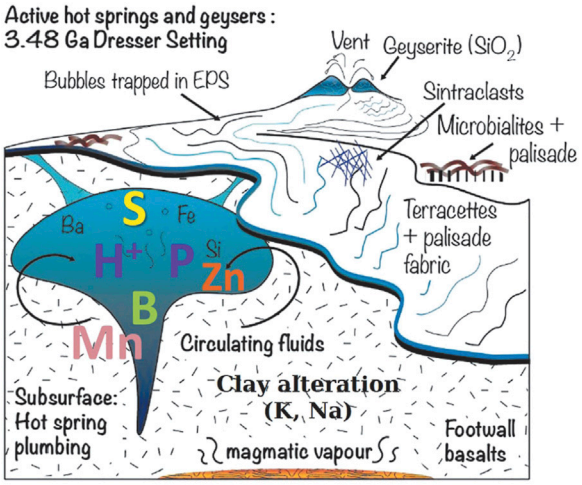


FIGURE 8
Schematic representation of the Dresser hot spring system and the elements it concentrated in hot spring pools and in the altered footwall (after van Kranendonk et al., 2021).

4.1 Subaerial springs

Recently, subaerial springs have been proposed as suitable locations for the emergence of life (van Kranendonk et al., 2018; Damer and Deamer, 2020; van Kranendonk et al., 2021). Today, subaerial springs are common in volcanic areas associated plate margin and intraplate, long-lived (10 s My) plume activity. During the Hadean, plumes on the hotter early Earth may have existed for longer, up to several 100 s Ma, judging by the ages of the Palaeoarchaean greenstone terranes, while individual volcanic eruptions and associated hydrothermal activity would have been more short-lived and episodic. For example, hot spring activity at Yellowstone over the North American mantle plume has been ongoing since at least the end of the last ice age (~15,000y) with the construction of large geyser cones taking place over thousands of years (USGS, 2022). A modern analogue of Hadean/Palaeoarchaean protocontinents with associated volcanic and hydrothermal activity is the Kerguelen plateau with its volcanic island archipelago. Characterised by mainly basaltic lavas with some felsic components intruded by plutonic rocks, the volcanic rocks date back to 39 Ma (Gagnevin et al., 2003), with individual lava complexes have a longevity of a couple of million years.

Palaeoarchaean suberial springs have been described from the Pilbara Greenstone Belt, while much of the hydrothermal activity in the 3.48 Ga Dresser Formation occurred under shallow water submarine conditions. In one case, some of the venting was associated with relatively restricted (<10 m) geyserite and siliceous sinter deposits in the North Pole Chert, typical of a subaerial hot spring field (Djokic et al., 2017; Van Kranendonk et al., 2018; Djokic et al., 2021) (Figure 8). Van Kranendonk et al., 2021 and references therein) describe a volcanic caldera setting characterised by voluminous, contemporaneous hydrothermal fluid circulation that formed a dense network of hydrothermal chert and baryte veins cutting through the underlying komatiitic pillow basalts and lowermost layers of the North Pole Chert. Van Kranendonk et al. (op.cit.) conclude that the sedimentary succession records "a transition upward from open marine, through a shallow, evaporative marine basin (caldera phase), to subaerial conditions with hot springs and fluvio-lacustrine deposition, followed by a return to deep marine conditions that resulted from caldera collapse with/without active extension (Djokic et al., 2017; Van Kranendonk et al., 2018; Djokic et al., 2021)".

The hydrothermal veins associated with the North Pole Chert succession are large, up to 50 m wide and also deep, penetrating down to depths of several kilometres. The tops of the veins are associated with the hot spring deposits and mineralised remnants of the hot spring pools (Van Kranendonk et al., 2007; Djokic et al., 2017; Van Kranendonk et al., 2018; Djokic et al., 2021). Fluid temperatures in the veins are estimated to have been about 350°C at depth and about 120°C at the palaeosurface (Harris et al., 2009). Interestingly, tourmaline crystals (containing boron) are associated with some of the hydrothermal deposits, and boron is of interest in prebiotic processes (cf. Ricardo et al., 2004), as noted above.

Although the individual Dresser springs were of limited dimension (Kranendonk et al., 2021), it is known that subaerial springs may occur in swarms of many hundreds or even thousands of individual vents, each with different physico-chemical parameters in terms of temperature, fluid chemistry (pH, ionic concentration, element composition), gas content, style of upwelling (passive flow, to bubbling, to geysers), that can vary widely over lateral scales of only a few meters (Campbell et al., 2015). These kinds of environments provide a wide range of possibilities for concatenating prebiotic reactions over periods of up to a couple of million years, according to the Kerguelen analogy.In an analysis of the chemical energy available at one subaerial spring (Obsidian Pool, Yellowstone), Shock et al. (1995) determined that the energy yields for different redox reactions involving reduced hydrothermal fluids depend largely on the electron acceptors. In this modern case, the electron acceptors include, O_2, nitrate, nitrite, elemental S, magnetite, hematite, goethite, sulfate, CO, and bicarbonate/CO_2 in order of the highest to lowest energy yields. On the Hadean Earth, O_2 would only have been available in very small quantities, formed by abiotic radiolysis of H_2O. Likewise, nitrate and sulfate were probably not present, and nitrite limited.

4.2 Submarine environments

Scenarios for the emergence of life in the submarine realm are mainly related to hydrothermal activity (Baross and Hofman, 1985; Russell and Hall, 1997; Martin et al., 2008). The submarine

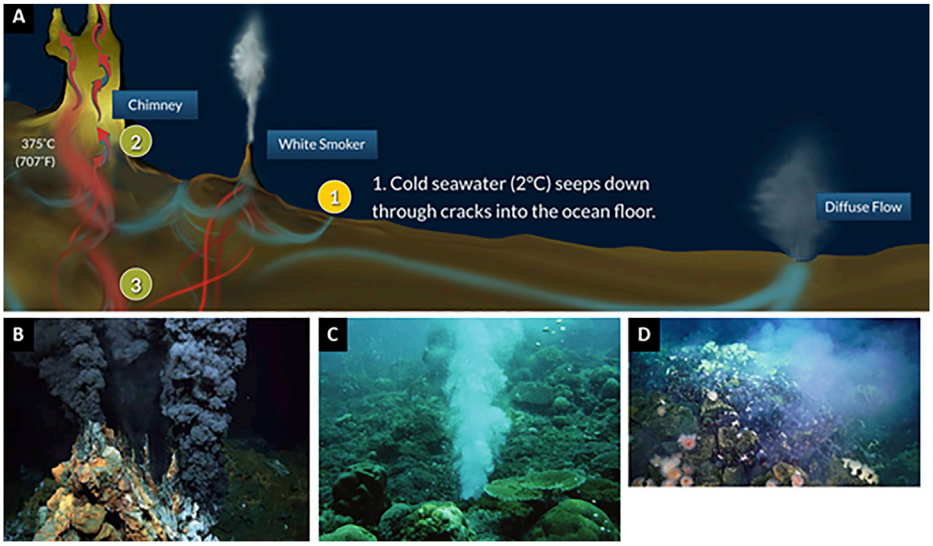


FIGURE 9
Deep sea hydrothermal vents today. **(A)** Sketch showing high temperature black smoker point sources (>350°C) on ridge axes, slightly lower temperature white smoker point sources (<300°C) off ridge axes, and more diffuse, low temperature (<100°C) sources further away from ridge axes. (sourced from the Woods Hole Oceanographic Institution). **(B)**. Black smoker vent (after Rogers et al., 2015). **(C)** White smoker vent (after Karlen et al., 2010). **(D)** Diffuse venting (sourced from Seveseas.media.org).

realm on the Hadean/Palaeoarchaean Earth comprised a wide range of environments with differing physico-chemical characteristics, and hosted above differing types of igneous substrates. These include the shallow water, volcano flanking environments on top of the oceanic plateaux that were the protocontinents (i.e. at littoral to subwave base water depths), as well as environments in the deep ocean associated with plume-related hot spots and, if they existed in the Hadean when life emerged, early tectonic spreading ridges. Redox reactions, such as those described at the beginning of Chapter 4 involving dissolved gases (e.g., CO_2, and H_2) (Shock, 1993; Shock and Schulte, 1998; McDermott et al., 2015) or metal sulphides as catalysts and H2S as a reductant (He et al., 2019) to form reduced carbon compounds (e.g., formate and acetate in the latter case).

4.2.1 Deep sea hydrothermal environments

There is a wide variety of hydrothermal systems forming in the deep sea environment, their physico-chemical characteristics controlled ultimately by the underlying igneous rock (and sediment) compositions, heat flow, and volcanic activity (Figure 9). They may be associated with actively spreading ridges and differ in style depending upon rate of ridge spreading, or they may occur at some distance from ridge axes, as well as above mid plate (or ridge/hot spot associations) mantle hot spots associated with submarine volcanism or seamounts. The hydrothermal systems may produce focussed vent edifices, as in the black smokers or some white smokers, or they may eject fluids in a more diffuse manner, especially for lower temperature systems that have a relatively strong component of intermixed seawater. Each style of venting, each hydrothermal field has different spatial dimensions and different longevity of activity, often with episodic activity. We have noted above that there is no preserved deep sea crust from the Hadean/Palaeoarchaean. However, hot spot activity and perhaps some form of early plate tectonics with spreading ridges, subduction and back arc basins may have been present (although there is a current consensus that plate tectonics did not start until after about 4 Ga, see Lammer et al., 2018; Dehant et al., 2019, or even later). Below we review the salient characteristics of the different types of deep-sea hydrothermal activity that could have been active during the Hadean.

Present day studies concerning the emergence of life in hydrothermal locations tend to concentrate on those representing processes occurring at or near high temperature (up to 400°C), acidic (pH 2–5), Fe-rich black smoker type vents that characterise spreading ridges (Corliss et al., 1981; Russell and Hall, 1997); and cooler, acidic Zn-rich white smoker vents (<300°C), also forming at or near ridge axes; or the even cooler (40°C–90°C), alkaline (pH 9–11), carbonate-rich vents (e.g., Russell, 2007; Martin et al., 2008) that form at some distance from the spreading ridges and emit hydrogen and methane. (Both the lower temperature white smoker and the even cooler alkaline vents are often referred to as “white smokers” in the literature). The latter environments are presently considered to be more likely for abiogenesis (Martin et al., 2008). These are environments characterised by well defined, focussed vents, as

well as more diffuse venting in the case of the white smokers. However, hydrothermal activity in the deep oceans is not just confined to ridge axes and areas off the ridge axes, it also occurs wherever there is volcanic activity, e.g. hot spots above mid plate mantle plumes or island arc calderas, for example. Note that, in this context, there is a significant amount of fluid flow (of a variety of temperatures) in three dimensions for up to hundreds of kilometres away from the spreading centre. This is demonstrated by the fact that the majority of the advective heat loss associated with the ridges actually occurs on their flanks, as well as in older crust and at lower temperatures than on-axis systems (Stein and Stein, 1992; Stein and Stein, 1994). This activity results in significant off-ridge hydrothermal circulation resulting in circulation of solute enriched fluids between the crust and ocean (e.g., Elderfield and Schult, 1996; Wheat and Mottl, 2004; Fisher and Wheat, 2010) and geochemical processes within the crust (e.g., Boschi et al., 2006; Bach and Früh-Green, 2010).

Black smoker vents occur in fields of many individual exit points directly associated with volcanic activity and distributed on scales of meters to tens of kilometres. They may be associated with point sources, or along linear, ridge-related fissures (Figure 9). In any one field, some vents may be active while others may be extinct. Pulses of hydrothermal activity will follow pulses of volcanic activity. Haymon et al. (2008) conclude that the life span of black smoker-type hydrothermal activity related to dyke intrusions at spreading ridges is relatively short, on the order of a few tens of years. On the other hand, models of melt extraction from a mantle plume produce volcanic pulses on timescales from ≤103 y (Schmeling, 2006), thus affecting the correlated hydrothermal activity above mid plate hot spots. For example, in the case of Hawaii and Iceland, hot spot volcanic activity varies over timescales of 101–103 years (Takada, 1999; Thordarson and Larsen, 2007).

Recently, fields of diffuse venting (Figure 9D) have been observed, for example the Von Damm vent field (between Cuba and South America), located away from the ridge axis, that exhale moderately low pH (6–7) fluids up to 215 C (Hodgkinson, 2015; Lough et al., 2019). Villafañe-Barajas and Colín-Garcia (2021) review submarine hydrothermal vents and their relevance for the origin of life, also emphasizing the widespread nature of diffusive venting.

(García M 2021) In these cases, infiltrating seawater mixes with subsurface hydrothermal fluids, in the process resulting in lower temperature, lower metal content, and less acidic pH fluids compared to the focused flow from chimneys (Lough et al., 2019). Diffuse, seepage of hydrothermal fluids is apparently widespread and makes up 60%–90% of the flux of hydrothermal effluent (German et al., 2016).

Scheirer et al. (2006) noted variations in temperature in both focussed and diffuse hydrothermal vents along the East Pacific Rise on the order of days to weeks. These were controlled by the presence of a shallow, subsurface reservoir of warm hydrothermal fluids that exited and mixed with ambient seawater during eruption/fissuring events at varying rates on a daily to weekly basis. Although this study related only to the temperature of the hydrothermal emissions, it is anticipated that other factors, such as pH and the concentration of dissolved trace metal species would be equally affected.

In terms of spatial scales and hydrothermal vent variability, a recent study of the la Scala vent field in the Woodlark Basin (NE of Australia) by Boulart et al. (2022) documented contemporaneous active and inactive areas, one comprising mainly diffuse vents over an area of 30 × 10 m, and a second area of vigorous black smoker vents in an area of 50 × 15 m. The latter is located above brecciated and altered basaltic rocks. While active venting is occurring in the black smoker area, it was also noted that a previous hydrothermal episode had occurred about 24,000 years ago, i.e. there are repetitive events on time scales of several 104 years.

Interestingly, experimental studies have recently demonstrated the formation mechanisms of the edifices precipitated around black and white smoker type vents. Cardoso and Cartwright (2017) demonstrated that black smoker growth is driven by thermal diffusion, while white smokers, similar to those of the Mid Atlantic Ridge field Lost City, are formed by much slower chemical diffusion. In the latter case, the increased contact between the effluent and the environmental fluids owing to the slower extrusion rate of the effluent and therefore the greater time available for interactions between the ambient seawater, results in fluid dynamics that leads to precipitation, producing a self-organized and self-assembled complex system, and allowing the controlled exchange of ions with the environment across a semipermeable membrane (cf. Russell et al., 1993).

Fields of white smoker vents are more long lasting because they are not directly connected to active volcanic activity and can be active for up to several 105 years. The Lost City white smoker field on the North Atlantic spreading ridge has been dated to more than 120,000 years (Ludwig et al., 2011), and has formed on an extremely slow spreading centre on top of ultramafic basement rocks. Present activity has been ongoing for about 30,000 years. The Lost City site is relatively large, covering about 500 m^2, and comprises both active and inactive vents locations.

4.2.2 Shallow water hydrothermal environments

We define shallow water environments as those that were above wave base. This is about 5–15 m for normal conditions today but up to 40 m for storm conditions. During the Hadean/Palaeoarchaean when higher temperatures, more frequent impacts and associated tsunamis (e.g., Lowe et al., 2003), as well as seismic and gravitational slumping would have created more unstable conditions, the storm wave base may have been deeper, perhaps up to 100 m. Shallow water hydrothermal environments during the Hadean/Palaeoarchaean were characterised by the influx of riverine runoff into the shallow

basins (cf. Hickman-Lewis et al., 2020b) that formed on top of the plateau-like protocontinents as "aprons" around exposed volcanic edifices. Thus, while the seawater in these basins has the positive Eu signature indicative of hydrothermal fluids that was global during these eras, and while other rare earth element indicators for episodically active, local hydrothermal input were common, the seawater was diluted by riverine input (op.cit.). In such shallow water basins, fresh water and seawater, infiltrating fractures in the basaltic crust, mixed to varying degrees with the subsurface hydrothermal fluids, resulting in an hydrothermal effluent with relatively low temperatures.

We will use the example of one, well-studied, sedimentary basin, the Josefsdal Chert, Barberton Greenstone Belt (3.33 Ga) (Westall et al., 2015a; Westall et al., 2018), to demonstrate hydrothermal activity in a shallow water basin (N.B., this basin is only one example, others exhibit different characteristics). The Josefsdal Chert is a sequence of volcaniclastic, chemical and biogenic sediments deposited on pillow lavas. In general, the sediments comprise three successive layers, during which hydrothermal activity fluctuated in importance depending upon the ambient volcanic setting. The lower layer consisting of volcanic ash fall and detrital volcanic components and was deposited directly on pillow lavas where it was very heavily dissected and infiltrated by hydrothermal silica, both during sedimentation and during early lithification of the sediments. The base of the unit comprises hyaloclastites of brecciated basaltic rock and hydrothermal silica, testifying to the explosive nature of the contact between seawater, hot lava and hydrothermal fluids. This sedimentary unit contains evidence of hydrothermal venting in a littoral, beach setting: here, phototrophic microbial biofilms and mats exposed on the beach surfaces were killed by hydrothermal outflow (as evidenced by the presence of still turgid filaments) and completely impregnated with silica (Westall et al., 2006b; Westall et al., 2011b).

Focussed venting took place during the sedimentation of this unit, as evidenced by vein density on the order of 1–2 m (vent features themselves are rare). Diffuse transport of hydrothermal fluids through the accumulating sediment was pervasive, as documented by the almost instantaneous silicification of the sediments, as well as the presence of mini veins and vents of cm size cutting through the sediments (Figures 9, 10). Hydrothermal effluent was also transported through large growth faults that controlled sedimentation thicknesses in many of the Palaeoarchaean basins (de Vries, 2004; de Vries et al., 2010; Westall et al., 2015a; cf Figure 3). Growth faults are normal faults that are active during sedimentation, enabling sediment to accumulate to great thicknesses on the downfault side, while condensed sediment sequences typify the upfault side. Their relevance is that they can be deep reaching and act as important conduits of hydrothermal fluids. Based on comparisons with sedimentation rates in modern volcanic environments, estimations for the duration of sedimentation (and hydrothermal activity) in this first phase, based on the thickness of the sediment layers, is of the order of 3.5–27 cm/ky. For a unit thickness of ~1 m, this means that sedimentation and contemporaneous hydrothermal activity continued, probably in an episodic manner, over periods of several 103 years. Volcanic and hydrothermal activity gradually waned and were replaced by a long period (105–106 y) of quiescence, during which chemical and biogenic sediments were deposited without being directly influenced by hydrothermal activity (Unit 2). This quiet period was brought to an end by renewed volcanic activity, producing once more point source injections of hydrothermal fluids which would influence the surrounding volcaniclastic sediments for up to 1–2 m (Unit 3) at the base of ashfall into very shallow waters (Unit 4).

Hydrothermal activity (point source and diffuse) was continuous throughout deposition of the basaltic volcanoclastic sedimentary units that could be traced laterally for up to a couple of kilometres (and was probably much more extensive, judging by similar volcanoclastic sedimentary deposits of the 3.472 Ga Middle Marker Chert from Barberton, Lanier and Lowe, 1982).

The type of hydrothermal activity in these shallow water environments appears to have been exclusively low temperature and silica rich (Hofmann and Harris, 2008; Westall et al., 2015a). Primary carbonate deposits, as in the deep sea white smokers of today, were not present, possibly because of the mainly acidic seawater conditions (note that carbonate did form as alteration products of pillow lavas directly exposed to seawater), or because if the relative concentrations of silicon and calcium in the hydrothermal fluids. Secondary carbonate as dolomite, ankerite or siderite were common diagenetic components of the sediments and precipitated contemporaneously with or after the silica gel. These shallow water environments characterised by lower temperature point and diffuse hydrothermal activity may have been suitable locations for the prebiotic chemistry and abiogenesis (Figure 10) (Westall et al., 2018).

High temperature, Fe-rich venting has not been documented in the shallow water environments, although there is one disputed: the iron-rich deposit in the Buck reef Chert (3.42 Ga) in Barberton, that has been described as an ironstone pod, similar to the Fe-rich deposits around deep sea black smokers (de Ronde et al., 1994). This particular deposit has subsequently been re-interpreted as a Quaternary spring, but the debate continues (Lowe and Byerly, 2003).

5 Open questions about the origin of life

Even if we have today a more coherent understanding of the environments of the early Earth, still we do not know how life began on our planet. It is generally assumed that life started in a

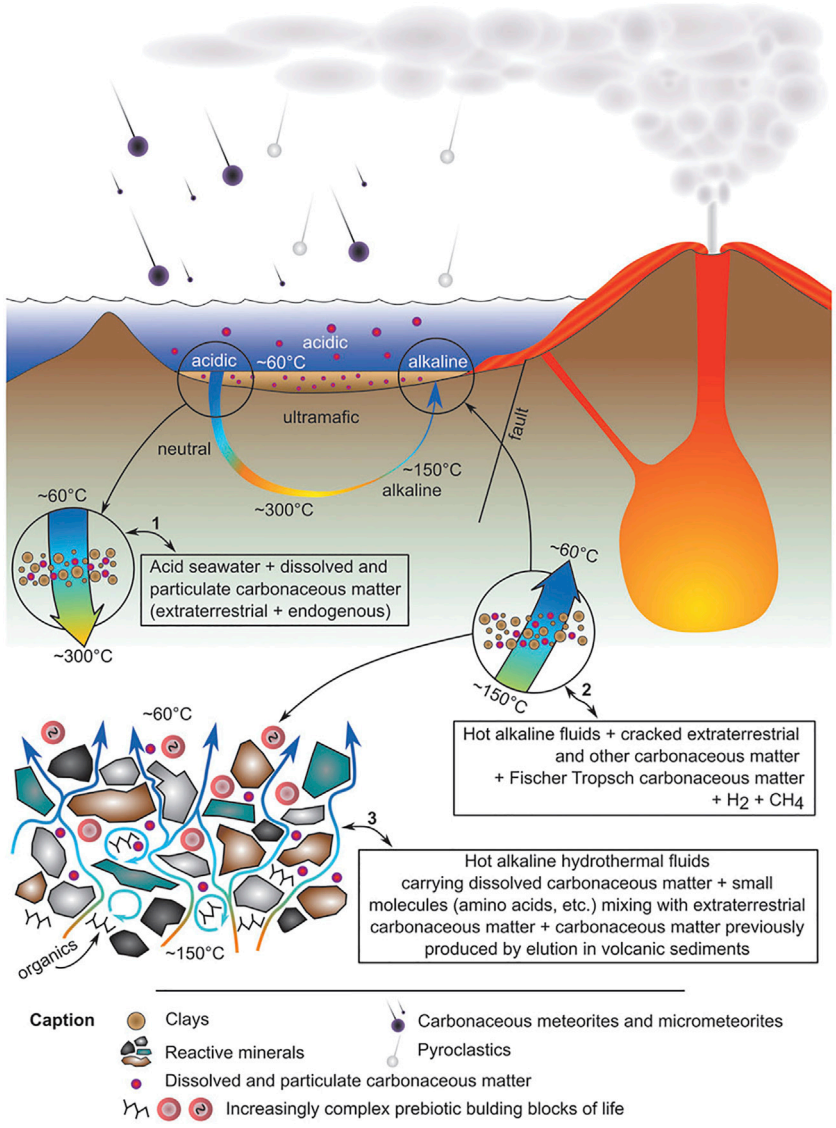


FIGURE 10
Circulation of hydrothermal fluids through shallow water volcanic sediments on the Hadean/palaeoarchaean Earth (after Westall et al., 2018). This scheme is based on data from well-preserved, Palaeoarchaean rocks of ~3.5–3.3 Ga. However, the geo-environmental situation of the Palaeoarchaean is considered to be analogous to that of the latter half of the Hadean, when life is believed to have emerged.

liquid water medium, because water is an ideal solvent (Brack, 1993). Water facilitates the formation of clay minerals *via* aqueous alteration of silicate minerals. Water is also a good heat dissipater, and for example, the products that are synthesized in hot vents could be rapidly quenched in the cooler surroundings thanks to this good heat conductivity. Imai et al. (1999) and Ogata et al. (2000) have demonstrated that oligomers of glycine can be formed under simulated thermal-quenching hydrothermal conditions. Water can also act as a discriminating driver for chemistry, as shown by the polymerization in water of a mixture of amino acids (including proteinaceous amino acids) containing both protein and non-protein amino acids, close to that found in the Murchison meteorite (Brack, 1987). Indeed, one of the most important aspects of water is its electrochemistry power (or redox chemistry), since metabolism can only occur in a polar fluid where electrons are easy to move around.

Theoretical considerations have been proposed for origins of life studies. The field, aiming to bridge chemistry and biology, must include "boundary conditions" as defined by Polanyi (Polanyi 1997) and commented by Paksi (Paksi 2014): structural boundary conditions as found in the field of

chemistry and control boundary conditions found in biology. De Duve mentioned "congruence", the notion that proto metabolism, the set of early chemical reactions that initiated and sustained incipient life, followed pathways that prefigured in many respects those of present-day, enzyme-catalyzed metabolism (de Duve 2003).

However, chemists have to run practical experiments in their laboratories. Even if the primitive environment is now better known, a reasonable scenario for the origin of life still faces the following pending questions:

5.1 An organic or mineral start to life?

It is generally assumed that primordial life was based on organic molecules containing carbon and hydrogen atoms associated with oxygen, nitrogen, and sulfur, the basis of current biology. However, some scientists imagined an alternative scenario based on inorganic substrates. Schneider (1977), for example, suggested that the complex dislocation networks found in some crystals could follow the definition of living units, proposing the idea of a crystalline physiology. According to Cairns-Smith (1982), there is no compelling reason to relate the first living cell to a last common ancestor made of organic molecules. He proposed that the first living systems, and the chemical evolution preceding them, could have been based on clay chemistry, different from life as we know it today. Although each step of this hypothetical sequence of events was developed in detail, the scenario has not been subsequently supported by experimental facts. Weiss (1981) published data that appeared to experimentally support Cairns-Smith's scenario but, unfortunately, the results could not be replicated. Since no satisfactory answers could be obtained, the clay-mediated replication hypothesis will not be further considered here.

5.2 Homemade organics from atmospheric gases, from hydrothermal systems or from impact processing?

In his historical experiment, Miller (1953) exposed a mixture of methane, ammonia, hydrogen, and water to spark discharges and silent electric discharge. In his initial experiment, he obtained three amino acids (glycine, alanine, and β-alanine) in the hot water within his experimental flask *via* the intermediary formation of hydrogen cyanide and aldehydes. The synthesis of amino acids occurred efficiently when using a reducing gas mixture containing significant amounts of hydrogen. However, the dominant view now is that the primitive atmosphere consisted mainly of CO_2, N_2, and H_2O, along with small amounts of CO and H_2 (Catling and Kasting, 2007). Only small yields of amino acids are formed in such a mixture (Schlesinger and Miller, 1983), though more recent studies show that the low yields previously reported could be the outcome of oxidation of the organic compounds during hydrolytic reprocessing by nitrite and nitrate produced in the reactions. The yield of amino acids is greatly increased when oxidation inhibitors, such as ferrous iron, are added prior to hydrolysis, suggesting that synthesis from neutral primitive atmospheres may have been more important than previously thought (Cleaves et al., 2008). Re-analyses of original sample residues from a previously unreported Stanley Miller electric discharge experiment, using a gaseous mixture of H_2S, CH_4, NH_3, and CO_2, were found to contain a total of 23 amino acids and 4 amines, including 7 organosulfur compounds (Parker et al., 2011). Miller also sparked a gas mixture of CH_4, NH_3 and H_2O while intermittently adding the condensing reagent cyanamide, and modern reanalysis of original sample residues of this experiment found a dozen amino acids, ten glycine-containing dipeptides, and three glycine-containing diketopiperazines (Parker et al., 2014).

The principal driver of organic synthesis is the disequilibrium between very different fluids. Thus, the reducing conditions in hydrothermal systems could have been also an important source of biomolecules on the primitive Earth (Holm and Andersson, 2005; Konn et al., 2015; Sojo et al., 2016; Damer and Deamer, 2020; Rimmer and Shorttle, 2019; White et al., 2020). The reducing environment results from substances dissolved in seawater interacting with inorganic compounds present in very hot crustal material (ultramafic rocks) that reduce them. These reduced compounds flow out of the hydrothermal system and the inorganic sulfides formed precipitate when they mix with cooler ocean water (4°C today but much warmer on the hotter, Hadean Earth, ~60°C–70°C). For example, hydrocarbons containing 16–29 carbon atoms have been detected in the Rainbow ultramafic hydrothermal system, Mid-Atlantic Ridge (Holm and Charlou, 2001). Hydrothermal vents are often disqualified as efficient reactors for the synthesis of bioorganic molecules, because of their high temperatures, but this applies only to the very hot (>300°C) black smokers, and not vast amount of oceanic crust that provides a three-dimensional substrate through which fluids are flowing, including the cooler (120°C) white smokers and diffuse venting. However, the products that are synthesized in hot vents are rapidly quenched in the surrounding cooler water, which may preserve those organics formed. Recently, hydrothermal-sedimentary environments have been suggested as prebiotic reactors for the origin of life (Westall et al., 2018).

Intense bombardment by meteors and comets probably caused some chemical reprocessing of the Earth's primitive atmosphere by impact shock chemistry (cf. Maher and Stevenson, 1988; Melosh and Vickery, 1989; Abramov et al., 2013; Mojzsis et al., 2019). This was demonstrated by Furukawa et al. (2009) who subjected a mixture of solid carbon, iron, nickel, water, and nitrogen to high-velocity impacts using a propellant gun. They recovered several organic molecules after the impact,

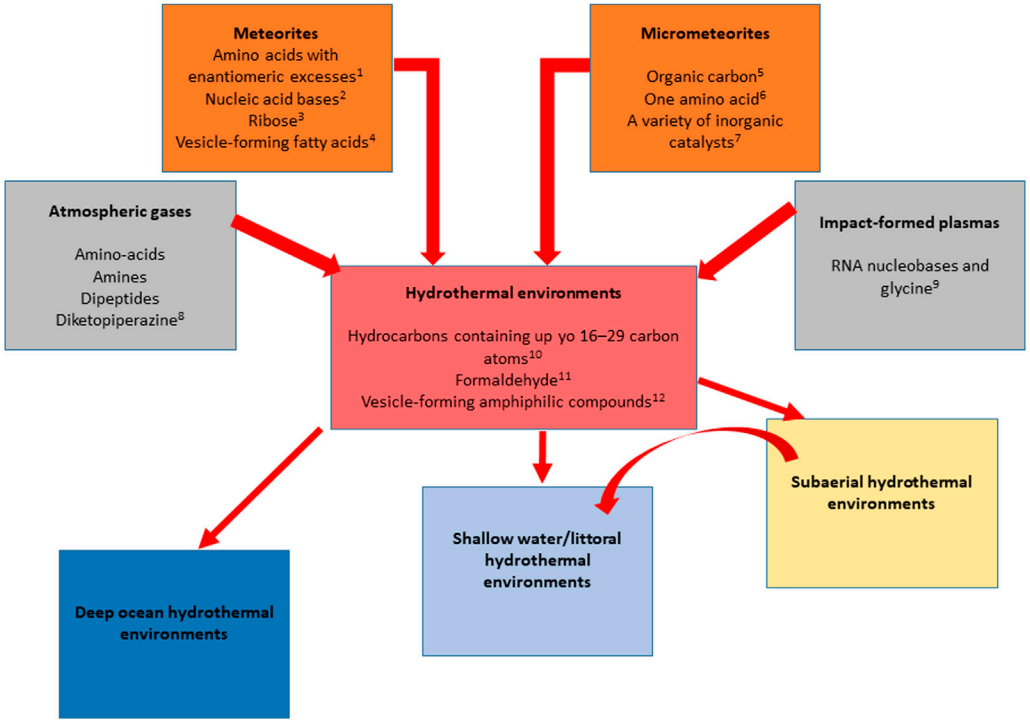


FIGURE 11
Block diagramme showing the prebiotic products arriving on, or formed on, the early Earth, and illustrating the interactions between different components and different environments. Note that there is no significant transfer of compounds between the subaerial and shallow water hydrothermal environments and those of the deep ocean; any transported molecules into the deep would be so diluted that their impact would be negligible. [1]Pizzarello et al. (2012), [2]Stoks and Schwartz (1982), [3]Furukawa et al. (2019), [4]Deamer (1998), [5]Dobrica et al. (2009), Rojas et al. (2021), [6]Matrajit et al. (2004), [7]Maurette (2006), [8]Parker et al. (2014), [9]Ferus et al. (2020), [10]Holm and Charlou (2001), [11]Herschy et al. (2014), [12]Damer and Deamer (2020).

including complex molecules, such as fatty acids and amines. Glycine, the simplest protein-building amino acid, was formed when the starting material contained ammonia, which is believed to have been formed during prior impacts on the early Earth. All the RNA canonical nucleobases and the simplest amino acid glycine were obtained when impacting a mixture of carbon monoxide, methane, and molecular nitrogen using a terawatt high-power laser system in the presence of montmorillonite (Ferus et al., 2020).

Figure 11 is a block diagramme showing the prebiotic products arriving on, or formed on, the early Earth, and illustrating the interactions between different components and different environments. Note that there is no significant transfer of compounds between the subaerial and shallow water hydrothermal environments and those of the deep ocean; any transported molecules into the deep would be so diluted that their impact would be negligible. We note here that some of the organic compounds found in the interstellar medium are fairly complex (up to 256 molecular species, Guélin and Cernicharo, 2022) and may form from radicals made by photolysis (high energy light, up to UV, Öberg, 2016). Some of this material may be the source of meteoritic material (although in the carbonaceous chondrite meteorites this was subsequently reprocessed by aqueous activity); the "kerogen" found in carbonaceous chondrites may be originally of interstellar origin (as shown by the carbon isotopic composition, Swart et al., 1983).

5.3 Imported extraterrestrial organics

A great number of organic molecules, including amino acids, have been found in carbonaceous chondrites. The Murchison meteorite, a CM2 type carbonaceous chondrite that fell in Australia in 1969, has been extensively analyzed (Shock and Schulte 1990; Pizzarello, 2007; Pizzarello and Shock, 2010; Pizzarello and Shock, 2017; Glavin et al., 2021). A combination of high-resolution analytical methods applied to the organic fraction of Murchison extracted under mild conditions documented its great chemical diversity of tens of

thousands of different molecular compositions and likely millions of diverse structures (Schmitt-Kopplin et al., 2010; Hertkorn et al., 2015). The soluble organic compounds found in the meteorite represent a diverse and abundant group of organics that range from small water-soluble compounds, such as amino acids and polyols, up to 30 carbon-long hydrocarbons.

Vesicle-forming fatty acids have been extracted from different carbonaceous meteorites (Yuen and Kvenvolden, 1973; Deamer, 1985; Deamer, 1998). Nucleic acid bases, purines and pyrimidines, have also been detected in the Murchison meteorite (Stoks and Schwartz, 1982), and ribose was recently identified in the soluble organic matter of carbonaceous chondrites (Furukawa et al., 2019). The total number of meteoritic amino acids detected is now about one hundred, including all the possible α-amino alkylamino acids up to seven carbon atoms, as well as large abundances of N-substituted, cyclic, β, γ, δ, and ε-amino acids, and eight protein-building amino acids. Most of the amino acids detected in carbonaceous chondrites are chiral but are present as racemic mixtures, i.e. L- and D-enantiomers are present in equal proportions. However, Cronin and Pizzarello (1997) found L-enantiomer excesses in six α-methyl-α-amino alkanoic acids from the Murchison (2.8%–9.2%) and Murray (1.0%–6.0%) carbonaceous chondrites, and an enantiomeric excess up to 18% has been measured for isovaline. These amino acids are either unknown or rare in the terrestrial biosphere and cannot therefore be attributed to terrestrial contamination (Pizzarello 2007). In addition, the endogeneity of D- and L-isovaline enantiomers is supported by carbon and hydrogen isotopic data (Pizzarello et al., 2003; Pizzarello and Huang, 2005). Large enantiomeric excesses up to 60% have also been found in Antartica Renazzo-type (CR) chondrites (Pizzarello et al., 2012). The meteoritic enantiomeric excesses figure prominently among the many hypotheses put forward to explain the emergence of homochirality, the one-handedness, of life. Looking for homochiral entities beyond the Earth would probably be one of the best ways to detect extraterrestrial life (Brack, 2019; Gleiser, 2022; Lee et al., 2022).

From dust collections in the Greenland and Antarctica ice sheet (Maurette, 1998; Maurette, 2006), it appears that the Earth captures interplanetary dust as micrometeorites at a rate of about 20,000 tons per year. The value has been lowered to 5,200 tons per year by analyzing recent melts of large volumes of ultra-clean snow at Dome C in Antarctica (Rojas et al., 2021). This value is much higher than the most reliable estimate of the meteorite flux, i.e., about 10 tons per year. At least, ~20 wt% of the micrometeorites survive unmelted upon atmospheric entry. As their kerogen fraction represents about 2.5 wt% of carbon, this amounts to a total mass of kerogen on the early Earth surface equivalent to a ~30 m thick global layer (Maurette and Brack, 2006), although this represents an accumulation rate of about 10–6 mm/y over a 500 Ma period. One amino acid, α-amino isobutyric acid, has been identified in Antarctic micrometeorites (Brinton et al., 1998; Matrajt et al., 2004). These grains contain also a high proportion of metallic sulfides, oxides and clay minerals (Folco and Cordier, 2015), a rich variety of inorganic catalysts which could have promoted the reactions of the carbonaceous material which led to the origin of life. Many similarities can be found between Antarctic micrometeorites and the comet Wild 2 samples, in terms of chemical, mineralogical, and isotopic compositions, and in the structure and composition of their carbonaceous matter (Dobrica et al., 2009).

Nevertheless, one of the issues with imported organic matter is what happens to it upon interaction with the early oceans? Certainly carbonaceous chondrites and micrometeorites would have served as raw material sources. However, in the hot, slightly acidic conditions of the early seawaters (especially in shallow water regions where mechanical disruption would have contributed to disaggregation of the exogenic rocks), this material would have been broken down and incorporated into the sedimentary deposits. While numerous layers of Palaeoarchaean extraterrestrial spherule beds are testimony to impacts and the presence of impacted rocks (Lowe et al., 2003; Lowe et al., 2014; Lowe and Byerly, 2015), the recent documentation of finely particulate extraterrestrial organic matter disseminated within early terrestrial sediment (Gourier et al., 2019) is testimony to the fine-scale incorporation of extraterrestrial organic matter into hydrothermal sediments and its availability for reworking, transformation and availability for prebiotic processes.

5.4 Primaeval soup or metabolism first for the inception of life?

Two distinct hypotheses have been put forward for the inception of life depending on the availability of organics. In a "metabolism-first" approach, the proponents of an autotrophic life call for the direct formation of simple molecules from carbon dioxide that rapidly evolve to life (Wächtershäuser, 2007). In this scenario, the energy is provided by chemical and thermal disequilibrium. Shock et al. (1995) and McCollom and Shock (1997) note that the geochemical constraints on organic formation and early metabolism are readily met in hydrothermal systems. The energy source required to reduce CO2 could have been provided by the oxidative formation of pyrite from iron sulphide and hydrogen sulphide, giving rise to a two-dimensional mineral surface metabolism. Along with the scenario proposed by Michael Russel (Martin et al., 2008), a laboratory set up simulating conditions prevailing in alkaline hydrothermal vents generated low yields of simple organics (Herschy et al., 2014). So far, the proponents of a "metabolism-first" approach have not been able to produce large enough precursor prebiotic molecules.

In the second hypothesis, the primeval soup scenario, also labelled "replication-first", preformed complex organic molecules accumulated in a warm little pond, ‹‹ à la ›› Darwin. Targeting a primitive living cell-like system, great efforts are still being deployed in laboratories to produce boundary molecules providing compartmentalization, protein enzymes and RNA, using either hydrothermal conditions, wet-dry cycling or minerals. Nevertheless, we note that wet-drying cycles are not necessary for dehydration reactions.

For compartmentalization, amphiphilic compounds assemble into membranous vesicles in hydrothermal hot spring water (Monnard and Walde, 2015; Milshteyn et al., 2018; Damer and Deamer, 2020; Deamer et al., 2019). Experiments that demonstrate how different prebiotically-available building blocks can become precursors of vesicle-forming phospholipids have been reviewed elsewhere (Fiore and Strazewski, 2016). Mixtures of C10–C15 single-chain amphiphiles form vesicles in aqueous solutions at temperatures of ~70 C in the presence of isoprenoids and under strongly alkaline conditions (Jordan et al., 2019). Vesicles functionalized with RNA and peptides could have been an interesting step towards the formation of early protocells (Izgu et al., 2016). The possible origins of a cellular life have been reviewed elsewhere (Schrum et al., 2010).

As for prebiotic peptides, seminal articles include the possible role of short peptides in the steps towards the formation of a protocell (Fishkis, 2007), the formation of peptides on oxide surfaces (Lambert et al., 2013; Kitadai et al., 2017), mineral surface chemistry control for origin of prebiotic peptides (Erastova et al., 2017), ester-mediated amide bond formation driven by wet–dry cycles (Forsythe et al., 2015), or the formation and self-assembly of long prebiotic polypeptides produced by the condensation of non-activated amino acids on oxide surfaces (Martra et al., 2014). Frenkel-Pinter et al. (2020) exhaustively reviewed the present state of the art in this question. More recently, abiotic synthesis and chain extension of unique peptide isomers from free amino acids have been obtained in aqueous microdroplets (Holden et al., 2022). Self-replicating catalytic amyloid peptides have been hypothesized as prebiotic informational and protometabolic entities (Maury, 2018; Rout et al., 2022).

RNA is generally claimed as being the most important molecule for the origin of life (Budin and Szostak, 2010; Bernhardt, 2012; Higgs and Lehman, 2015; Benner et al., 2020). In this respect, significant progress has been made in "one-pot" syntheses. Sutherland's team, in a relevant example of system chemistry, showed that precursors of ribonucleotides, amino acids and lipids can all be derived by the reductive homologation of hydrogen cyanide and some of its derivatives, and thus that all the cellular subsystems could have arisen simultaneously through common chemistry (Patel et al., 2015). A systems chemistry approach has been recently developed (Strazewski, 2019). The synthesis of pyrimidine nucleosides driven solely by wet-dry cycles has also been reported (Becker et al., 2018). In the presence of phosphate-containing minerals, 5′-mono- and diphosphates also form selectively in one-pot reactions (Becker et al., 2019). Diamidophosphate efficiently phosphorylates a wide variety of potential building blocks, nucleosides/nucleotides, amino acids and lipid precursors, under aqueous conditions. Significantly, higher-order structures, oligonucleotides, peptides and liposomes, are formed under the same phosphorylation reaction conditions (Gibard et al., 2018). Abiotic synthesis of purine and pyrimidine ribonucleosides has been obtained in aqueous microdroplets (Nam et al., 2018). The chemistry of abiotic nucleotide synthesis has been reviewed elsewhere (Yadav et al., 2020).

Despite these encouraging results, it still seems unlikely that life could have started with RNA molecules, because whole RNA strands are not simple enough and yet too difficult to assemble under prebiotic conditions. Our preliminary approximation to answer the question asked is that a promising avenue will perhaps consist in far from equilibrium, open wet-dry cycling organic reactions, occurring on mineral surfaces under hydrothermal-like conditions, and generating a "replication-first" scenario in the primeval soup.

5.5 Which time scale for the emergence of life?

How long did it take for life to emerge? We do not know, but the fact that the process had to be driven by a kinetic momentum such that there could be no turning back, no undoing of the haphazard combination of molecules to form a protocell, means that it had to have occurred on shorter time scales rather than longer ones, because of the more unstable surface conditions reigning on the early Earth. However, given the uncertainty of our understanding of conditions on the early Earth, all that we can state is that, not knowing how RNA-based life appeared, it is not possible to estimate the time necessary for life to emerge (Orgel, 1998). This also depends on exactly which steps in the emergence of life are considered in the processes leading to the appearance of an auto-catalytic entity. Do we include the time required for the creation of the prebiotic building bricks of life, or simply the assemblage of the pre-formed molecules? In the latter case, the process could have started very rapidly or even instantaneously, like solutions that suddenly crystallize upon the addition of a crystal acting as a germ. There is also the possibility that life may have emerged more than once in different places on the planet but could have been extinguished by changing environmental parameters, such as a catastrophic meteoroid or asteroid impact (more common during the Hadean (Marchi et al., 2021), and re-emerged. Could life have emerged simultaneously in numerous locations, where one form was more efficient than the others and took over? These are questions we may not be able to answer on Earth, but perhaps future exploration

of Mars, where rocks more than 4 Ga are preserved, may provide some insight into this possibility.

We take, as the context for these considerations, our initial premises, that life emerged after the Moon-forming impact about 4.51 Ga (Barboni et al., 2017) and was well-established and relatively evolved by about 3.5 Ga (Noffke et al., 2013; Hickman-Lewis et al., 2018) if not before (e.g., Hassenkam et al., 2017). Zircon crystals offer strong evidence for crustal water at about 4.19 Ga but habitable conditions conducive to prebiotic chemistry probably existed well before (Sleep et al., 2014). How much time did it take between the emergence of life and advanced life? This would have been a process, not an event, rather than individual, isolated cells, perhaps some kind of symbionts?

5.6 What to do next to succeed in creating a simple life form in a test tube?

By demonstrating in 1953 that it was possible to form amino acids—the building blocks of proteins - from methane, Stanley Miller generated the ambitious hope that chemists will be able to create life in a test tube. Despite the tremendous efforts of chemists tackling the problem, the dream has not yet been accomplished. Obviously, experiments should adhere to geologically relevant environments pertinent to the early Earth. A rational methodology involving chemical systems on heterogeneous geological surfaces has been proposed, based on the relationship between prebiotic chemistry and the geological conditions of the Hadean Earth (Dass et al., 2016). A recently-created consortium of researchers seeking to understand the origins of life on Earth and in the Universe will hopefully advance the field (Müller et al., 2022). Novel apparatuses for incorporating natural selection processes have been also proposed (Root-Bernstein and Brown, 2022).

For future research, one approach could be far-from-equilibrium wet-dry cycling (either subaerial exposure or dehydration through chelation to mineral surfaces) of organic reactions occurring repeatedly and iteratively at mineral surfaces under hydrothermal-like conditions, an approach merging geology and chemistry (although we noted above that wet-dry cycling is not essential to concentrating organic molecules). Also, information from biology shows that early life flourished at relatively high temperatures (Woese, 1987; Miller and Lazcano, 1995; Gaucher et al., 2003; Camprubi et al., 2019 and references therein). Slight enantiomeric excesses of the reactants, as well as clays and other minerals, should be used. Concerning clays, they formed as soon as liquid water was present on Earth's surface through aqueous alteration of the early volcanic rocks. Bernal (1949) listed the advantageous features of clays, i.e., their ordered arrangement, large adsorption capacity, shielding capacity against sunlight, ability to concentrate organic chemicals, and ability to serve as polymerization templates. Given the importance of catalysts for prebiotic chemistry, minerals, such as clays were probably the most efficient partners of the primitive game of life (Brack, 2013; Hansma, 2022; Kloprogge and Hartman 2022) but other minerals with reactive surfaces were present on the early Earth, such as spinels, could have been equally useful (Iqubal et al., 2016). The use of microdroplets represents also an interesting avenue for the synthesis of both peptides and nucleotides (e.g., Wang et al., 2021).

6 Concluding reflections

In the first place, constraining the time period in which life emerged on Earth is difficult. We have taken as boundaries for this event the oldest evidence for liquid water at the Earth's surface (onset of habitability conditions) and the oldest evidence for life.

The oldest evidence of water comes from the oldest zircon crystals having formed by fractionation of melts in the crust that date back to 4.19 Ga (Whitehouse et al., 2017 and references therein), although oxygen and silicon isotope measurements of older zircons (4.4–4.3 Ga) indicate interaction of hydrothermal fluids with crustal materials (Mojzsis et al., 2001; Wilde et al., 2001; Trail et al., 2018). However, there is a debate about the real age of these older zircons (Whitehouse et al., 2017). Modelling suggests that liquid water on Earth's surface could have existed at habitable temperatures earlier than 4.2 Ga (Sleep, 2010; Sleep et al., 2014). Further constraints on habitability can be obtained by comparative planetology, using Mars as a model to better understand surface processes in the terrestrial planets during the first billion years: the very ancient geological record on Earth has been erased on our very active planet, by plate tectonics, and atmosphere, hydrosphere and biosphere interactions with the surface, and there is a consequent lack of preservation of the Hadean crust. Conversely, on Mars, all major geologic activity came to an end more than 3 Ga ago, and therefore the preservation of the early habitability environmental conditions is better (e.g., Fairén et al., 2010). Decades of robotic exploration have confirmed that early Mars was, indeed, as habitable as early Earth (Domagal-Goldman et al., 2016; McMahon et al., 2018; Moser et al., 2019).

There is suggestive evidence for life on Earth (carbon isotope ratios, hydrocarbon molecules) in highly metamorphosed rocks from the 3.75 Ga Isua terrane, West Greenland (Rosing, 1999; Hassenkam et al., 2017) and possibly in older rocks from 4.1 Ga (carbon isotope rations, Bell et al., 2015). Certainly, by 3.5 Ga life was well established on Earth and had diversified to include chemotrophs as well as phototrophs (e.g., Byerly et al., 1986; Walsh, 1992; Hofmann et al., 1999; Tice and Lowe, 2004; Westall et al., 2015a; Hickman-Lewis and Westall, 2021). We can conclude that life likely emerged after about 4.3 Ga, but

certainly before 3.75 Ga, a time uncertainty equivalent to the entire Phanerozoic.

Prebiotic chemistry needs input from biology, geology and geochemistry in order to constrain the conditions of the environments in which life might have emerged on Earth. While large-scale processes, such as formation of emerged landmasses and proto-continents, as well as tectonic regime and fluid flow, formed the backdrop, it is in the local-scale environment that life emerged. These local-scale environments were characterised by mainly mafic and ultramafic rocks, as well as volcanic rocks of more felsic composition. They, in turn, were defined by specific mineralogical associations and geochemical compositions (notably important trace elements), including the alteration products of these protoliths (clays, anatase, other). While each local environment would have been characterised by local conditions in terms of temperature, pH, or ionic concentrations, the gradients essential for driving prebiotic chemistry would have been provided by heterogeneity in the local conditions driven by physical factors (tides, drying cycles, hydrothermal effusions) as well as chemical and other gradients.

One important factor generally neglected in prebiotic chemistry is the significance of hydrothermal silica gel in the early terrestrial aqueous environments. Owing to the more or less ubiquitous hydrothermal activity, silica-rich effluent was common in the early oceans. Added to this was silica from devititrification of the volcanic sediments, silica input *via* rivers, and the concentration of silica in shallow waters by dehydration (as an evaporate mineral). The presence of silica gel could have been an important pathway for the concentration of, and dehydration of, prebiotic molecules. It also places constraints on the rates of processes occurring in submarine (and subaerial spring) environments because dehydration and lithification of the local environment (hydrothermal vents and surrounding sediments) was generally fast.

We note the advantages for the emergence of life that would be provided subaerial environments, such as hot springs located around the edges of collapsed volcanic calderas, where subaerially-exposed surfaces gave access to UV radiation that provided the energy for initiating the prebiotic processes (if it was necessary, and not all share this view). While individual springs may not be long lasting, they occur in swarms that, collectively, can last several tens of thousands of years. We also consider hydrothermal systems (vents and their surrounding sediments), where reactive mineral surfaces were important for prebiotic reactions. Such systems probably existed in the deep ocean (for which there is no crustal preservation), and definitely existed on the shallow water platforms that were the protocontinents, as well as in the littoral beach area. All of these individual environments have individual life times but on different scales. While hydrothermal activity associated with mantle plumes may be ongoing for more than tens of thousands of years, individual springs/vents last for tens of years, but occur in swarms.

On top of these context, considerations still remain that represent fundamental, outstanding questions about the origin of life. In the first place, was the start to life inorganic or organic? The evidence seems both and inorganic and an organic start, metabolism coming from inorganic chemistry but the carbonaceous components clearly from organic sources. While many of the organic molecules used in early prebiotic processes were of extraterrestrial origin, there is growing awareness of the endogenous formation of organic molecules in the environment of the early Earth, in the atmosphere, in the crust, or in the clouds of volcanoes. One of the major controversies concerns the timing of the appearance of early metabolism or of the first replication molecule, or even the bounding membranes. How long did it take for the first cells to emerge? We simply do not know. The answer depends on how an RNA world appeared and on exactly which steps in the emergence of life are considered in the processes leading to the appearance of an auto-catalytic entity. It also depends on the time required for the creation of the prebiotic building bricks of life or simply the assemblage of the pre-formed molecules. In the latter case, the process could have started very rapidly or even instantaneously, like solutions that suddenly crystallize upon the addition of a crystal acting as a trigger.

Author contributions

FW devised the study; all authors participated in the writing.

Acknowledgments

FW acknowledges ISSI Bern for providing a nicely quiet environment for cogitation and writing. AGF was supported by the Project "MarsFirstWater", European Research Council Consolidator Grant No. 818602. The authors are very grateful to the reviewers for their insightful comments.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abramov, O., Kring, D. A., and Mojzsis, S. J. (2013). The impact environment of the Hadean Earth. *Geochemistry* 73, 227–248. doi:10.1016/j.chemer.2013.08.004

Adam, Z. R., Hongo, Y., Cleaves, H. J., Yi, R., Fahrenbach, A. C., Yoda, I., et al. (2018). Estimating the capacity for production of formamide by radioactive minerals on the prebiotic Earth. *Sci. Rep.* 8, 265–268. doi:10.1038/s41598-017-18483-8

Albarede, F., and Juteau, M. (1984). Unscrambling the lead model ages. *Geochim. Cosmochim. Acta* 48, 207–212. doi:10.1016/0016-7037(84)90364-8

Allegr, C. J., Manhes, G., and Göpel, C. (2008). The major differentiation of the Earth at ~4.45 Ga. *Earth Planet Sci. Lett.* 267, 386–398.

Allwood, A. C., Rosing, M. T., Flannery, D. T., Hurowitz, J. A., and Heirwegh, C. M. (2018). Reassessing evidence of life in 3, 700-million-year-old rocks of Greenland. *Nature* 563, 241–244. doi:10.1038/s41586-018-0610-4

Allwood, A. C., Walter, M. R., Kamber, B. S., Marshall, C. P., and Burch, I. W. (2006). Stromatolite reef from the early archaean era of Australia. *Nature* 441, 714–718. doi:10.1038/nature04764

Appel, P. W. U., Fedo, C. M., Moorbath, S., and Myers, J. S. (1998). Recognizable primary volcanic and sedimentary features in a low-strain domain of the highly deformed, oldest known (≈ 3.7–3.8 Gyr) Greenstone Belt, Isua, West Greenland. *Terra nova.* 10, 57–62. doi:10.1046/j.1365-3121.1998.00162.x

Arndt, N. T., Barnes, S. J., and Lesher, C. M. (2008). *Komatiite*. Cambridge, UK: Cambridge Univ. Press., 487.

Arndt, N. T., and Nisbet, E. G. (2012). Processes on the young Earth and the habitats of early life. *Ann. Rev. Earth Planet Sci.* 40, 521–549. doi:10.1146/annurev-earth-042711-105316

Asphaug, E., Emsenhuber, A., Cambioni, S., Gabriel, T. S., and Schwartz, S. R. (2021). Collision chains among the terrestrial planets. III. Formation of the moon. *Planet Sci. J.* 2, 200. doi:10.3847/psj/ac19b2

Avice, G., and Marty, B. (2014). The iodine-plutonium-xenon age of the Moon-Earth system revisited. *Roy. Soc. Lond Ser. A* 372, 20130260. doi:10.1098/rsta.2013.0260

Bach, W., and Früh-Green, G. L. (2010). Alteration of the oceanic lithosphere and implications for seafloor processes. *Elements* 6, 173–178. doi:10.2113/gselements.6.3.173

Bahcall, J. N., Pinsonneault, M. H., and Basu, S. (2001). Solar models: Current epoch and time dependences, neutrinos, and helioseismological properties. *Astrophysical J.* 555, 990.

Barboni, M., Boehnke, P., Keller, B., Kohl, I. E., Schoene, B., Young, E. D., et al. (2017). Early formation of the moon 4.51 billion years ago. *Sci. Adv* 3, e1602365. doi:10.1126/sciadv.1602365

Barnes, S. J., and Arndt, N. T. (2019). "Distribution and geochemistry of komatiites and basalts through the Archean," in *Earth's oldest rocks*. Editor M. J. van Kranendonk, 103–132.

Baross, J. A., and Hoffman, S. E. (1985). Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Orig. Life Evol. Biosph.* 15, 327–345. doi:10.1007/bf01808177

Becker, S., Feldmann, J., Wiedemann, S., Okamura, H., Schneider, C., Iwan, K., et al. (2019). Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. *Science* 366, 76–82. doi:10.1126/science.aax2747

Becker, S., Schneider, C., Okamura, H., Crisp, A., Amatov, T., Dejmek, M., et al. (2018). Wet-dry cycles enable the parallel origin of canonical and non-canonical nucleosides by continuous synthesis. *Nat. Commun.* 9, 163–171. doi:10.1038/s41467-017-02639-1

Bell, E. A., Boehnke, P., Harrison, T. M., and Mao, W. L. (2015). Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proc. Nat. Acad. Sci.* 112, 14518–14521. doi:10.1073/pnas.1517557112

Benner, S. A., Bell, E. A., Biondi, E., Brasser, R., Carell, T., Kim, H-J., et al. (2020). When did life likely emerge on Earth in an RNA-first process? *ChemSystemsChem* 2, e1900035. doi:10.1002/syst.201900035

Bernal, J. D. (1949). The physical basis of life. *Proc. Phys. Soc. Sect. A* 62, 537–558. doi:10.1088/0370-1298/62/9/301

Bernhardt, H. S. (2012). The RNA world hypothesis: The worst theory of the early evolution of life (except for all the others)a. *Biol. Direct* 7, 23. doi:10.1186/1745-6150-7-23

Bolhar, R., Kamber, B. S., Moorbath, S., Whitehouse, M. J., and Collerson, K. D. (2005). Chemical characterization of Earth's most ancient clastic metasediments from the Isua greenstone belt, Southern West Greenland. *Geochim. Cosmochim. Acta* 69, 1555–1573. doi:10.1016/j.gca.2004.09.023

Boschi, C., Früh-Green, G. L., Delacour, A., Karson, J. A., and Kelley, D. S. (2006). Mass transfer and fluid flow during detachment faulting and development of an oceanic core complex, Atlantis Massif (MAR 30 °N). *Geochem Geophys Geosyst* 7, Q01004. doi:10.1029/2005GC001074

Boulart, C., Rouxel, O., Scalabrin, C., Le Meur, P., Pelleter, E., Poitrimol, C., et al. (2022). Active hydrothermal vents in the Woodlark Basin may act as dispersing centres for hydrothermal fauna. *Commun. Earth Environ.* 3, 64–16. doi:10.1038/s43247-022-00387-9

Boyet, M., and Carlson, R. W. (2005). 142Nd evidence for early (>4.53 Ga) global differentiation of the silicate Earth. *Science* 309, 576–581. doi:10.1126/science.1113634

Brack, A. (2019). "Chemical biosignatures at the origins," in *Biosignatures for Astrobiology*. Editors B. Cavalazzi and F. Westall (Springer), 3–21.

Brack, A. (2013). "Clay minerals and the origin of life," in *Handbook of clay science*. Editors F. Bergaya and G. Lagaly (Amsterdam, Netherlands: Elsevie), 507–518.

Brack, A. (1993). Liquid water and the origin of life. *Orig. Life Evol. Biosph.* 23, 3–10. doi:10.1007/bf01581985

Brack, A. (1987). Selective emergence and survival of early polypeptides in water. *Orig. Life Evo. Biosph.* 17, 367–379. doi:10.1007/bf02386475

Brasie, M. D., Matthewman, R., McMahon, S., and Wacey, D. (2011). Pumice as a remarkable substrate for the origin of life. *Astrobiology* 11, 725–735. doi:10.1089/ast.2010.0546

Brasier, M. D., Matthewman, R., McMahon, S., Kilburn, M. R., and Wacey, D. (2013). Pumice from the ~3460Ma apex basalt, western Australia: A natural laboratory for the early biosphere. *Precambrian Res.* 224, 1–10. doi:10.1016/j.precamres.2012.09.008

Brasser, R., Mojzsis, S. J., Werner, S. C., and Abramov, O. (2021). A new estimate for the age of highly-siderophile element retention in the lunar mantle from late accretion. *Icarus* 361, 114389. doi:10.1016/j.icarus.2021.114389

Brinton, K. L. F., Engrand, C., Glavin, D. P., Bada, J. L., and Maurette, M. (1998). A search for extraterrestrial amino acids in carbonaceous Antarctic micrometeorites. *Orig. Life Evol. Biosph.* 28, 413–424. doi:10.1023/a:1006548905523

Budin, I., and Szostak, J. W. (2010). Expanding roles for diverse physical phenomena during the origin of life. *Annu. Rev. Biophys.* 39, 245–263. doi:10.1146/annurev.biophys.050708.133753

Byerly, G. R., Lower, D. R., and Walsh, M. M. (1986). Stromatolites from the 3, 300–3, 500-myr Swaziland supergroup, Barberton mountain land, South Africa. *Nature* 319, 489–491. doi:10.1038/319489a0

Cairns-Smith, A. G. (1982). *Genetic takeover*. Cambridge, U.K: Cambridge University Press.

Cairns-Smith, A. G. (1966). The origin of life and the nature of the primitive gene. *J. Theor. Biol.* 10, 53–88. doi:10.1016/0022-5193(66)90178-0

Campbell, K. A., Guido, D. M., Gautret, P., Foucher, F., Ramboz, C., and Westall, F. (2015). Geyserite in hot-spring siliceous sinter: Window on earth's hottest terrestrial (paleo)environment and its extreme life. *Earth Sci. Rev.* 148, 44–64. doi:10.1016/j.earscirev.2015.05.009

Camprubí, E., De Leeuw, J. W., House, C. H., Raulin, F., Russell, M. J., Spang, A., et al. (2019). The emergence of life. *Space Sci. Rev* 215, 56–53. doi:10.1007/s11214-019-0624-8

Caracausi, A., Avice, G., Burnard, P. G., Füri, E., and Marty, B. (2016). Chondritic xenon in the Earth's mantle. *Nature* 533, 82–85. doi:10.1038/nature17434

Cardoso, S. S., and Cartwright, J. H. (2017). On the differing growth mechanisms of black-smoker and Lost City-type hydrothermal vents. *Proc. Roy. Soc. Lond A* 473, 20170387. doi:10.1098/rspa.2017.0387

Caro, G., Bourdon, B., Birck, J. L., and Moorbath, S. (2003). 146Sm-142Nd evidence from Isua metamorphosed sediments for early differentiation of the Earth's mantle. *Nature* 423, 428–432. doi:10.1038/nature01668

Catalá, T. S., Shorte, S., and Dittma, T. (2021). Marine dissolved organic matter: A vast and unexplored molecular space. *App Microbiol. Biotechnol.* 105, 7225–7239. doi:10.1007/s00253-021-11489-3

Cates, N. L., and Mojzsis, S. J. (2007). Pre-3, 750 Ma supracrustal rocks from the Nuvvuagittuq supracrustal belt, northern Québec. *Earth Planet Sci. Lett.* 255, 9–21. doi:10.1016/j.epsl.2006.11.034

Catling, D., and Kasting, J. F. (2007). "Planetary atmospheres and life," in *Planets and life*. Editors W. T. SullivanIII and J. A. Baross (Cambridge, U.K.: Cambridge University Press), 91–116.

Cavosie, A. J., Valley, J. W., Wilde, S. A., and E.i.m.f. (2005). Magmatic δ18O in 4400–3900 ma detrital zircons: A record of the alteration and recycling of crust in the early archean. *Earth Planet. Sci. Lett.* 235, 663–681. doi:10.1016/j.epsl.2005.04.028

Cleaves, H. J., Chalmers, J. H., Lazcano, A., Miller, S. L., and Bada, J. L. (2008). A reassessment of prebiotic organic synthesis in neutral planetary atmospheres. *Orig. Life Evol. Biosph.* 8, 105–115. doi:10.1007/s11084-007-9120-3

Cnossen, I., Sanz-Forcada, J., Favata, F., Witasse, O., Zegers, T., and Arnold, N. F. (2007). Habitat of early life: Solar X-ray and UV radiation at Earth's surface 4–3.5 billion years ago. *Geophys. Res. Planets* 112 (E2).

Connelly, J. N., Bizzarro, M., Krot, A. N., Nordlund, Å., Wielandt, D., and Ivanova, M. A. (2012). The absolute chronology and thermal processing of solids in the solar protoplanetary disk. *Science* 338, 651–655. doi:10.1126/science.1226919

Connelly, J. N., and Bizzarro, M. (2016). Lead isotope evidence for a young formation age of the Earth–Moon system. *Earth Planet Sci. Lett.* 452, 36–43. doi:10.1016/j.epsl.2016.07.010

Corliss, J. B., Baross, J. A., and Hofman, S. E. (1981). An hypothesis concerning the relationship between submarine hot springs and the origin of life on Earth. *Oceanol. Acta* 4, 59–69.

Costerton, J., Lewandowski, D. E., Caldwell, D. E., Korber, D. R., and Lappin-Scott, H. M. (1995). Microbial biofilms. *Ann. Rev. Microbiol.* 49, 711–745. doi:10.1146/annurev.mi.49.100195.003431

Cronin, J. R., and Pizzarello, S. (1997). Enantiomeric excesses in meteoritic amino acids. *Science* 275, 951–955. doi:10.1126/science.275.5302.951

Dalrymple, G. B. (2001). The age of the earth in the twentieth century: A problem (mostly) solved. *Geol. Soc. Lond* 190, 205–221. doi:10.1144/gsl.sp.2001.190.01.14

Damer, B., and Deamer, D. (2020). The hot spring hypothesis for an origin of life. *Astrobiology* 20, 429–452. doi:10.1089/ast.2019.2045

Dass, A. V., Hickman-Lewis, K., Brack, A., Kee, T. P., and Westall, F. (2016). Stochastic prebiotic chemistry within realistic geological systems. *ChemistrySelect* 1, 4906–4926. doi:10.1002/slct.201600829

Dass, A. V., Jaber, M., Brack, A., Foucher, F., Kee, T. P., Georgelin, T., et al. (2018). Potential role of inorganic confined environments in prebiotic phosphorylation. *Life* 8, 7. doi:10.3390/life8010007

De Duve, C. (2003). A research proposal on the origin of life. *Orig. Life Evol. Biosphere* 33, 559–574. doi:10.1023/a:1025760311436

de Ronde, C. E. J., de Wit, M. J., and Spooner, E. T. C. (1994). Early Archean (>3.2 Ga) Fe-oxide-rich, hydrothermal discharge vents in the Barberton greenstone belt, South Africa. *GSA Bull.* 106, 86–104. doi:10.1130/0016-7606(1994)106<0086:EAGFOR>2.3.CO;2

de Vries, S. T., Nijman, W., and de Boer, P. L. (2010). Sedimentary geology of the palaeoarchaean Buck ridge (South Africa) and kittys gap (western Australia) volcano-sedimentary complexes. *Precamb Res.* 183, 749–769. doi:10.1016/j.precamres.2010.09.005

de Vries, S. T. (2004). *Early archaean sedimentary basins: Depositional environment and hydrothermal systems*. Doctoral dissertation. Utrecht: University Utrecht.

Deamer, D., Damer, B., and Kompanichenko, V. (2019). Hydrothermal chemistry and the origin of cellular life. *Astrobiology* 19, 1523–1537. doi:10.1089/ast.2018.1979

Deamer, D. W. (1998) *Membrane compartments in prebiotic evolution*. In: A. Brack (ed) p. 189–205.

Deamer, D. W. (1985). Boundary structures are formed by organic components of the Murchison carbonaceous chondrite. *Nature* 317, 792–794. doi:10.1038/317792a0

Dehant, V., Debaille, V., Dobos, V., Gaillard, F., Gillmann, C., Goderis, S., et al. (2019). Geoscience for understanding habitability in the solar system and beyond. *Space Sci. Rev* 215, 42–48. doi:10.1007/s11214-019-0608-8

Dittmar, T., Koch, B., Hertkorn, N., and Kattner, G. (2008). A simple and efficient method for the solid-phase extraction of dissolved organic matter (SPE-DOM) from seawater. *Limnol. Oceanogr. Methods* 6, 230–235. doi:10.4319/lom.2008.6.230

Djokic, T., Van Kranendonk, M. J., Campbell, K. A., Havig, J., Walter, M. R., and Guido, D. M. (2021). A reconstructed subaerial hot spring field in the *3.5 billion-year-old Dresser Formation, North Pole Dome, Pilbara craton, western Australia. *Astrobiology* 21, 1–38. doi:10.1089/ast.2019.2072

Djokic, T., Van Kranendonk, M. J., Campbell, K. A., Walter, M. R., and Ward, C. R. (2017). Earliest signs of life on land preserved in ca. 3.5 Ga hot spring deposits. *Nat. Commun.* 8, 15263. doi:10.1038/ncomms15263

Dobrica, E., Engrand, C., Duprat, J., Gounelle, M., Leroux, H., Quirico, E., et al. (2009). Connection between micrometeorites and Wild 2 particles: From antarctic snow to cometary ices. *Meteorit. Planet Sci.* 44, 1643–1661. doi:10.1111/j.1945-5100.2009.tb01196.x

Dodd, M. S., Papineau, D., Grenne, T., Slack, J. F., Rittner, M., Pirajno, F., et al. (2017). Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* 543, 60–64. doi:10.1038/nature21377

Domagal-Goldman, S. D., Wright, K. E., Arney, G., Atri, D., Azua-Bustos, A., Bowman, J. S., et al. (2016). The Astrobiology primer v2.0. *Astrobiology* 16, 561–653. doi:10.1089/ast.2015.1460

Duverger, A., Berg, J. S., Busigny, V., Guyot, F., Bernard, S., and Miot, J. (2020). Mechanisms of pyrite formation promoted by sulfate-reducing bacteria in pure culture. *Front. Earth Sci.* 8, 588310. doi:10.3389/feart.2020.588310

Duverger, A., Bernard, S., Viennet, J-C., Miot, J., and Busigny, V. (2021). Formation of pyrite spherules from mixtures of biogenic FeS and organic compounds during experimental diagenesis. *Geochem, Geophys, Geosyst* 22, e2021GC010056. doi:10.1029/2021GC010056

Elderfield, H., and Schult, A. (1996). Mid-ocean ridge hydrothermal fluxes and the chemical composition of the ocean. *Ann. Rev. Earth Planet Sci.* 24, 191–224. doi:10.1146/annurev.earth.24.1.191

Elliott, B. A. (2018). Petrogenesis of heavy rare earth element enriched rhyolite: Source and magmatic evolution of the round top laccolith, trans-pecos, Texas. *Trans-Pecos, Tex. Miner.* 8, 423. doi:10.3390/min8100423

Emsenhuber, A., Asphaug, E., Cambioni, S., Gabriel, T. S., and Schwartz, S. R. (2017). Collision chains among the terrestrial planets. II. an asymmetry between Earth and venusErastova V, Degiacomi MT, Fraser DG, Greenwell HCMineral surface chemistry control for origin of prebiotic peptides. *Planet. Sci. journalNature Commun* 28 (5), 1992033. doi:10.1038/s41467-017-02248-y

Erastova, V., Degiacomi, M. T., Fraser, D. G., and Greenwell, H. C. (2017). Mineral surface chemistry control for origin of prebiotic peptides. *Nat. Commun.* 8, 2033. doi:10.1038/s41467-017-02248-y

Fairén, A. G., Davila, A. F., Gago-Duport, L., Amils, R., and McKay, C. P. (2009). Stability against freezing of aqueous solutions on early Mars. *Nature* 459, 401–404. doi:10.1038/nature07978

Fairén, A. G., Davila, A. F., Lim, D., Bramall, N., Bonaccorsi, R., Zavaleta, J., et al. (2010). Astrobiology through the ages of mars: The study of terrestrial analogues to understand the habitability of mars. *Astrobiology* 10, 821–843. doi:10.1089/ast.2009.0440

Fairén, A. G., Fernández-Remolar, D., Dohm, J. M., Baker, V. R., and Amils, R. (2004). Inhibition of carbonate synthesis in acidic oceans on early Mars. *Nature* 431, 423–426. doi:10.1038/nature02911

Fedo, C. (2000). Setting and origin for problematic rocks from the >3.7 Ga Isua greenstone belt, southern west Greenland: Earth's oldest coarse clastic sediments. *Precamb Res.* 101, 69–78. doi:10.1016/S0301-9268(99)00100-X

Fegley, B., and Schaefer, L. (2013). "Chemistry of the earth's earliest atmosphere," in *Treatise on geochemistry* (Elsevier), 13. eprint arXiv:1210.0270.

Ferris, J. P., Edelson, E. H., Mount, N. M., and Sullivan, A. E. (1979). The effect of clays on the oligomerization of HCN. *J. Mol. Evol.* 13, 317–330. doi:10.1007/bf01731372

Ferris, J. P., Hill, A. R., Liu, R., and Orgel, L. E. (1996). Synthesis of long prebiotic oligomers on mineral surfaces. *Nature* 381, 59–61. doi:10.1038/381059a0

Ferus, M., Rimmer, P., Cassone, G., Knızek, A., Civis, S., Sponer, J. E., et al. (2020). One-pot hydrogen cyanide-based prebiotic synthesis of canonical nucleobases and glycine initiated by high-velocity impacts on early Earth? *Astrobiology* 20, 1476–1488. doi:10.1089/ast.2020.2231

Fiore, M., and Strazewski, P. (2016). Prebiotic lipidic amphiphiles and condensing agents on the early Earth. *Life* 6, 17. doi:10.3390/life6020017

Fisher, A. T., and Wheat, C. G. (2010). Seamounts as conduits for massive fluid, heat, and solute fluxes on ridge flanks. *Oceanography* 23, 74–87.

Fishkis, M. (2007). Steps towards the formation of a protocell: The possible role of short peptides. *Orig. Life Evol. Biosph.* 37, 537–553. doi:10.1007/s11084-007-91114

Folco, L., and Cordier, C. (2015). Micrometeorites. *EMU Notes Mineralogy* 15, 253–297. doi:10.1180/EMU-notes.15.9

Forsythe, J. G., Yu, S-S., Mamajanov, I., Grover, M. A., Krishnamurthy, R., Fernández, F. M., et al. (2015). Ester-mediated amide bond formation driven by wet–dry cycles: A possible path to polypeptides on the prebiotic earth. *Angew. Chem.* 54, 9871–9875. doi:10.1002/anie.201503792

Foucher, F., Westall, F., Brandstatter, F., Demets, R., Parnell, J., Cockell, C. S., et al. (2010). Testing the survival of microfossils in artificial martian sedimentary meteorites during entry into earth's atmosphere: The STONE 6 experiment. *Icarus* 207, 616–630. doi:10.1016/j.icarus.2009.12.014

Franck, S. (1998). Evolution of the global mean heat flow over 4.6 Gyr. *Tectonophysics* 291, 9–18. doi:10.1016/s0040-1951(98)00027-4

Frenkel-Pinter, M., Samanta, M., Ashkenasy, G., and Leman, L. J. (2020). Prebiotic peptides: Molecular hubs in the origin of life. *Chem. Rev.* 120, 4707–4765. doi:10.1021/acs.chemrev.9b00664

Furukawa, Y., Chikaraishi, Y., Ohkouchi, N., Ogawa, N. O., Glavin, D. P., Dworkin, J. P., et al. (2019). Extraterrestrial ribose and other sugars in primitive meteorites. *PNAS* 116, 24440–24445. doi:10.1073/pnas.1907169116

Furukawa, Y., Sekine, T., Oba, M., Kakegawa, T., and Nakazawa, H. (2009). Biomolecule formation by oceanic impacts on early Earth. *Nat. Geosci.* 2, 62–66. doi:10.1038/ngeo383

Gagnevin, D., Ethien, R., Bonin, B., Moine, B., Féraud, G., Gerbe, M. C., et al. (2003). Open-system processes in the Genesis of silica-oversaturated alkaline rocks of the rallier-du-baty peninsula, kerguelen archipelago (Indian ocean). *J. Volcanol. Geotherm. Res.* 123, 267–300. doi:10.1016/S0377-0273(02)00509-7

Gale, A., Dalton, C. A., Langmuir, C. H., Su, Y., and Schilling, J. G. (2013). The mean composition of ocean ridge basalts. *Geochem Geophys Geosyst* 14, 489–518. doi:10.1029/2012gc004334

Gallo, T. A., and Klein, L. C. (1988). Effect of dehydration on the viscosity of sol-gel processed silica. *J. Non-Cryst Solids* 100, 429–434. doi:10.1016/0022-3093(88)90059-2

Gaucher, E. A., Thomson, J. M., Burgan, M. F., and Benner, S. A. (2003). Inferring the palaeoenviron-ment of ancient bacteria on the basis of resurrected proteins. *Nature* 425, 285–288. doi:10.1038/nature01977

German, C. R., Casciotti, K. A., Dutay, J. C., Heimbürger, L. E., Jenkins, W. J., Measures, C. I., et al. (2016). Hydrothermal impacts on trace element and isotope ocean biogeochemistry. *Phil. Trans. R. Soc. A.* 374, 20160035. doi:10.1098/rsta.2016.0035

Gibard, C., Bhowmik, S., Karki, M., Kim, E-K., and Krishnamurthy, R. (2018). Phosphorylation, oligomerization and self-assembly in water under potential prebiotic conditions. *Nat. Chem.* 10, 212–217. doi:10.1038/nchem.2878

Glavin, D. P., Elsila, J. E., Mclain, H. L., Aponte, J. C., Parker, E. T., Dworkin, J. P., et al. (2021). Extraterrestrial amino acids and L-enantiomeric excesses in the CM 2 carbonaceous chondrites Aguas Zarcas and Murchison. *Planet Sci.* 56, 148–173. doi:10.1111/maps.13451

Gleiser, M. (2022). Biological homochirality and the search for extraterrestrial biosignatures. *Orig. Life Evol. Biosph.* 52, 93–104. doi:10.1007/s11084-022-09623-w

Gourier, D., Binet, L., Calligaro, T., Capelli, S., Vezin, H., Bréhéret, J., et al. (2019). Extraterrestrial organic matter preserved in 3.33 Ga sediments from Barberton, South Africa. *Geochim. Cosmochim. Acta* 258, 207–225.

Greer, J., Caro, G., Cates, N. L., Tropper, P., Bleeker, W., Kelly, N. M., et al. (2020). Widespread poly-metamorphosed Archean granitoid gneisses and supracrustal enclaves of the Southern Inukjuak domain, Québec (Canada). *Lithos* 364, 105520. doi:10.1016/j.lithos.2020.105520

Griffin, W. L., Belousova, E. A., O'Neill CosyO'Reilly, S. Y., Malkovets, V., Pearson, N. J., et al. (2014). The world turns over: Hadean–Archean crust–mantle evolution. *Lithos* 189, 2–15. doi:10.1016/j.lithos.2013.08.018

Guélin, M., and Cernicharo, J. (2022). Organic molecules in interstellar space: Latest advances. *Front. Astron. Space Sci.* 9, 787567. doi:10.3389/fspas.2022.787567

Halliday, A., Rehkämper, M., Lee, D. C., and Yi, W. (1996). Early evolution of the earth and moon: New constraints from Hf-W isotope geochemistry. *Earth Planet Sci. Lett.* 142, 75–89. doi:10.1016/0012-821x(96)00096-9

Hansell, D. A., Carlson, C. A., Repeta, D. J., and Reiner, S. (2009). Dissolved organic matter in the ocean: A controversy stimulates new insights. *Oceanography* 22, 202–211. doi:10.5670/oceanog.2009.109

Hansma, H. G. (2022). DNA and the origins of life in micaceous clay. *Biophys. J.* 121, 301–307. doi:10.3390/life12020301

Harris, A. C., White, N. C., McPhie, J., Bull, S. W., Line, M. A., Skrzeczynski, R., et al. (2009). Early archean hot springs above epithermal veins, North Pole, western Australia: New insights from fluid inclusion microanalysis. *Econ. Geol.* 104, 793–814. doi:10.2113/gsecongeo.104.6.793

Harrison, T. M., Blichert-Toft, J., Müller, W., Albarede, F., Holden, P., and Mojzsis, S. J. (2005). Heterogeneous hadean hafnium: Evidence of continental crust at 4.4 to 4.5 Ga. *Science* 310, 1947–1950. doi:10.1126/science.1117926

Harrison, T. M., Schmitt, A. K., McCulloch, M. T., and Lovera, O. M. (2008). Early (> 4.5 Ga) formation of terrestrial crust: Lu-Hf, δ18O, and Ti thermometry results for hadean zircons. *Earth Planet Sci. Lett.* 268, 476–486. doi:10.1016/j.epsl.2008.02.011

Hartman, H. (1975). Speculations on the origin and evolution of metabolism. *J. Molec Evol.* 4, 359–370. doi:10.1007/bf01732537

Hassenkam, T., Andersson, M., Dalby, K., Mackenzie, D., and Rosing, M. (2017). Elements of Eoarchean life trapped in mineral inclusions. *Nature* 548, 78–81. doi:10.1038/nature23261

Hawkesworth, C. J., and Cawood, P. A. (2020). The evolution of the continental crust and the onset of plate tectonics. *Front. Earth Sci.* 8, 326. doi:10.3389/feart.2020.00326

Haymon, R., White, S. M., Baker, E. T., Anderson, P. G., Macdonald, K. C., and Resing, J. A. (2008). High-resolution surveys along the hot spot–affected galapagos spreading center: 3. Black smoker discoveries and the implications for geological controls on hydrothermal activity. *Geochem Geophys Geosyst* 9, Q12006. doi:10.1029/2008GC002114

He, R., Hu, B., Zhong, H., Jin, F., Fan, J., Hu, Y. H., et al. (2019). Reduction of CO2 with H2S in a simulated deep-sea hydrothermal vent system. *Chem. Commun.* 55, 1056–1059. doi:10.1039/c8cc08075e

Herschy, B., Whicher, A., Camprubi, E., Watson, C., Dartnell, L., Ward, J., et al. (2014). An origin-of-life reactor to simulate alkaline hydrothermal vents. *J. Mol. Evol.* 79, 213–227. doi:10.1007/s00239-014-9658-4

Hertkorn, N., Benner, R., Frommberger, M., Schmitt-Kopplin, P., Witt, M., Kaiser, K., et al. (2006). Characterization of a major refractory component of marine dissolved organic matter. *Geochim Cosmochim Acta* 70, 2990–3010. doi:10.1016/j.gca.2006.03.021

Hertkorn, N., Harir, M., and Schmitt-Kopplin, P. (2015). Nontarget analysis of Murchison soluble organic matter by high-field NMR spectroscopy and FTICR mass spectrometry. *Magn. Res. Chem.* 53, 754–768. doi:10.1002/mrc.4249

Hickman-Lewis, K., Cavalazzi, B., Foucher, F., and Westall, F. (2018). Most ancient evidence for life in the Barberton greenstone belt: Microbial mats and biofabrics of the ~3.47 Ga Middle marker horizon. *Precamb Res.* 312, 45–67.

Hickman-Lewis, K., Cavalazzi, B., Sorieul, S., Gautret, P., Foucher, F., Whitehouse, M. J., et al. (2020a). Metallomics in deep time and the influence of ocean chemistry on the metabolic landscapes of Earth's earliest ecosystems. *Sci. Rep.* 10, 4965. doi:10.1038/s41598-020-61774-w

Hickman-Lewis, K., Gourcerol, B., Westall, F., Manzini, D., and Cavalazzi, B. (2020b). Reconstructing Palaeoarchaean microbial biomes flourishing in the presence of emergent landmasses using trace and rare Earth element systematics. *Precamb Res.* 342, 105689. doi:10.1016/j.precamres.2020.105689

Hickman-Lewis, K., and Westall, F. (2021). A southern African perspective on the co-evolution of early life and environments. *South Afr. J. Geol.* 124, 225–252. doi:10.25131/sajg.124.0016

Higgs, P. G., and Lehman, N. (2015). The RNA world: Molecular cooperation at the origins of life. *Nat. Rev. Genet.* 16, 7–17. doi:10.1038/nrg3841

Hodgkinson, M. R. S. (2015). *The geological controls on the Von Damm vent field*. Southampton: University of Southampton. Doctoral dissertation.

Hofmann, A., and Harris, C. (2008). Silica alteration zones in the Barberton greenstone belt: A window into subseafloor processes 3.5–3.3 Ga ago. *Chel Geol.* 257, 221–239. doi:10.1016/j.chemgeo.2008.09.015

Hofmann, H. J., Grey, K., Hickman, A. H., and Thorpe, R. (1999). Origin of 3.45 Ga coniform stromatolites in warrawoona group, western Australia. *Geol. Soc. Am. Bull.* 111, 1256–1262. doi:10.1130/0016-7606(1999)111<1256:oogcsi>2.3.co;2

Holden, D. T., Morato, N. M., and Cooks, R. G. (2022). Aqueous microdroplets enable abiotic synthesis and chain extension of unique peptide isomers from free amino acids. *PNAS* 119, e2212642119. doi:10.1073/pnas.2212642119

Holm, N. G., and Andersson, E. M. (2005). Hydrothermal simulation experiments as a tool for studies of the origin of life on Earth and other terrestrial planets: A review. *Astrobiology* 5, 444–460. doi:10.1089/ast.2005.5.444

Holm, N. G., and Charlou, J-L. (2001). Initial indications of abiotic formation of hydrocarbons in the Rainbow ultramafic hydrothermal system, mid-Atlantic ridge. *Earth Planet Sci. Lett.* 191, 1–8. doi:10.1016/s0012-821x(01)00397-1

Iler, R. K. (1979). *The chemistry of silica*. Wiley, 866.

Imai, E. I., Honda, H., Hatori, K., Brack, A., and Matsuno, K. (1999). Elongation of oligopeptides in a simulated submarine hydrothermal system. *Science* 283, 831–833. doi:10.1126/science.283.5403.831

Iqubal, M. A., Sharma, R., and Kamaluddin, K. (2016). Surface interaction of ribonucleic acid constituents with spinel ferrite nanoparticles: A prebiotic chemistry experiment. *RSC Adv.* 6, 68574–68583. doi:10.1039/c6ra12247g

Izgu, E. C., Björkbom, A., Kamat, N. P., Lelyveld, V. S., Zhang, W., Jia, T. Z., et al. (2016). N-Carboxyanhydride-Mediated fatty acylation of amino acids and peptides for functionalization of protocell membranes. *J. Am. Chem. Soc.* 138, 16669–16676. doi:10.1021/jacs.6b08801

Jacobsen, S. B. (2005). The Hf-W isotopic system and the origin of the earth and moon. *Ann. Rev. Earth Planet Sci.* 33, 531–570. doi:10.1146/annurev.earth.33.092203.122614

Jerome, C. A., Kim, H. J., Mojzsis, S. J., Benner, S. A., and Biondi, E. (2022). Catalytic synthesis of polyribonucleic acid on prebiotic rock glasses. *Astrobiology* 22, 629–636. doi:10.1089/ast.2022.0027

Johansen, A., Ronnet, T., Bizzarro, M., Schiller, M., Lambrechts, M., Nordlund, A., et al. (2021). A pebble accretion model for the formation of the terrestrial planets in the Solar System. *Sci. Adv.* 7, eabc0444. doi:10.1126/sciadv.abc0444

Jordan, S. F., Rammu, H., Zheludev, I. N., Hartley, A. M., Maréchal, A., and Lane, N. (2019). Promotion of protocell self-assembly from mixed amphiphiles at the origin of life. *Nat. Ecol. Evol.* 3, 1705–1714. doi:10.1038/s41559-019-1015-y

Karlen, D., Price, R., Pichler, T., and Garey, J. (2010). Changes in benthic macrofauna associated with a shallow-water hydrothermal vent gradient in Papua New Guinea. *Pac. Sci.* 64, 391–404. doi:10.2984/64.3.391

Kitadai, N., Oonishi, H., Umemoto, K., Usui, T., Fukushi, K., and Nakashima, S. (2017). Glycine polymerization on oxide minerals. *Orig. Life Evol. Biosph.* 47, 123–143. doi:10.1007/s11084-016-9516-z

Kleine, T., Münker, C., Mezger, K., and Palme, H. (2002). Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf-W chronometry. *Nature* 418, 952–955. doi:10.1038/nature00982

Kleine, T., Palme, H., Mezger, K., and Halliday, A. N. (2005). Hf-W Chronometry of lunar metals and the age and early differentiation of the Moon. *Science* 310, 1671–1674. doi:10.1126/science.1118842

Kleine, T., Touboul, M., Bourdon, B., Nimmo, F., Mezger, K., Palme, H., et al. (2009). Hf–W chronology of the accretion and early evolution of asteroids and terrestrial planets. *Geochim. Cosmochim. Acta* 73, 5150–5188. doi:10.1016/j.gca.2008.11.047

Kloprogge, J. T., and Hartman, H. (2022). Clays and the origin of life: The experiments. *Life* 12, 259–313. doi:10.3390/life12020259

Knauth, L. P. (2005). Temperature and salinity history of the Precambrian ocean: Implications for the course of microbial evolution. *Palaeogeo Palaeoclimatol. Palaeoecol.* 219, 53–69. doi:10.1016/j.palaeo.2004.10.014

Konn, C., Charlou, J. L., Holm, N. G., and Mousis, O. (2015). The production of methane, hydrogen, and organic compounds in ultramafic-hosted hydrothermal vents of the Mid-Atlantic Ridge. *Astrobiology* 15, 381–399. doi:10.1089/ast.2014.1198

Korenaga, J. (2018). Crustal evolution and mantle dynamics through Earth history. *Phil Trans. Roy. Soc. A* 376 (2132), 20170408. doi:10.1098/rsta.2017.0408

Lagaly, G. (1984). Clay-organic interactions. *Phil. Trans. R. Soc. Lond. A* 311, 315–332.

Lambert, J-F., Jaber, M., Georgelin, T., and Stievano, L. (2013). A comparative study of the catalysis of peptide bond formation by oxide surfaces. *Phys. Chem. Chem. Phys.* 15, 13371–13380. doi:10.1039/C3CP51282G

Lammer, H., Zerkle, A. L., Gebauer, S., Tosi, N., Noack, L., Scherf, M., et al. (2018). Origin and evolution of the atmospheres of early venus, earth and mars. *Astron Astrophys. Rev.* 26, 2–72. doi:10.1007/s00159-018-0108-y

Lan, Z., Kamo, S. L., Roberts, N. M., Sano, Y., and Li, X. H. (2022). A Neoarchean (ca. 2500 ma) age for jaspilite-carbonate bif hosting purported micro-fossils from the Eoarchean (> 3750 ma) Nuvvuagittuq supracrustal belt (Québec, Canada). *Precamb Res.* 377, 106728. doi:10.1016/j.precamres.2022.106728

Lanier, W. P., and Lowe, D. R. (1982). Sedimentology of the Middle marker (3.4 Ga), onverwacht group, transvaal, South Africa. *Precamb Res.* 18, 237–260. doi:10.1016/0301-9268(82)90012-2

Ledevin, M., Arndt, N., Chauvel, C., Jaillard, E., and Simionovici, A. (2019). The sedimentary origin of black and white banded cherts of the Buck Reef, Barberton, South Africa. *Geosciences* 9, 424. doi:10.3390/geosciences9100424

Lee, C., Weber, J. M., Rodriguez, L. E., Sheppard, R. Y., Barge, L. M., Berger, E. L., et al. (2022). Chirality in organic and mineral systems: A review of reactivity and alteration processes relevant to prebiotic chemistry and life detection missions. *Symmetry* 14, 460–509. doi:10.3390/sym14030460

Lough, A. J., Connelly, D. P., Homoky, W. B., Hawkes, J. A., Chavagnac, V., Castillo, A., et al. (2019). Diffuse hydrothermal venting: A hidden source of iron to the oceans. *Front. Mar. Sci.* 6, 329. doi:10.3389/fmars.2019.00329

Lourenço, D. L., Rozel, A. B., Gerya, T., and Tackley, P. J. (2018). Efficient cooling of rocky planets by intrusive magmatism. *Nat. Geosc* 11, 322–327. doi:10.1038/s41561-018-0094-8

Lowe, D. R. (1999). "Petrology and sedimentology of cherts and related silicified sedimentary rocks in the Swaziland Supergroup," in *Geologic Evolution of the Barberton Greenstone Belt, South Africa: Geol Soc America Special Paper* 29, 83–114.

Lowe, D. R., Byerly, G. R., Kyte, F. T., Shukolyukov, A., Asaro, F., and Krull, A. (2003). Spherule beds 3.47–3.24 billion years old in the Barberton greenstone belt, South Africa: A record of large meteorite impacts and their influence on early crustal and biological evolution. *Astrobiology* 3, 7–48. doi:10.1089/153110703321632408

Lowe, D. R., and Byerly, G. (2003). Ironstone pods in the archean Barberton greenstone belt, South Africa: Earth's oldest seafloor hydrothermal vents reinterpreted as quaternary subaerial springs. *Geology* 31, 909. doi:10.1130/G19664.1

Lowe, D. R., Byerly, G. R., and Kyte, F. T. (2014). Recently discovered 3.42-3.23 Ga impact layers, Barberton belt, South Africa: 3.8 Ga detrital zircons, archean impact history, and tectonic implications. *Geology* 42, 747–750. doi:10.1130/g35743.1

Lowe, D. R., and Knauth, L. P. (1977). Sedimentology of the Onverwacht Group (3.4 billion years), Transvaal, South Africa, and its bearing on the characteristics and evolution of the early Earth. *J. Geol.* 85, 699–723. doi:10.1086/628358

Lowe, D. R., and Byerly, G. R. (2015). Geologic record of partial ocean evaporation triggered by giant asteroid impacts, 3.29-3.23 billion years ago. *Geology* 43, 535–538. doi:10.1130/g36665.1

Ludwig, K. A., Shen, C-C., Kelley, D. S., Cheng, H., and Edwards, R. L. (2011). U–Th systematics and 230Th ages of carbonate chimneys at the Lost City Hydrothermal Field. *Geochim. Cosmochim. Acta* 75, 1869–1888. doi:10.1016/j.gca.2011.01.008

Maher, K. A., and Stevenson, D. J. (1988). Impact frustration of the origin of life. *Nature* 331, 612–614. doi:10.1038/331612a0

Manhes, G., Allegre, C. J., Dupre, B., and Hamelin, B. (1979). Lead-lead systematics, the "age of the Earth" and the chemical evolution of our planet in a new representation space. *Earth Planet Sci. Lett.* 44, 91–104. doi:10.1016/0012-821x(79)90011-6

Marchi, S., Drabon, N., Schulz, T., Schaefer, L., Nesvorny, D., Bottke, W. F., et al. (2021). Delayed and variable late Archaean atmospheric oxidation due to high collision rates on Earth. *Nat. Geosci.* 14, 827–831. doi:10.1038/s41561-021-00835-9

Martin, W., Baross, J., Kelley, D., and Russell, M. J. (2008). Hydrothermal vents and the origin of life. *Nat. Rev. Microbio* 6, 805–814. doi:10.1038/nrmicro1991

Martra, G., Deiana, C., Sakhno, Y., Barberis, I., Fabbiani, M., Pazzi, M., et al. (2014). The formation and self-assembly of long prebiotic oligomers produced by the condensation of unactivated amino acids on oxide surfaces. *Angew. Chem.* 53, 4759–4762. doi:10.1002/ange.201311089

Matrajt, G., Pizzarello, S., Taylor, S., and Brownlee, D. (2004). Concentration and variability of the AIB amino acid in polar micrometeorites: Implications for the exogenous delivery of amino acids to the primitive Earth. *Meteorit. Planet Sci.* 39, 1849–1858. doi:10.1111/j.1945-5100.2004.tb00080.x

Maurette, M., and Brack, A. (2006). Cometary petroleum in hadean time? Meteoritics planet. *Science* 41, 5247.

Maurette, M. (1998). Carbonaceous micrometeorites and the origin of life. *Orig. Life Evol. Biosph.* 28, 385–412. doi:10.1023/a:1006589819844

Maurette, M. (2006). *Micrometeorites and the mysteries of our origins*. Heidelberg, Germany: Springer Berlin.

Maurice, M., Tosi, N., Schwinger, S., Breuer, D., and Kleine, T. (2020). A long-lived magma ocean on a young Moon. *Sci. Adv.* 6, eaba8949. doi:10.1126/sciadv.aba8949

Maury, C. P. J. (2018). Amyloid and the origin of life: Self-replicating catalytic amyloids as prebiotic informational and protometabolic entities. *Cell Mol. Life Sci.* 75, 1499–1507. doi:10.1007/s00018-018-2797-9

McCollom, T. M. (2000). Geochemical constraints on primary productivity in submarine hydrothermal vent plumes. *Deep-Sea Res. Part I Oceanogr. Res. Pap.* 47, 85–101. doi:10.1016/s0967-0637(99)00048-5

McCollom, T. M., and Shock, E. L. (1997). Geochemical constraints on chemolithoautotrophic metabolism by microorganisms in seafloor hydrothermal systems. *Geochim. Cosmochim. Acta* 61, 4375–4391. doi:10.1016/s0016-7037(97)00241-x

McDermott, J. M., Seewald, J. S., German, C. R., and Sylva, S. P. (2015). Pathways for abiotic organic synthesis at submarine hydrothermal fields. *PNAS* 112, 7668–7672. doi:10.1073/pnas.1506295112

McMahon, S., Bosak, T., Grotzinger, J. P., Milliken, R. E., Summons, R. E., Daye, M., et al. (2018). A field guide to finding fossils on Mars. *J. Geophys Res. Planets* 123, 1012–1040. doi:10.1029/2017je005478

McMahon, S. (2019). Earth's earliest and deepest purported fossils may be iron-mineralized chemical gardens. *Proc. R. Soc. B Biol. Sci.* 286, 20192410. doi:10.1098/rspb.2019.2410

Melosh, H., and Vickery, A. (1989). Impact erosion of the primordial atmosphere of Mars. *Nature* 338, 487–489. doi:10.1038/338487a0

Micela, G. (2002). "Evolution of stellar coronal activity on the main sequence," in *The evolving sun and its influence on planetary environments* 269, 107.

Micela, G., Sciortino, S., Serio, S., Vaiana, G. S., Bookbinder, J., Golub, L., et al. (1985). Einstein X-ray survey of the pleiades-the dependence of X-ray emission on stellar age. *Astrophysical J.* 292, 172–180.

Micela, G., Sciortino, S., Vaiana, G. S., Schmitt, J. H. M. M., Stern, R. A., Harnden Jr, F. R., et al. (1988). The Einstein observatory survey of stars in the Hyades cluster region. *Astrophysical J.* 325, 798–819.

Miller, S. L. (1953). A production of amino acids under possible primitive earth conditions. *Science* 117, 528–529. doi:10.1126/science.117.3046.528

Miller, S. L., and Lazcano, A. (1995). The origin of life--did it occur at high temperatures? *J. Mol. Evol.* 41, 689–692. doi:10.1007/bf00173146

Milshteyn, D., Damer, B., Havig, J., and Deamer, D. (2018). Amphiphilic compounds assemble into membranous vesicles in hydrothermal hot spring water but not in seawater. *Life* 8, 11. doi:10.3390/life8020011

Mojzsis, S. J., Brasser, R., Kelly, N. M., Abramov, O., and Werner, S. C. (2019). Onset of giant planet migration before 4480 million years ago. *Astrophys. J.* 881, 44. doi:10.3847/1538-4357/ab2c03

Mojzsis, S. J., Harrison, T. M., and Pidgeon, R. T. (2001). Oxygen-isotope evidence from ancient zircons for liquid water at the Earth's surface 4, 300 myr ago. *Nature* 409, 178–181. doi:10.1038/35051557

Monnard, P-A., and Walde, P. (2015). Current ideas about prebiological compartmentalization. *Life* 5, 1239–1263. doi:10.3390/life5021239

Morbidelli, A., Lunine, J. I., O'Brien, D. P., Raymond, S. N., and Walsh, K. J. (2012). Building terrestrial planets. *Ann. Rev. Earth Planet Sci.* 40, 251–275. doi:10.1146/annurev-earth-042711-105319

Moser, E., Arcuri, G. A., Reinhard, D. A., White, L. F., Darling, J. R., Barker, I. R., et al. (2019). Decline of giant impacts on Mars by 4.48 billion years ago and an early opportunity for habitability. *Nat. Geosc* 12, 522–527. doi:10.1038/s41561-019-0380-0

Moyen, J. F., and Martin, H. (2013). Forty years of TTG research. *Lithos* 148, 312–336. doi:10.1016/j.lithos.2012.06.010

Muehlenbachs, K., and Clayton, R. N. (1976). Oxygen isotope composition of the oceanic crust and its bearing on seawater. *J. Geophys Res.* 81, 4365–4369. doi:10.1029/jb081i023p04365

Mukhopadhyay, S. (2012). Early differentiation and volatile accretion recorded in deep-mantle neon and xenon. *Nature* 486, 101–104. doi:10.1038/nature11141

Müller, U. F., Elsila, J., Trail, D., DasGupta, S., Giese, C-C., Walton, C. R., et al. (2022). Frontiers in prebiotic chemistry and early Earth environments. *Orig. Life Evol. Biosph.* 52, 165–181. doi:10.1007/s11084-022-09622-x

Nam, I., Nam, H. G., and Zare, R. N. (2018). Abiotic synthesis of purine and pyrimidine ribonucleosides in aqueous microdroplets. *PNAS* 115, 36–40. doi:10.1073/pnas.1718559115

Nijman, W., and De Vries, S. T. (2004). "Early Archaean crustal collapse structures and sedimentary basin dynamics," in *The precambrian earth: Tempos and events. Developments in precambrian geology*. Editors P. G. Eriksson, W. Altermann, D. R. Nelson, W. U. Mueller, and O. Catuneanu (Amsterdam: Elsevier), 12, 139–155.

Nijman, W., Kloppenburg, A., and de Vries, S. T. (2017). Archaean basin margin geology and crustal evolution: An east Pilbara traverse. *J. Geol. Soc.* 174, 1090–1112. doi:10.1144/jgs2016-127

Noffke, N., Christian, D., Wacey, D., and Hazen, R. M. (2013). Microbially induced sedimentary structures recording an ancient ecosystem in the ca. 3.48 billion-year-old Dresser Formation, Pilbara, Western Australia. *Astrobiology* 13, 1103–1124. doi:10.1089/ast.2013.1030

Nutman, A. P., Bennett, V. C., Friend, C. R., Van Kranendonk, M. J., and Chivas, A. R. (2016). Rapid emergence of life shown by discovery of 3, 700-million-year-old microbial structures. *Nature* 537, 535–538. doi:10.1038/nature19355

O'Neil, J., Carlson, R. W., Paquette, J-L., and Francis, D. (2012). formation age and metamorphic history of the Nuvvuagittuq greenstone belt. *Precambrian Res.* 220-221, 23–44. doi:10.1016/j.precamres.2012.07.009

O'Neill, C., Jellinek, A. M., and Lenardic, A. (2007). Conditions for the onset of plate tectonics on terrestrial planets and moons. *Earth Planet Sci. Lett.* 261, 20–32. doi:10.1016/j.epsl.2007.05.038

Ogata, Y., Imai, E. I., Honda, H., Hatori, H. K., and Matsuno, K. (2000). Hydrothermal circulation of seawater through hot vents and contribution of interface chemistry to prebiotic synthesis. *Orig. Life Evol. Biosph.* 30, 527–537. doi:10.1023/a:1026543825011

O'Neil, J., Francis, D., and Carlson, R. W. (2011). Implications of the Nuvvuagittuq greenstone belt for the formation of Earth's early crust. *J. Pet.* 52, 985–1009. doi:10.1093/petrology/egr014

Orange, F., Westall, F., Disnar, J. R., Prieur, D., Bienvenu, N., Le Romancer, M., et al. (2009). Experimental silicification of the extremophilic ArchaeaPyrococcus abyssiandMethanocaldococcus jannaschii: Applications in the search for evidence of life in early earth and extraterrestrial rocks. *Geobiology* 7, 403–418. doi:10.1111/j.1472-4669.2009.00212.x

Orcel, G., Phalippou, J., and Hench, L. L. (1986). Structural changes of silica xerogels during low temperature dehydration. *J. Non-Crystalline Solids* 88, 114–130. doi:10.1016/s0022-3093(86)80092-8

Orgel, L. E. (1998). The Origin of Life-How long did it take? *Orig. Life Evol. Biosph.* 28, 91–96. doi:10.1023/a:1006561308498

Osinski, G. R., Cockell, C. S., Pontefract, A., and SapersHM (2020). The role of meteorite impacts in the origin of life. *Astrobiology* 20, 1121–1149. doi:10.1089/ast.2019.2203

Oberg, K. I. (2016). Photochemistry and astrochemistry: Photochemical pathways to interstellar complex organic molecules. *Chem. Rev.* 116 (17), 9631–9663. doi:10.1021/acs.chemrev.5b00694

Ozima, M., and Podosek, F. A. (1999). Formation age of Earth from 129I/127I and 244Pu/238U systematics and the missing Xe. *J. Geophysl Res.* 104 (25), 25493–25499. doi:10.1029/1999jb900257

Paksi, D. (2014). The concept of boundary conditions. *Polanyiana* 1-2, 5–20.

Papineau, D., She, Z., Dodd, M. S., Iacoviello, F., Slack, J. F., Hauri, E., et al. (2022). Metabolically diverse primordial microbial communities in Earth's oldest seafloor-hydrothermal jasper. *Scie Adv.* 8, eabm2296. doi:10.1126/sciadv.abm2296

Parker, E. T., Cleaves, H. J., Dworkin, J. P., Glavin, D. P., Callahan, M., Aubrey, A., et al. (2011). Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment. *PNAS* 108, 5526–5531. doi:10.1073/pnas.1019191108

Parker, E. T., Zhou, M., Burton, A. S., Glavin, D. P., Dworkin, J. P., Krishnamurthy, R., et al. (2014). A plausible simultaneous synthesis of amino acids and simple peptides on the primordial Earth. *Angew. Chem.* 53, 8132–8136. doi:10.1002/anie.201403683

Pascal, R., Pross, A., and Sutherland, J. D. (2013). Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biol.* 3, 130156. doi:10.1098/rsob.130156

Patel, B. H., Percivalle, C., Ritson, D. J., Duffy, C. D., and Sutherland, J. D. (2015). Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem.* 7, 301–307. doi:10.1038/nchem.2202

Perchuk, A. L., Gerya, T. V., Zakharov, V. S., and Griffin, W. L. (2020). Building cratonic keels in Precambrian plate tectonics. *Nature* 586, 395–401. doi:10.1038/s41586-020-2806-7

Pizzarello, S., and Huang, Y. (2005). The deuterium enrichment of individual amino acids in carbonaceous meteorites: A case for the presolar distribution of biomolecule precursors. *Geochim. Cosmochim. Acta* 69, 599–605. doi:10.1016/j.gca.2004.07.031

Pizzarello, S., Schrader, D. L., Monroe, A. A., and Lauretta, D. S. (2012). Large enantiomeric excesses in primitive meteorites and the diverse effects of water in cosmochemical evolution. *PNAS* 109, 11949–11954. doi:10.1073/pnas.1204865109

Pizzarello, S., and Shock, E. (2017). Carbonaceous chondrite meteorites: The chronicle of a potential evolutionary path between stars and life. *Orig. Life Evol. Biosph.* 47, 249–260. doi:10.1007/s11084-016-9530-1

Pizzarello, S., and Shock, E. (2010). The organic composition of carbonaceous meteorites: The evolutionary story ahead of biochemistry. *Cold Spring Harb. Persp Biol.* 2, a002105. doi:10.1101/cshperspect.a002105

Pizzarello, S. (2007). The chemistry that preceded life's origin: A study guide from meteorites. *Chem. Biodivers.* 4, 680–693. doi:10.1002/cbdv.200790058

Pizzarello, S., Zolensky, M., and Turk, K. A. (2003). Nonracemic isovaline in the Murchison meteorite: Chiral distribution and mineral association. *Geochim. Cosmochim. Acta* 67, 1589–1595. doi:10.1016/s0016-7037(02)01283-8

Polanyi, M. (1997). "Life transcends physics and chemistry," in *Michael Polanyi: Society, economics, philosophy. Selected papers*. Editor R. T. Allen (New Brunswick, London: Transaction Publishers), 283–297.

Polat, A., Hofmann, A. W., Münker, C., Regelous, M., and Appel, P. W. U. (2003). Contrasting geochemical patterns in the 3.7–3.8 Ga pillow basalt cores and rims, Isua greenstone belt, southwest Greenland: Implications for postmagmatic alteration processes. *Geochim. Cosmochim. Acta* 67, 441–457. doi:10.1016/s0016-7037(02)01094-3

Pollack, G. H. (2001). *Cells, gels and the engines of life*. Seattle, Washington, USA: Ebner and Sons Publishers, 305.

Raymond, S. N. (2021). A terrestrial convergence. *Nat. Astron.* 5, 875–876. doi:10.1038/s41550-021-01488-9

Ricardo, A., Carrigan, M. A., Olcott, A. N., and Benner, S. A. (2004). Borate minerals stabilize ribose. *Science* 303, 196. doi:10.1126/science.1092464

Rimmer, P. B., and Shorttle, O. (2019). Origin of life's building blocks in Carbon and Nitrogen rich surface hydrothermal vents. *Life* 9, 12. doi:10.3390/life9010012

Robins, B., Sandstå, N. R., Furnes, H., and de Wit, M. (2010). Flow banding in basaltic pillow lavas from the early archean hooggenoeg formation, Barberton

greenstone belt, South Africa. *Bull. Volcanol.* 72, 579–592. doi:10.1007/s00445-009-0341-3

Rogers, A. D., Brierley, A., Croot, P., Cunha, M. R., Danovaro, R., Devey, C., et al. (2015). Delving deeper: Critical challenges for 21st century deep-sea research. in *position paper 22 of the European marine board*. Editors K. E. Larkin, K. Donaldson, and N. McDonough Ostend, Belgium 22, 224 ISBN 978-94-920431-1-5.

Rojas, J., Duprat, J., Engrand, C., Dartois, E., Delauche, L., Godard, M., et al. (2021). The micrometeorite flux at Dome C (Antarctica), monitoring the accretion of extraterrestrial dust on Earth. *Earth Planet Sci. Lett.* 560, 116794–116805. doi:10.1016/j.epsl.2021.116794

Root-Bernstein, R., and Brown, A. W. (2022). Novel apparatuses for incorporating natural selection processes into origins-of-life experiments to produce adaptively evolving chemical ecosystems. *Life* 12, 1508–1528. doi:10.3390/life12101508

Rosing, M. T. (1999). ^{13} C-depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from west Greenland. *Science* 283, 674–676. doi:10.1126/science.283.5402.674

Rossel, P. E., Stubbins, A., Rebling, T., Koschinsky, A., Hawkes, J. A., and Dittmar, T. (2017). Thermally altered marine dissolved organic matter in hydrothermal fluids. *Org. Geochem.* 110, 73–86. doi:10.1016/j.orggeochem.2017.05.003

Rout, S. K., Rhyner, D., Riek, R., and Greenwald, J. (2022). Prebiotically plausible autocatalytic peptide amyloids. *Chem. Eur. J.* 28, e202103841. doi:10.1002/chem.202103841

Rozel, A., Golabek, G. J., Jain, C., Tackley, P. J., and Gerya, T. (2017). Continental crust formation on early Earth controlled by intrusive magmatism. *Nature* 545, 332–335. doi:10.1038/nature22042

Russell, M. J., Daniel, R. M., and Hall, A. J. (1993). On the emergence of life via catalytic iron-sulphide membranes. *Terra nova.* 5, 343–347. doi:10.1111/j.1365-3121.1993.tb00267.x

Russell, M. J., and Hall, A. J. (1997). The emergence of life from iron monosulphide bubbles at a submarine hydrothermalredox and pH front. *J. Geol. Soc. Lond* 154, 377–402. doi:10.1144/gsjgs.154.3.0377

Russell, M. J. (2007). The alkaline solution to the emergence of life: Energy, entropy and early evolution. *Acta Biotheor.* 55, 133–179. doi:10.1007/s10441-007-9018-5

Saha, A., Yi, R., Fahrenbach, A. C., Wang, A., and Jia, T. Z. (2022). A physicochemical consideration of prebiotic microenvironments for self-assembly and prebiotic chemistry. *Life* 12, 1595. doi:10.3390/life12101595

Sasselov, D. D., Grotzinger, J. P., and Sutherland, J. D. (2020). The origin of life as a planetary phenomenon. *Sci. Adv.* 6, eaax3419. doi:10.1126/sciadv.aax3419

Schaefer, L., and Fegley, B. (2010). Chemistry of atmospheres formed during accretion of the Earth and other terrestrial planets. *Icarus* 208, 438–448. doi:10.1016/j.icarus.2010.01.026

Scheirer, D. S., Shank, T. M., and Fornari, D. J. (2006). Temperature variations at diffuse and focused flow hydrothermal vent sites along the northern East Pacific Rise. *Geochem Geophys Geosyst* 7, 3. doi:10.1029/2005gc001094

Schlesinger, G., and Miller, S. L. (1983). Prebiotic synthesis in atmospheres containing CH4, CO, and CO2. *J. Molec Evol.* 19, 376–382. doi:10.1007/bf02101642

Schmeling, H. (2006). A model of episodic melt extraction for plumes. *J. Geophys Res.* 111, B03202. doi:10.1029/2004JB003423

Schmitt-Kopplin, P., Gabelica, Z., Gougeon, R. D., Fekete, A., Kanawati, B., Harir, M., et al. (2010). High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. *PNAS* 107, 2763–2768. doi:10.1073/pnas.0912157107

Schneider, J. (1977). A model for a non-chemical form of life: Crystalline physiology. *Orig. Life Evol. Biosphere* 8, 33–38. doi:10.1007/bf00930937

Schreiber, U., Locker-Grütjen, O., and Mayer, C. (2012). Hypothesis: Origin of life in deep-reaching tectonic faults. *Orig. Life Evol. Biosph.* 42, 47–54. doi:10.1007/s11084-012-9267-4

Schrum, J. P., Zhu, T. F., and Szostak, J. W. (2010). The origins of cellular life. *Cold Spring Harb. Perspect. Biol.* 2, a002212. doi:10.1101/cshperspect.a002212

Schubert, G., Turcotte, D. L., and Olson, P. (2001). *Mantle convection in the earth and planets 2 volume set*. New York: Cambridge Univ. Press.

Seewald, J. S., Zolotov, M. Y., and McCollom, T. (2006). Experimental investigation of single carbon compounds under hydrothermal conditions. *Geochim. Cosmochim. Acta* 70, 446–460. doi:10.1016/j.gca.2005.09.002

Sheilds, G. A. (2007). "The marine carbonate and chert isotope records and they implications for tetctonics, life and climate on early Earth," in *Earth's oldest rocks. Dev. In precambrian geology*. Editors M. J. van Kranendonk, R. H. Smithies, and V. C. Bennet (Elsevier), 971–983.

Shock, E. L., and Schulte, M. D. (1990). Summary and implications of reported amino acid concentrations in the Murchison meteorite. *Geochim. Cosmochim. Acta* 54, 3159–3173. doi:10.1016/0016-7037(90)90131-4

Shock, E. L. (1993). Hydrothermal dehydration of aqueous organic compounds. *Geochim. Cosmochim. Acta* 57, 3341–3349. doi:10.1016/0016-7037(93)90542-5

Shock, E. L., McCollom, T., and Schulte, M. D. (1995). Geochemical constraints on chemolithoautotrophic reactions in hydrothermal systems. *Orig. Life Evol. Biosph.* 25, 141–159. doi:10.1007/BF01581579

Shock, E. L., and Schulte, M. D. (1998). Organic synthesis during fluid mixing in hydrothermal systems. *J Geophys Res. Planets* 103, 28513–28527. doi:10.1029/98je02142

Sim, S. J., Stegman, D. R., and Coltice, N. (2016). Influence of continental growth on mid-ocean ridge depth. *Geochem Geophys Geosyst* 17, 4425–4437. doi:10.1002/2016GC006629

Sizova, E., Gerya, T., Stüwe, K., and Brown, M. (2015). Generation of felsic crust in the archean: A geodynamic modeling perspective. *Precamb Res.* 271, 198–224. doi:10.1016/j.precamres.2015.10.005

Sleep, N. H. (2016). Asteroid bombardment and the core of Theia as possible sources for the Earth's late veneer component. *Geochem Geophys Geosyst* 17, 2623–2642. doi:10.1002/2016GC006305

Sleep, N. H. (2010). The hadean-archaean environment. *Cold Spring Harb. Perspect. Biol.* 2, a002527. doi:10.1101/cshperspect.a002527

Sleep, N. H., Zahnle, K., and Neuhoff, P. S. (2001). Initiation of clement surface conditions on the earliest Earth. *PNAS* 98, 3666–3672. doi:10.1073/pnas.071045698

Sleep, N. H., Zahnle, K. J., and Lupu, R. E. (2014). Terrestrial aftermath of the moon-forming impact. *Phil Trans. Roy. Soc. A* 372 (2024), 20130172. doi:10.1098/rsta.2013.0172

Smithies, R. H., Champion, D. C., and van Kranendonk, M. J. (2007). "The oldest well-preserved felsic volcanic rocks on earth: Geochemical clues to the early evolution of the Pilbara supergroup and implications for the growth of a paleoarchean continent," in *Earth's oldest rocksIn Precambrian Geology*. Editors M. J. van Kranendonk, R. H. Smithies, and V. C. Bennet (Elsevier), 339–367.

Sojo, V., Herschy, B., Whicher, A., Camprubi, E., and Lane, N. (2016). The origin of life in alkaline hydrothermal vents. *Astrobiology* 16, 181–197. doi:10.1089/ast.2015.1406

Staudacher, T., and Allegre, C. J. (1982). Terrestrial xenology. *Earth Planet Sci. Lett.* 60, 389–406. doi:10.1016/0012-821x(82)90075-9

Stein, C. A., and Stein, S. (1992). A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature* 359, 123–129. doi:10.1038/359123a0

Stein, C. A., and Stein, S. (1994). Constraints on hydrothermal heat flux through the oceanic lithosphere from global heat flow. *J. Geophys Res.* 99, 3081–3095. doi:10.1029/93jb02222

Stoks, P. G., and Schwartz, A. W. (1982). Basic nitrogen-heterocyclic compounds in the Murchison meteorite. *Geochim. Cosmochim. Acta* 46, 309–315. doi:10.1016/0016-7037(82)90222-8

Strazewski, P. (2019). The essence of systems chemistry. *Life* 9, 60–69. doi:10.3390/life9030060

Swart, P. K., Grady, M. M., Pillinger, C. T., Lewis, R. S., and Anders, E. (1983). Interstellar carbon in meteorites. *Science* 220 (4595), 406–410. doi:10.1126/science.220.4595.406

Takada, A. (1999). Variations in magma supply and magma partitioning: The role of tectonic settings. *J. Volcanol. Geotherm. Res.* 93, 93–110. doi:10.1016/S0377-0273(99)00082-7

Takahagi, W., Seo, K., Shibuya, T., Takano, Y., Fujishima, K., Saitoh, M., et al. (2019). Peptide synthesis under the alkaline hydrothermal conditions on Enceladus. *ACS Earth Space Chem.* 3, 2559–2568. doi:10.1021/acsearthspacechem.9b00108

Tartèse, R., Chaussidon, M., Gurenko, A., Delarue, F., and Robert, F. (2017). Warm Archaean oceans reconstructed from oxygen isotope composition of early-life remnants. *Geochem Perspect. Lett.* 3, 55–65. doi:10.7185/geochemlet.1706

Taylor, D. J., McKeegan, K. D., and Harrison, T. M. (2009). Lu-Hf zircon evidence for rapid lunar differentiation. *Earth Planet Sci. Lett.* 279, 157–164. doi:10.1016/j.epsl.2008.12.030

Thiemens, M. M., Sprung, P., Fonseca, R. O. C., Leitzke, F. P., and Münker, C. (2019). Early Moon formation inferred from hafnium-tungsten systematics. *Nat. Geosc* 12, 696–700. doi:10.1038/s41561-019-0398-3

Thordarson, T., and Larsen, G. (2007). Volcanism in Iceland in historical time: Volcano types, eruption styles and eruptive history. *J. Geodyn.* 43, 118–152. doi:10.1016/j.jog.2006.09.005

Tice, M. M., and Lowe, D. R. (2004). Photosynthetic microbial mats in the 3, 416-Myr-old ocean. *Nature* 431, 549–552. doi:10.1038/nature02888

Touboul, M., Kleine, T., Bourdon, B., Palme, H., and Wieler, R. (2009). Tungsten isotopes in ferroan anorthosites: Implications for the age of the Moon and lifetime of its magma ocean. *Icarus* 199, 245–249. doi:10.1016/j.icarus.2008.11.018

Trail, D., Boehnke, P., Savage, P. S., Liu, M. C., Miller, M. L., and Bindeman, I. (2018). Origin and significance of si and o isotope heterogeneities in phanerozoic, archean, and hadean zircon. *PNAS* 115, 10287–10292. doi:10.1073/pnas.1808335115

Trevors, J. T., and Pollack, G. H. (2005). Hypothesis: The origin of life in a hydrogel environment. *Prog. Biophys. Mol. Biol.* 89, 1–8. doi:10.1016/j.pbiomolbio.2004.07.003

Ueda, H., Shibuya, T., Sawaki, Y., Shozugawa, K., Makabe, A., and Takai, K. (2021). Chemical nature of hydrothermal fluids generated by serpentinization and carbonation of komatiite: Implications for H2-rich hydrothermal system and ocean chemistry in the early Earth. *Geochem Geophys Geosyst* 22, e2021GC009827. doi:10.1029/2021GC009827

USGS (2022). Usgs. Available: https://www.usgs.gov/volcanoes/yellowstone.

Valley, J. W., Cavosie, A. J., Ushikubo, T., Reinhard, D. A., Lawrence, D. F., Larson, D. J., et al. (2014). Hadean age for a post-magma-ocean zircon confirmed by atom-probe tomography. *Nat. Geosc* 7, 219–223. doi:10.1038/ngeo2075

van den Boorn, S. H., van Bergen, M. J., Nijman, W., and Vroon, P. Z. (2007). Dual role of seawater and hydrothermal fluids in early archean chert formation: Evidence from silicon isotopes. *Geology* 35, 939–942. doi:10.1130/g24096a.1

Van Kranendonk, M. J., Baumgartner, R., Djokic, T., Ota, T., Steller, L., Garbe, U., et al. (2021). Elements for the origin of life on land: A deep-time perspective from the Pilbara Craton of Western Australia. *Astrobiology* 21, 39–59. doi:10.1089/ast.2019.2107

Van Kranendonk, M. J., Djokic, T., Poole, G., Tadbiri, S., Steller, L., and Baumgartner, R. (2018). "Depositional setting of the fossiliferous, c. 3480 ma Dresser Formation, Pilbara craton: A review," in *Earth's oldest rocks*. Editors M. J. Van Kranendonk, V. Bennett, and E. Hoffmann (USA: Elsevier), 985–1006.

Van Kranendonk, M. J., Smithies, R. H., Hickman, A. H., and Champion, D. C. (2007). "Paleoarchean development of continental nucleus: the East Pilbara terrane of the Pilbara craton, western Australia," in *Earth's oldest rocksDev. In Precambrian Geology*. Editors M. J. van Kranendonk, R. H. Smithies, and V. C. Bennet (Elsevier), 15, p307–p338.

Van Zuilen, M. A., Lepland, A., Teranes, J., Finarelli, J., Wahlen, M., and Arrhenius, G. (2003). Graphite and carbonates in the 3.8 Ga old Isua supracrustal belt, southern West Greenland. *Precamb Res.* 126, 331–348. doi:10.1016/s0301-9268(03)00103-7

Vanderhaeghe, O., Guergouz, C., Fabre, C., Duchêne, S., and Baratoux, D. (2019). Secular cooling and crystallization of partially molten Archaean continental crust over1Ga. *ComptesRendus. Géoscience, Académiedes Sci.* 351, 562–573. doi:10.1016/j.crte.2019.07.002

Vauchez, A., Tommasi, A., and Mainpric, D. (2012). Faults (shear zones) in the Earth's mantle. *Tectonophysics* 558-559, 1–27. doi:10.1016/j.tecto.2012.06.006

Vennemann, T. W., and Smith, H. S. (1999). "Geochemistry of mafic and ultramafic rocks in the type section of the kromber formation, Barberton greenstone belt, South Africa," in *Geologic evolution of the Barberton greenstone Belt, South Africa, geological society of America special paper 329*. Editors D. L. Lowe and G. R. Byerly (Boulder, CO: Geological Society of America), 133–149.

Villafañe-Barajas, S. A., and Colín-García, M. S. (2021). Hydrothermal vent systems: the relevance of dynamic systems in chemical evolution and prebiotic chemistry experiments. *Int. J. Astrobiol.* 20, 427–434. doi:10.1017/S1473550421000331

Wächtershäuser, G. (2007). On the chemistry and evolution of the pioneer organism. *Chem. Biodiv* 4, 584–602. doi:10.1002/cbdv.200790052

Walsh, M. W. (1992). Microfossils and possible microfossils from the early archean onverwacht group, Barberton mountain land, South Africa. *Precamb Res.* 54, 271–293. doi:10.1016/0301-9268(92)90074-x

Wang, W., Qiao, L., He, J., Ju, Y., Yu, K., Kan, G., et al. (2021). Water microdroplets allow spontaneously abiotic production of peptides. *J. Phys. Chem. Lett.* 12, 5774–5780. doi:10.1021/acs.jpclett.1c01083

Weiss, A. (1981). Replication and evolution in inorganic systems. *Ang. Chem.* 20, 850–860. doi:10.1002/anie.198108501

Westall, F., Campbell, K. A., Bréhéret, J. G., Foucher, F., Gautret, P., Hubert, A., et al. (2015b). Archean (3.33 Ga) microbe-sediment systems were diverse and flourished in a hydrothermal context. *Geology* 43, 615–618. doi:10.1130/g36646.1

Westall, F., Cavalazzi, B., Lemelle, L., Marrocchi, Y., Rouzaud, J. N., Simionovici, A., et al. (2011b). Implications of *in situ* calcification for photosynthesis in a ~3.3 Ga-old microbial biofilm from the Barberton greenstone belt, South Africa. *Earth Planet Sci. Lett.* 310, 468–479. doi:10.1016/j.epsl.2011.08.029

Westall, F., de Ronde, C. E. J., Southam, G., Grassineau, N., Colas, M., Cockell, C., et al. (2006b). Implications of a 3.472–3.333 Gyr-old subaerial microbial mat from the Barberton greenstone belt, South Africa for the UV environmental conditions on the early Earth. *Phil Trans. Roy. Soc. Lond B* 361, 1857–1876. doi:10.1098/rstb.2006.1896

Westall, F., de Vries, S. T., Nijman, W., Rouchon, V., Orberger, B., Pearson, V., et al. (2006a). "The 3.466 Ga Kitty's gap chert, an early archaean microbial ecosystem," in *Processes on the early earth*. Editors W. U. Reimold and R. Gibson (Geol Soc Amer Spec Pub), 405, 105–131.b.

Westall, F., De Wit, M. J., Dann, J., Van Der Gaast, S., De Ronde, C., and Gerneke, D. (2001). Early Archaean fossil bacteria and biofilms in hydrothermally influenced, shallow water sediments, Barberton Greenstone Belt, South Africa. *Precamb Res.* 106, 91–112.

Westall, F., Foucher, F., Bost, N., Bertrand, M., Loizeau, D., Vago, J. L., et al. (2015a). Biosignatures on mars: What, where and how? Implications for the search for martian life. *Astrobiology* 15, 998–1029. doi:10.1089/ast.2015.1374

Westall, F., Hickman-Lewis, K., Hinman, N., Gautret, P., Campbell, K. A., Bréhéret, J. G., et al. (2018). A hydrothermal-sedimentary context for the origin of life. *Astrobiology* 18, 259–293. doi:10.1089/ast.2017.1680

Westall, F., Höning, D., Avice, G., Gentry, D., Gerya, T., Gillmann, C., et al. (2022). The habitability of venus and a comparison to early Earth. *Space Sci. Rev.* doi:10.1002/essoar.10512790.1

Westall, F., Steele, A., Toporski, J., Walsh, M., Allen, C., Guidry, S., et al. (2000). Polymeric substances and biofilms as biomarkers in terrestrial materials: Implications for extraterrestrial samples. *J. Geophys Res. Planets* 105 (24), 24511527–24524527. doi:10.1029/2000je001250

Westall, F., Foucher, F., Cavalazzi, B., de Vries, S. T., Nijman, W., Pearson, V., et al. (2011a). Volcaniclastic habitats for early life on earth and mars: A case study from ~3.5Ga-old rocks from the Pilbara, Australia. *Planet Space Sci.* 59, 1093–1106. doi:10.1016/j.pss.2010.09.006

Wheat, C. G., and Mottl, M. J. (2004). "Geochemical fluxes through mid-ocean ridge flanks," in *Hydrogeology of the oceanic lithosphere*. Editors E. E. Davis and H. Elderfield (Cambridge, UK: Cambridge University Press), 627–658.

White, L. M., Shibuya, T., Vance, S. D., Christensen, L. E., Bhartia, R., Kidd, R., et al. (2020). Simulating serpentinization as it could apply to the emergence of life using the jpl hydrothermal reactor. *Astrobiology* 20, 307–326. doi:10.1089/ast.2018.1949

Whitehouse, M. J., Nemchin, A. A., and Pidgeon, R. T. (2017). What can hadean detrital zircon really tell us? A critical evaluation of their geochronology with implications for the interpretation of oxygen and hafnium isotopes. *Gondwana Res.* 51, 78–91. doi:10.1016/j.gr.2017.07.007

Wilde, S. A., Valley, J. W., Peck, W. H., and Graham, C. M. (2001). Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. *Nature* 409, 175–178. doi:10.1038/35051550

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.* 51, 221–271. doi:10.1128/mr.51.2.221-271.1987

Yadav, M., Kumar, R., and Krishnamurthy, R. (2020). Chemistry of abiotic nucleotide synthesis. *Chem. Rev.* 120, 4766–4805. doi:10.1021/acs.chemrev.9b00546

Yin, Q., Jacobsen, S. B., Yamashita, K., Blichert-Toft, J., Telouk, P., and Albarede, F. (2002). A short timescale for terrestrial planet formation from Hf-W chronometry of meteorites. *Nature* 18, 949–952. doi:10.1038/nature00995

Yuen, G. U., and Kvenvolden, K. A. (1973). Monocarboxylic acids in Murray and Murchison carbonaceous meteorites. *Nature* 246, 301–303. doi:10.1038/246301a0

Zahnle, K. J., Lupu, R., Dobrovolskis, A., and Sleep, N. H. (2015). The tethered moon. *Earth Planet. Sci. Lett.* 427, 74–82. doi:10.1016/j.epsl.2015.06.058

Zark, M., Christoffers, J., and Dittmar, T. (2017). Molecular properties of deep-sea dissolved organic matter are predictable by the central limit theorem: Evidence from tandem FT-ICR-MS. *Mar. Chem.* 191, 9–15. doi:10.1016/j.marchem.2017.02.005

Zawaski, M. J., Kelly, N. M., Orlandini, O. F., Nichols, C. I., Allwood, A. C., and Mojzsis, S. J. (2020). Reappraisal of purported ca. 3.7 Ga stromatolites from the Isua supracrustal belt (West Greenland) from detailed chemical and structural analysis. *Earth Planet Sci. Lett.* 545, 116409. doi:10.1016/j.epsl.2020.116409

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 10 February 2023
DOI 10.3389/fspas.2023.1075052

Check for updates

OPEN ACCESS

EDITED BY
Alberto Fairén,
Center for Astrobiology (CSIC), Spain

REVIEWED BY
Teresa Fornaro,
Osservatorio Astrofisico di Arcetri (INAF), Italy
Jennifer Claire Stern,
Goddard Space Flight Center, National Aeronautics and Space Administration, United States

*CORRESPONDENCE
Arif H. Ansari,
a.h.ansari@bsip.res.in,

SPECIALTY SECTION
This article was submitted to Astrobiology, a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 20 October 2022
ACCEPTED 31 January 2023
PUBLISHED 10 February 2023

CITATION
Ansari AH (2023), Detection of organic matter on Mars, results from various Mars missions, challenges, and future strategy: A review.
Front. Astron. Space Sci. 10:1075052.
doi: 10.3389/fspas.2023.1075052

COPYRIGHT
© 2023 Ansari. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Detection of organic matter on Mars, results from various Mars missions, challenges, and future strategy: A review

Arif H. Ansari *

Birbal Sahni Institute of Palaeosciences, Lucknow, India

This review paper summarizes the literature on the organic matter detection by various Mars lander/rover missions, in order to understand the progress towards dealing with methodological challenges in the analysis of the Martian regolith and drilled mudstone samples. This paper shows that Martian missions are so far successful in detecting simple and some complex organic molecules, but their origin i.e., whether sourced by cosmic dust, meteoric bombardment, geochemical reduction of inorganic carbon during hydrothermal or igneous activity, or produced biologically in the ancient habitable Martian deposition environment, remains unknown. The preservation of organic matter in the Martian depositional realm has also been found as one of the biggest hurdles in its search. Therefore, upcoming ExoMars mission has been equipped with the instruments that would be able to drill and retrieve 2 m subsurface cores for organic matter analysis, with the assumption that the subsurface samples would have better chances of preserving original organic matter from the disintegration by ultraviolet (UV) radiation, galactic cosmic rays, and solar energetic particles. In addition to the method used for organic matter detection in previous missions [simple pyrolysis-GCMS and the use of combination of thermal combustion and derivatization (thermochemolysis)-GCMS], other alternative organic matter detection methodologies i.e., Raman spectroscopy (laser 523) plus deep resonant Raman and fluorescence spectroscopy are used in Mars 2020 Perseverance rover and will be used in ExoMars mission as well. Learning from the past and upcoming Mars missions will help in developing strategies and tools for the future Martian missions with goal to better understand it is ancient habitability.

KEYWORDS

Mars, organic matter detection, Vikings, Phoenix, Curiosity, ExoMars, extraterrestrial life, Perseverance

Introduction

There is compelling evidence that during the first half-billion years (4 to 3.5 Ga), Mars was wetter i.e., its surface and shallow subsurface hosted liquid water (Nisbet and Sleep, 2001; Baker, 2006; Greenwood et al., 2008). As a result, Mars has become the most promising target for astrobiological studies. However, the major question is how long this water must persist for life to evolve. Based on the Earth's geological record the estimated upper limit for the origins of life is considered 500 million years (Davis and McKay, 1996). Using a simple climate model, Davis and McKay (1996) determined that liquid water habitats on Mars could have survived for up to 500 million years under relatively thin ice covers with an atmospheric temperature below freezing point. Thus, the timeframe for liquid water on early Mars appears

to be equivalent to the maximum timeframe for the emergence of life on Earth (Davis and McKay, 1996). In the last 20 years, the scientific search for life on Mars has been taken on two major aspects: 1) exploring the Martial geological records for probable evidences of extinct life, 2) looking for signs of extant microbial life on Mars surface and subsurface environments. Progress has been made in both the directions. However, the search for extinct life on Mars has received more attention (Carrier et al., 2020) mainly due to the fact that early Martian condition was relatively more hospitable for life (Fairén et al., 2010).

There is a general consensus that life, if it exists elsewhere, is based on carbon chemistry, needs nitrogen, highly information-rich organic compounds, energy, substrates that allow for the synthesis of the organic compounds, and liquid water (Marais and Walter, 1999; Schulze-Makuch et al., 2002; Stern and Jedrzejas, 2008). Therefore, understanding how organic substances are synthesized abiogenically throughout the universe also piques our intrinsic curiosity (Oparin, 1965; Lemmon, 1970; Becker et al., 1999; Sephton and Hazen, 2013). Organic molecule signatures of early life on Earth are largely destroyed by tectonic recycling. Such tectonic recycling of early biotic or pre-biotic organic signatures on Mars seems unlikely as no evidence of past tectonic activity has been recorded on this planet (Mahaffy, 2008). However, the presence of superoxide radicals (Yen et al., 1999; Yen et al., 2000), oxidized halides (Zent and McKay, 1994; Hecht et al., 2009), and radiation (Kminek and Bada, 2006) on the Martian surface may have transformed organic molecules, reducing the diversity of such compounds from earlier periods and limiting our ability to describe their origin. It is nonetheless important to note that the type of organic material found on Mars has a direct bearing on two most important scientific questions: 1) how life arose on Earth and, 2) how life in the universe may have originated. For this reason, several missions have been sent or are under preparation to Mars for the detection of organic compounds in surface/subsurface sediments and the atmosphere.

Several Martian meteorites and Martian surface sediments have been found to contain organic carbon, but their origin remains unclear (Biemann et al., 1977; Mazur et al., 1978; Steele et al., 2012; Glavin et al., 2013; Freissinet et al., 2015; Eigenbrode et al., 2018; Franz et al., 2020; Millan et al., 2022a; Millan et al., 2022b; Stern et al., 2022). There have been a number of sources proposed for the detected organic compounds: 1) terrestrial contamination (Biemann et al., 1977; Mazur et al., 1978; Jull et al., 1998); 2) chondritic meteoric inputs (Becker et al., 1999); 3) direct precipitation from aqueous fluids (Steele et al., 2007; Ming et al., 2014; Eigenbrode et al., 2018; Franz et al., 2020; House et al., 2022; Stern et al., 2022); and, 4) biotic remains (McKay et al., 1996). By confirming and understanding the origin and formation of organic carbon on Mars, we will be able to improve our understanding of its putative carbon cycle, the availability of carbon for biotic chemistry, life detection, and how to detect organic compounds on future extraterrestrial missions (McKay et al., 1996). This review article is focused on the published results of Martian hydrocarbon detection missions: Vikings, Phoenix, Curiosity, and Perseverance, highlighting the challenges in detecting indigenous organic carbon on Martian surface, determining their origin, and developing strategy for the future missions.

Existence of water and its role in the evolution and persistence of life on ancient Mars

In contrast to contemporary Mars, which is an aeolian planet containing intermittent liquid water on its surface (likely brines), ancient Mars appears to have had localized habitats that were compatible with the needs of putative early life similar to that existed on Earth (Grotzinger et al., 2014; Grotzinger et al., 2015). Based on crater counts, water channels on the Martian surface are estimated to have formed intermittently between 0.5 and 3.5 billion years ago (Malin, 1976; Masursky et al., 1977). This ancient fluvial areas are unlikely to have been relevant to the current life but might have played a significant role in the evolution of organic compounds and the possible existence of past life. However, the existence of liquid water on the Martian surface was probably sporadic and might have been too brief to allow life to emerge and flourish. So even if a large quantity of water might have flowed during these episodes, it would have vanished in a matter of weeks due to high evaporation rates (Masursky et al., 1977). However, the observations do not rule out the possibility that, in local oases, a steady production of liquid water could have counterbalanced the high evaporation rates and permitted the long-term existence of liquid water (Mazur et al., 1978; Grotzinger et al., 2015).

Findings from Martian meteorites

The presence of organics in meteorites apparently ejected from Mars has been investigated in several studies (McSween Jr, 1994; McKay et al., 1996; Sephton et al., 2002; Steele et al., 2007; Steele et al., 2012; Steele et al., 2016; Eigenbrode et al., 2018). These include reports of organic volatile substances such as benzene, toluene, C2 alkylbenzene, and benzonitrile found by pyrolysis GCMS in Nakhla (Sephton et al., 2002) and polycyclic aromatic hydrocarbons (PAHs) in the Antarctic Martian meteorite ALH 84001 (McKay et al., 1996). Sephton et al. (2002) identified no organic pyrolysis products in their samples of ALH 84001, although they did discover aromatic organics in EET A79001. In this type of study, the amount of terrestrial contribution to Martian meteorites' organic matter during their residence on Earth (102–104 years) was a primary concern. For PAHs (Becker et al., 1997), amino acids (Bada et al., 1998), and acid-insoluble organic material (Jull et al., 2000), this issue has received considerable attention. Nevertheless, the extent of terrestrial contamination can be assessed by analyzing the environment from which the meteorite was collected, looking for molecules that appear to have been produced abiotically, and measuring carbon and hydrogen isotopes accurately (Mahaffy, 2008; Elsila et al., 2011; Aponte et al., 2014; Elsila et al., 2016; Glavin et al., 2021).

There has been evidence of macromolecular carbon (MMC) in Martian meteorites dating back from 4.2 Ga to 190 Ma (Borg and Draper, 2003; Lapen et al., 2010). Hirschmann and Withers (2008) postulated that carbon cycling in the Martian environment was dominated by its consumption in the reduced mantle which would have limited the release of CO_2 to the Martian atmosphere. The presence of MMC species in reduced igneous rocks (Steele et al., 2012) supported the central tenet of Hirschmann and Withers (2008). In contrast, the absence of MMC in inclusions in the most oxidized

TABLE 1 Major Mars Lander missions and their organic analysis results.

Mission	Year of landing	Experiment site	Instrument	Pyrolysis temperature	Method	Detected organic molecules	References
Viking	1975	Chryse Planitia, Utopia Panitia	GCMS	up to 500°C	Pyrolysis	No trace of hydrocarbons	Biemann et al. (1977), Mazur et al. (1978)
						0.7–6.5 ppm organic carbon	Reanalysis by Navarro-González et al. (2010)
Phoenix	2008	Vastitas Borealis, the arctic plains of Mars	GCMS	up to 820°C	Pyrolysis	No trace of hydrocarbons	Lauer et al. (2009)
Curiosity	2011	Rocknest aeolian deposits, Gale Crater	GCMS and Tunable Laser Spectrometer (TLS)	up to 835°C	Derivatization + Pyrolysis	Chloromethane, Dichloromethane, Trichloromethane, Chloromethylpropene, and Chlorobenzene	Glavin et al. (2013), Leshin et al. (2013), Ming et al. (2014)
		Sheepbed Lacustrine mudstones, Murray Formation, Pahrump Hills, Gale Crater	GCMS and TLS	up to 820°C	Derivatization + Pyrolysis	Dichloroethane, Dichloropane, Dichlorobutane, Thiophene, Methylthiophene, Benzene, Chlorobenzene, Toluene, Benzoic Acid, Menthanethiol, Dimethylsuplhide	Eigenbrode et al. (2018), Freissinet et al. (2015), Szopa et al. (2020)
		Glen Torridon, Gale Crator	GCMS and TLS	up to 850°C	Derivatization + Pyrolysis	Benzoic Acid, Phenol, Benzene, Toluene, Trimethyl- and Tetramethylbenzene, Naphthalene, Methylnaphthalen, Pentamethylbenzene, Benzoic Acid Methylester, Dimethyl-, Trimethyl-, and Tetramethylbenzenamine, Dihydronaphthalene, 2-Butylthiophene, and Benzothiophene	Millan et al. (2022a); Millan et al. (2022b)
Perseverance	2020	Jezero Crater, Nili Fossae region	Deep Ultraviolet Fluorescence and Raman spectrometry	No-pyrolysis	Deep Ultraviolet Fluorescence and Raman spectrometer scanning	Benzene, Naphthalene	Scheller et al. (2022)

sample (Nakhla) illustrates the possibility that the redox state of the Martian magmas might affect the preservation of these carbon species during crystallization (Sephton et al., 2002). This has implications for the distribution of MMC on Mars surface and within its crust (Michalski and Niles, 2010; Morris et al., 2010).

No evidence of CH_4 or CO_2 has been reported from igneous inclusions of Martian Meteorites (Steele et al., 2012). Although there is some skepticism about its existence at all (Zahnle et al., 2011), CH_4 has been found in the Martian atmosphere and has been explained by both abiotic and biotic processes (Atreya et al., 2007; Mumma et al., 2009; Webster et al., 2018). According to Steele et al. (2012), occurrence of MMC in igneous rocks could be used as a supporting evidence for abiotic generation of CH_4 (Atreya et al., 2007; Mumma et al., 2009). The discovery of ~190 million years old MMC-containing Martian meteorite, suggests that reduced organic carbon production on Mars has been active in the recent past and it is possible that the Martian carbon cycle may still be active (Steele et al., 2012). Furthermore, according to Steele et al. (2012), primary organic carbon is almost always present in Martian basaltic rocks. It was transported to the surface over the majority of Martian geologic history, as recently as ~190 million years ago, through igneous mechanisms rather than biological ones. Photolysis of organic carbon (MMC) in the surface sediments or rocks can be a possible abiotic source of CH_4 in Martian atmosphere (Keppler et al., 2012; Schuerger et al., 2012).

Findings from various Martian missions

Viking mission

The main goal of the Viking mission was to look for signs of current or past life on Mars (Biemann et al., 1977; Mazur et al., 1978). Viking found clear evidence of the presence of water vapour and water ice, as well as strong inferences that there was once a significant amount of underlying permafrost on Mars. But it found no indication of existing liquid water with the high chemical potential necessary for the survival of terrestrial life, which is in line with the known pressure-temperature relations on the planet's surface (Mazur et al., 1978). The elemental analysis of the atmosphere and regolith revealed that the chemical elements typically thought to be necessary for terrestrial life are present (Mazur et al., 1978). Viking's GCMS studies carried by Biemann et al. (1977) and Mazur et al. (1978) found no unambiguous indigenous organic compound. According to Biemann et al. (1977) and Mazur et al. (1978), minor traces of volatile chloro-hydrocarbons detected in Martian regolith samples (Table 1) by the Viking's GCMS

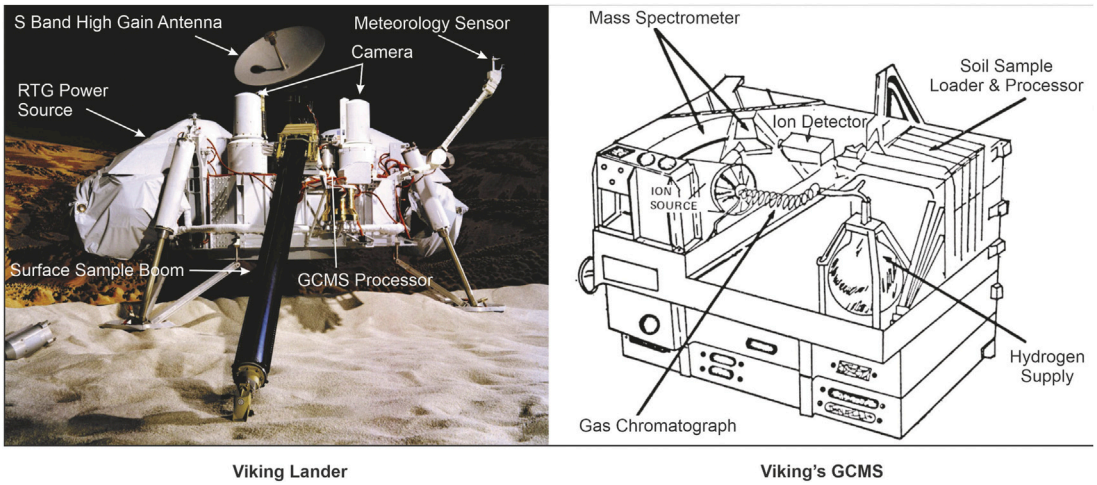


FIGURE 1
Viking Lander and GCMS instruments [sources: Huidobro et al. (2022) and NASA (2010)].

(Figure 1) was a result of terrestrial contamination. Furthermore, they stated that if volatile organic compounds were present in the regolith samples, they were either limited to substances like methane, which were undetectable or only detectable at low sensitivity levels, or they were present in concentrations below the detection limit of the instrument.

The reasons for the negative results of Viking's GCMS in detecting the presence of indigenous organic compounds were debated in subsequent studies (Benner et al., 2000; Glavin et al., 2013). According to Benner et al. (2000) and Glavin et al. (2013) the Viking's GCMS design was not optimized for the detection of certain potential hydrocarbons found in Martian meteorites (Sephton et al., 2002) which could have been present in the Martian samples. Benner et al. (2000) re-examined the Viking's GCMS results in light of what is known about the oxidation of organic compounds more generally as well as the makeup of organics that are likely to arrive on Mars *via* meteorite. They came to the conclusion that non-volatile salts of benzenecarboxylic acids, as well as possibly oxalic and acetic acid, should be metastable intermediates of meteoritic hydrocarbons under oxidizing conditions. For the Viking's GCMS, the salts of these non-volatile organic acids would have been largely undetectable, as the organic molecule must first go through a GC in order to access the MS, which is only possible for volatile molecules. Under pyrolysis, oxalic acid produces carbon dioxide, carbon monoxide, and water (Lewis et al., 2021). In fact, all of these were detected in the Viking's GCMS, but all these molecules are also present in the Martian atmosphere. Thus, Benner et al. (2000) and Mahaffy (2008) presented three possible reason for the failure of Viking's GCMS in detecting indigenous organic matter in the Martian regolith's: 1) the design of the gas-processing system that safeguarded the vacuum ion pumps decreased the GCMS's sensitivity for light organic compounds by a factor of 1,000; 2) the Viking GC column would not have transmitted numerous kinds of polar organic compounds, such as carboxylic acids, which are probably oxidation products of aliphatic and aromatic hydrocarbons; 3) instead of less permeable materials that would have been better protected from oxidants, the Viking sample-acquisition device could only collect loosely consolidated particles samples.

After the discovery of perchlorate in the Martian arctic soils by the Phoenix lander (Hecht et al., 2009), Navarro-González et al. (2010) ran an analogous experiment in which they heated the Mars-like soil from the Atacama desert containing 32 ppm of organic carbon with 1% magnesium perchlorate and demonstrated that nearly all of the organic present in the samples were decomposed to CO_2 and water, with the production of chloromethane and dichloromethane in trace amounts. Navarro-Gonzalez reconsidered the analysis results and suggested the burning of Martian soil sample would essentially decompose all the organics into water and CO_2. Based on this experiment they developed a chemical kinetics model for the degree of oxidation and chlorination of organics in the Viking oven and re-analysed the Viking results. According to Navarro-González et al. (2010) re-analysis results, ≤0.1% perchlorate and 1.5–6.5 ppm organic carbon were present in the sample of Viking landing site 1, and ≤1% perchlorate and 0.7–2.6 ppm organic carbon were present in the sample of Viking landing site 2 (Table 1).

Phoenix mission

Phoenix mission touched down on Mars' polar surface in March 2008 on a valley floor covered by the Scandia Formation, which is thought to be an ancient deposit from the Amazonian period (Arvidson et al., 2008; Smith et al., 2009). This formation encircles the northern edge of a shield volcano called Alba Patera (Arvidson et al., 2008). Phoenix heated soil and ice samples in the Thermal and Evolved-Gas Analyzer (TEGA) (Figure 2) in order to look for organic compounds (Boynton et al., 2009). For a mass range of 2–140 daltons, TEGA was made up of eight differential scanning calorimeter (DSC) ovens that were connected with a mass spectrometer (Boynton et al., 2009). The TEGA DSC monitored endothermic and exothermic processes when samples were heated from ambient temperature to roughly 1,000°C. Throughout the heating process, the mass spectrometer simultaneously detected all evolved gases, including any organic compounds and their byproducts.

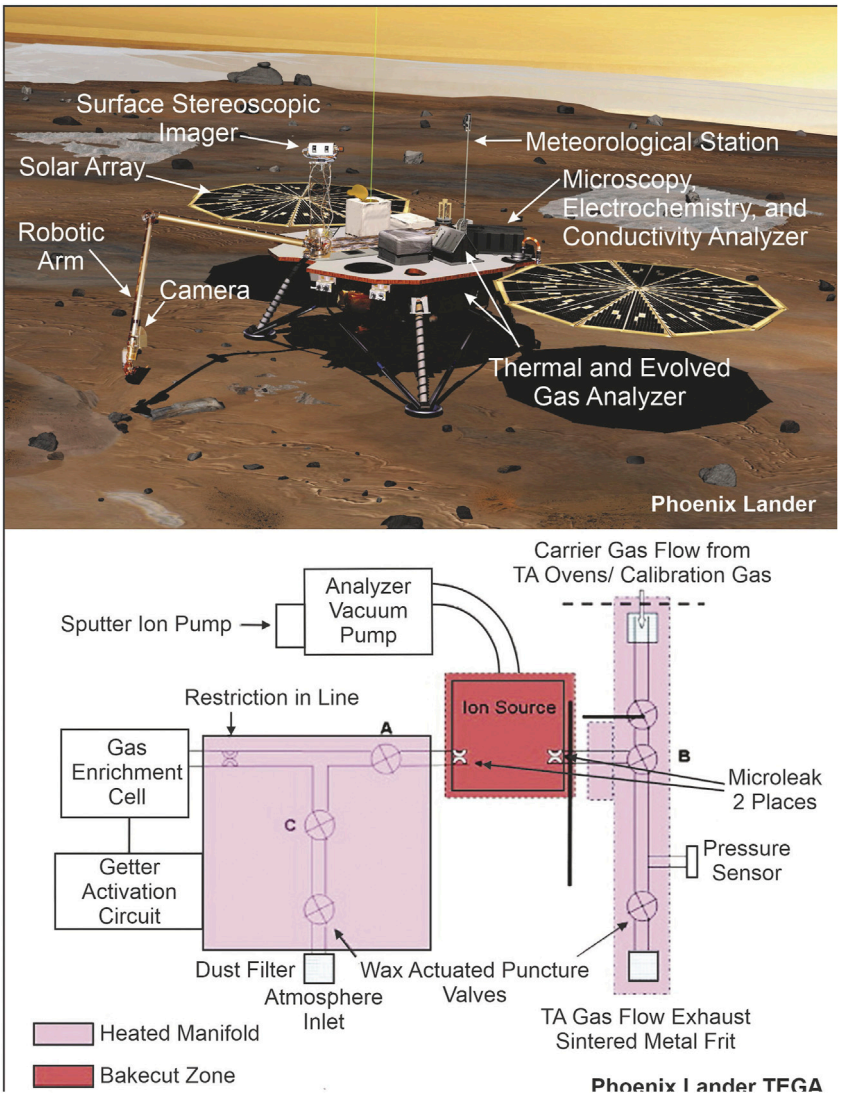


FIGURE 2
Phoenix Lander and TEGA instruments [sources: Huidobro et al. (2022) and NASA (2008)].

The evolved gas analysis (EGA) using Phoenix TEGA data did not reveal any organic fragments (Lauer et al., 2009; Ming et al., 2009). A perchlorate salt was found in the soil of Mars by the Microscopy, Electrochemical and Conductivity Analyzer (MECA) and Wet Chemistry Lab (WCL) (Hecht et al., 2009). It is probable that the O_2 generated during the thermal decomposition of the perchlorate salt ignited organic fragments evolved in the temperature range of 300°C–600°C. During the heating, CO_2 was produced as a byproduct of the burning of organic molecules. According to Ming et al. (2014), the possible causes of low temperature release of CO_2 in Phoenix TEGA could be desorption of adsorbed CO_2, thermal breakdown of Fe- and Mg-carbonates, and combustion of organic molecules in the Martian soils.

Curiosity mission

This is one of the strongest and long-standing missions which landed on Mars (in August 2012) in Aeolis Palus area inside Gale Crater. The main goal of Curiosity is to search the possible traces of life on Mars using the Sample Analysis Mars (SAM) lab suite it carried (Figure 3). SAM onboard the Curiosity rover has the capability to detect complex hydrocarbon molecules having astrobiological implication, by coupling the EGA and GCMS with two type of wet chemistry methods: 1) organic derivatization by adding N-methyl-N-(ter-butyldimethylsilyl) trifluoroacetamide (MTBSTFA) and dimethylformamide (DMF) in 4:1 by volume, 2) organic derivatization by adding 25% tetramethylammonium hydroxide (TMAH) in methanol (Millan et al., 2022a; Millan et al., 2022b). Glavin et al. (2013) reported first SAM-EGA-GCMS data of a Martian solid sample, from Rocknest aeolian deposit of Gale Crater and provided an extensive discussion on the possible origins of chlorinated hydrocarbons detected by SAM. Their studies detected several chlorinated chemical species above EGA background concentration including hydrochloric acid, chloromethane, and dichloromethane (Table 1). As these chlorinated organic species were not detected in the blank EGA

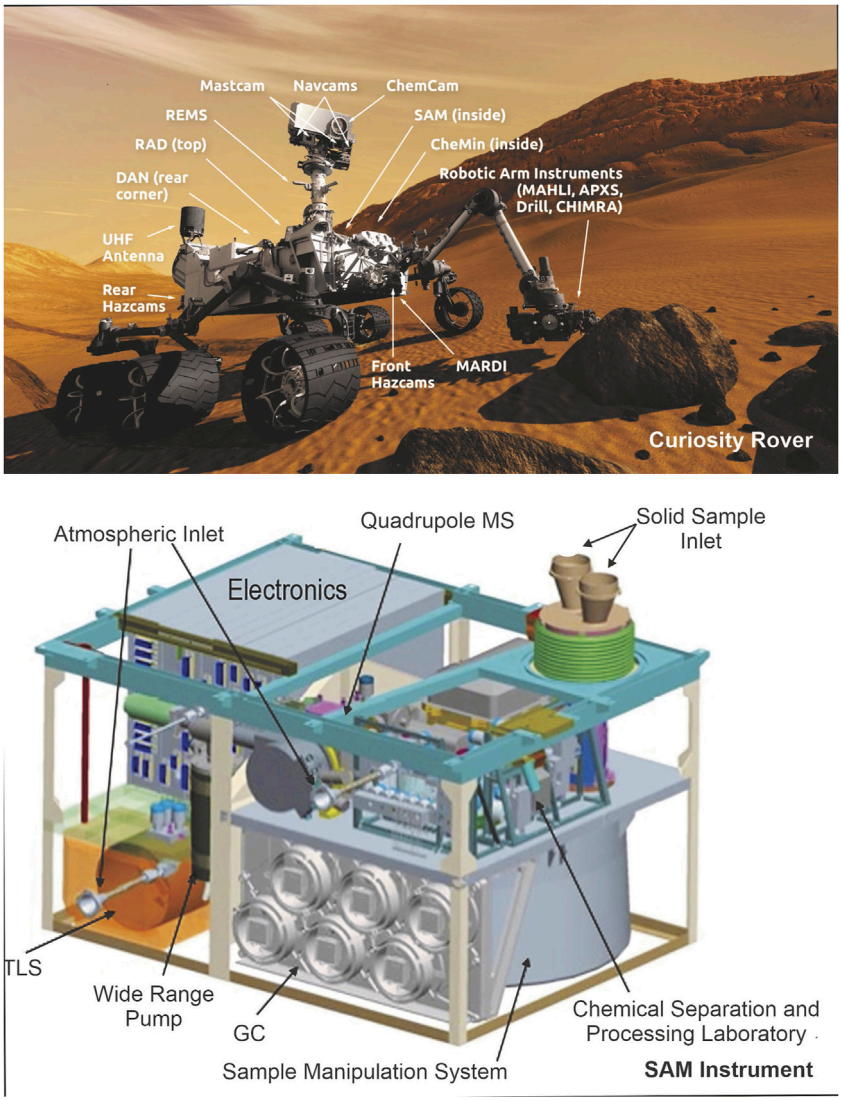


FIGURE 3
Curiosity Rover and SAM instruments [sources: Huidobro et al. (2022) and NASA (2012)].

run of SAM, the source of this organic carbon was combustion of reduced carbon in the presence of oxygen from oxychlorine. In the blank run 1,3-bis (1,1-dimethylethyl)-1,1,3,3-tetramethyldisiloxane and tert-butyldimethylsilanol were detected by GCMS. These two compounds are the fragmentary products of MTBSTFA (Glavin et al., 2013; Leshin et al., 2013; Millan et al., 2022a; Millan et al., 2022b). Therefore, it is assumed that the MTBSTFA and DMF vapor was perhaps leaked and reacted either with terrestrial or Martian water during the pyrolysis. A major difference between the peaks of 1,3-bis(1,1-dimethylethyl)-1,1,3,3-tetramethyldisiloxane for blank run and Rocknest sample run was that during the pyrolysis of the Rocknest sample, the peak intensity of 1,3-bis(1,1-dimethylethyl)-1,1,3,3-tetramethyldisiloxane sharply decreased with increase in O_2 peak. This according to Glavin et al. (2013) was a result of MTBSTFA decomposition along with the chlorate minerals present in the samples which may have contributed to some chlorinated hydrocarbons. However, GCMS analysis of Rocknest samples after pyrolysis detected chlorinated hydrocarbons such as chloromethane, dichloromethane, trichloromethane, choromethylpropene, and chlorobenzene (Table 1) significantly above the background level recorded in black run (Glavin et al., 2013). The concentration of chlorohydrocarbons detected by EGA-GCMS was more than 50 times higher than detected by the Viking-GCMS (Biemann et al., 1977). But, this comparison of detected concentration by the Viking and SAM is less relevant as unlike SAM, Viking did not carry derivatizing agent MTBSTFA. Nevertheless EGA-GCMS analysis at various pyrolysis temperature cut (>500°C) demonstrated that light chlorinated hydrocarbons can be effectively purged from SAM gas processing unit between the analyses. Chlorobenzene was also detected in the blank GCMS run of SAM and to some level above in the Rocknest samples. Its most likely source has been traced to thermal degradation of SAM organic trap (Tenax TA) containing 2,6-diphenylphenylene oxide which can produce Benzene, Toluene, and lesser amount of biphenyl. Hence its sourcing from Rocknest sample was suggested unlikely (Glavin et al., 2013).

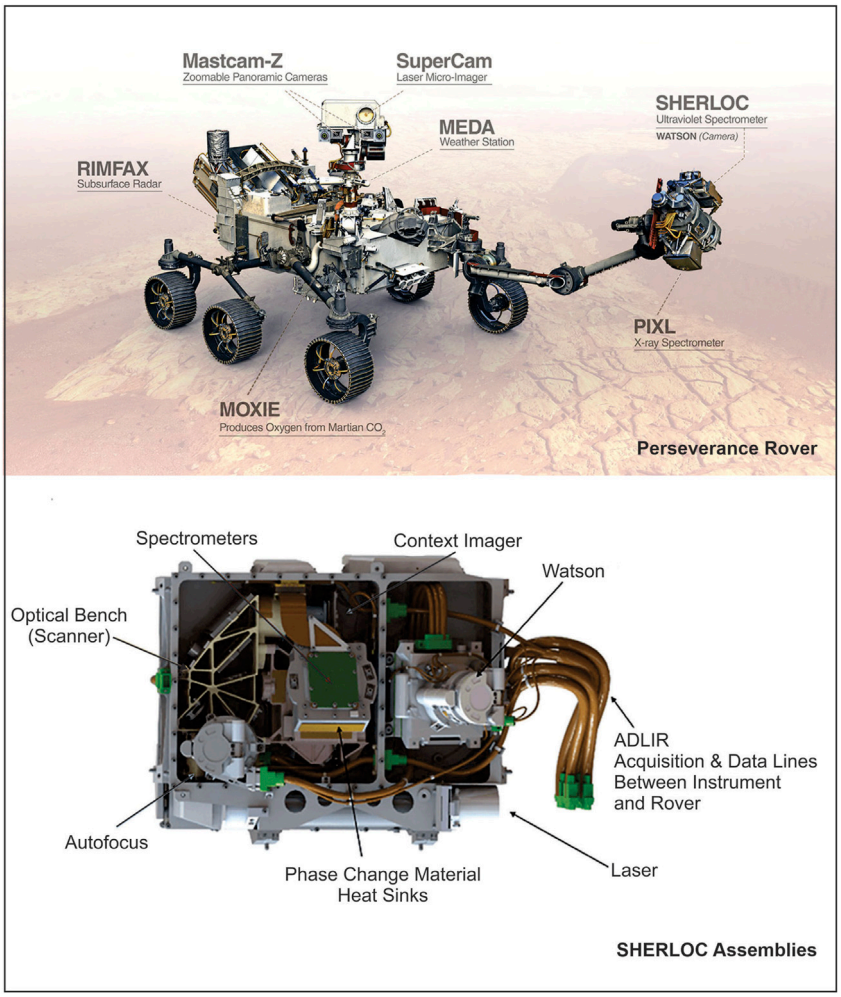


FIGURE 4
Perseverance Rover and SHERLOC instruments [Huidobro et al. (2022) and NASA (2020)].

Based on the results from EGA-GCMS, Freissinet et al. (2015) reported detection of 150–300 ppb chlorobenzene and up to 70 ppb C2-C4 dichloroalkanes in several locations of the Cumberland drill hole in the Sheepbed Mudstone at Yellowknife Bay (Table 1). Their findings indicate that when combined with EGA-GCMS data from multiple scooped and drilled samples, blank runs, and laboratory analogs, the relatively high concentrations of dichloroalkanes and chlorobenzene cannot be a result of the instrument's internal sourcing. The Sheepbed Mudstone may contain chlorohydrocarbons as such (Szopa et al., 2020), but they are thought to be the products of chemical reactions occurring during pyrolysis between Martian oxychloride and the sample's organic compounds. The finding of chlorinated organic molecules close to the surface indicates that reduced organic material with covalent bonds has endured despite the presence of several oxidants and high-energy radiation exposure due to the thin Martian atmosphere. Thus, the outcomes of the Freissinet et al. (2015) studies made a significant contribution to identifying possible preservation windows for reduced organic molecules on the Martian surface. In addition, this perspective on ancient Mars offers some context for a potentially habitable environment and represents a first step toward comprehending the existence and variety of potential prebiotic or biotic chemical fingerprints.

To circumvent the perchlorate degradation product of organic matter in the Martian sample analysis, Eigenbrode et al. (2018) analyzed the gases produced exclusively above 400°C, as salt present in the Martian regolith breakdown during heating to a temperature of 200°C in the SAM. Their results provided unambiguous evidence of organic molecules such as thiophenic, aromatic, and aliphatic hydrocarbon from the drill samples of Gale Crater. These sulfur-bearing hydrocarbons are considered to have originated from Mars surface (i.e. igneous, hydrothermal, atmospheric, or biological) and indicates presence of refractory organic materials in the samples (Freissinet et al., 2015; Eigenbrode et al., 2018).

The MTBSTFA derivatization experiment on Mars performed on Ogunquit Beach dune samples of Gale Crater detected several molecules containing benzoic acid, phenol, ammonia and phosphoric acid structures (Millan et al., 2022a; Millan et al., 2022b). Such molecules were either sourced from organic molecules on surface and subsurface environment due to strong oxidizing condition or from MTBSTFA present in SAM

TABLE 2 Lab based flash pyrolysis (at > 500°C) products of derivatization agents MTBSTFA and DMF (He, 2021).

Derivatization agent	Pyrolysis temperature (°C)	Derived organic compound	Reference
MTBSTFA	600	Difluorodimethylsilane, 4-methyl-1-hexene, Tert-Butyldimethylsiyl fluoride, Tert-Butyldimethyksilyl 2,2,2- trifluoroacetate, O-(trimethylsilyl)phenol, Tert-Butyl-isocyanatodimethylsilane, N-methyl-N-(trimethylsilytrifluoroacetamide, 1,3-ditert-butyl-1,1,3,3,- termethyldisiloxane, Pentamethyldisiloxane, Bis(trimethylsiloxy)methylsilane, Tris (trimethylsilyl) borate, Bis [tert-butyl (dimethyl)silyl]oxalate	He (2021)
	850	Methyltrifluorosilane, Tert-Butyldimethylfluorodilane, Fluorobenzene, 1, 4-difluoro-Benzene, Benzene, Trifluoromethybenzene, N-methyltrifluoroacetamide, Pyridine, Toluene, Isocyanatotrimethylsilane, O-(Trimethyl)phenol, Ethylbenzene, Tert-Butyl-isocyanatodimethylsilane, Styrene, 2-Propenylbenzene, 3-methyl-Benzonitrile,2 -Methyl-naphthelene, Biphenyl, Acenaphthylene, Phenanthrene, Pyreneetc.	
DMF	600	N, N-Dimethylformamide, Trimethylamine, Dimehtylamine, Ethanolamine, N,N,N′,N′-Tetramethylmethanediamine, Dimethylaminoacetonitrile, Styrene, Hydracrylonitrile, Benzene, Naphthaleneetc.	
	800	N, N-Dimethylformamide, Styrene, Hydracrylonitrile, Benzene, Naphthaleneetc.	

background or both (Millan et al., 2022a; Millan et al., 2022b). First TMAH experiment on Mars was performed at Mary Anning (MA) drill site in the Glen Torridon region of Gale Crater that detected a range of hydrocarbons e.g., benzene, toluene, trimethyl- and tetramethyl-benzene, naphthalene, methylnaphthalene, pentamethyl-benzene, benzoic acid methyl ester, dimethyl-, trimethyl-, and tetramethyl-benzenamine, dihydronaphthalene, 2-butyl-thiophene, and benzothiophene (Millan et al., 2022a; Millan et al., 2022b) (Table 1). As TMAH is a methylating agent that break polar bonds and methylate pyrolysis products *in situ*, suggests that hydrocarbons might have derived from macromolecules which was methylated by TMAH thermochemolysis. In this way pentamethyl benzene could be a product of multimethylation during TMAH reaction. Benzoic acid is detected both in SAM background and Mars samples, therefore its TMAH methylated product benzoic acid methyl ester is of currently unknown sources. The benzenamine may have originated on Mars or it could be a result of benzene reacting again with the N in TMAH. In addition, thiophene has been found in both refractory and mature organic phases in Martian meteorites (Steele et al., 2016; Eigenbrode et al., 2018). Hence, thiophenes found on Mars are thought to be pyrolysis breakdown products from various macromolecules.

Perseverance

The Perseverance rover (Figure 4) was launched by NASA in 2020 (Farley et al., 2020) in an effort to determine the potential habitability of Mars and possible evidence of past life. This mission landed (on February 2021) at Jazero Crater, a 45 km diameter impact structure in the Nili Fossae region of Mars, where a lake and river delta existed approximately 3.5 billion years ago (Mangold et al., 2007; Goudge et al., 2012; Goudge et al., 2015; Goudge et al., 2017; Farley et al., 2020). The main instrument onboard Perseverance to detect signs of past life is SHERLOC (Scanning Habitable Environments with Raman and Luminescence for Organics and Chemicals) (Hollis et al., 2021; Hollis et al., 2022). SHERLOC (Figure 4) uses a deep ultraviolet (DUV) laser to detect fluorescence emissions from aromatic compounds, as well as Raman scattered photons to identify organic functional groups, chemicals, and minerals (Bhartia et al., 2021). This enables SHERLOC to improve Martian studies by analyzing the abundance and variety of organic matter at the fine scale in different minerals to comprehend the potential for organic matter preservation in the Martian surface soils/outcrops, and help resolving organic biosignatures such as hopanes, steranes, and other organic macromolecules (Abbey et al., 2017; Bhartia et al., 2021). SHERLOC's micro-mapping mode allows non-destructive, sub-picogram sensitive organic detection (Bhartia et al., 2010) to identify smaller-scale features. In order to confirm that all the observed spectral characteristics are intrinsic to the material, the spectrometer's dust cover will occasionally be opened, and the instrument laser will be fired into the starry night sky (Bhartia et al., 2021). This method makes it possible to assess and eliminate any slight spectrum contamination from the optics or signature from Mars dust pollution deposited on optics from the data processing pipeline. Raman and Fluorescence data from first 208 Martian days of Perseverance on Mars has recently been published by Scheller et al. (2022). These analysis were conducted on three rock targets in two lithological units of the Jazero Crater floor. Reports of detailed scanning of the three samples by Scheller et al. (2022) demonstrated presence of carbonate, sulfate, phosphate, amorphous silicate bearing minerals in association with aromatic organic compounds most likely single- and double-ring aromatics. Carbonation of olivine in presence of briny water has been interpreted as an origin of carbonate minerals, whereas sulfate and perchlorate in the subsurface samples are indicated to have formed by precipitation from briny water. Presence of aromatic organic compounds in all the three targeted rocks is suggested by DUV analysis however, Raman analysis was unable to detect presence of organic compounds most likely due to insufficient organic matter concentration.

Future Mars sample return

For more than 4 decades, the planetary science community has placed a high focus on Mars Sample Return (MSR) (Beaty et al., 2019). By analyzing Mars samples in terrestrial laboratories, we could gain a

greater understanding of Mars than we could achieve with just *in situ* missions. Perseverance is capable of collecting samples to return to Earth. In support of its goal to collect sample for possible future Earth return, Perseverance carries an entirely new sampling and catching subsystem (SCS) (Townsend et al., 2022). SCS is used to collect rock core and regolith samples into individual ultraclean and sterile tubes, to photo-document and assess the collected volume of each sample, and to hermetically seal each tube. The sample tube can be manipulated through a sophisticated internal robotic mechanism called the Adaptive Catching Assembly (ACA) (Townsend et al., 2022), which can place the tube on the surface individually, at any time, and in one or more locations from which they can be retrieved. To maximize the likelihood of success of the Mars Sample Return (MSR) campaign duplicates have been collected for each particular target rock. One copy will be deposited at the end of the prime mission in the Jazero deposit identified at Three Forks, to ensure their availability for return to Earth in case Perseverance fail in the future. The second copy will be carried in the rover during its extended mission, with aim of collecting additional samples that the rover will directly deliver to the lander of the retrieval mission. The Mars 2020 science team has identified two hypothetical sample caches in order to examine: 1) the geology of the Jezero Crater, which will be collected during the prime mission, and, 2) the ancient crust outside of the Jezero Crater, which will be collected during a potential extended mission (Herd et al., 2022). The recent successful sample collection by the Perseverance rover and ongoing work by the National Aeronautics and Space Administration (NASA) and European Space Agency (ESA) on developing missions that could retrieve and transport the samples to Earth have put MSR on track to become a reality (Kminek et al., 2022).

Preservation of organic matter on Mars surface

Comets, asteroids, meteorites, and interplanetary dust particles deliver organic compounds to planetary surfaces abiotically (Chyba and Sagan, 1992; Ehrenfreund et al., 2011). Based on scaling the terrestrial influx to a Martian scenario and measurements of the Martian atmosphere, the estimated current influx of abiotic organic molecules to the surface of Mars is around 100–300 tons per year (Eigenbrode et al., 2018). Given that majority of the Martian surface is covered by billions of years old rocks (Carr and Head, 2010), it must have received abundance of organic molecules over this time span. But various Mars lander missions so far found it challenging to detect organic molecules unambiguously on the Martian surface. This suggests a more rapid degradation (by ionic radiation or oxidizing agents) of organic matter than accumulation. However, the crystal structures of minerals found on Mars, such as sulfates and clay minerals, may retain organic molecules, shielding them from the hostile environment (Keil and Mayer, 2014).

The results of the Mars missions to date mostly point to the existence of powerful oxidants and ultraviolet light, which can significantly shorten the residence period of organic matter on the Martian surface and represent the biggest barrier to its identification using *in situ* techniques. However, recent evidences of organic compounds from SAM observations (Glavin et al., 2013; Freissinet et al., 2015; Eigenbrode et al., 2018; Millan et al., 2022a; Millan et al., 2022b) demonstrate the possibilities and ways to detect *in situ* organic matter on Mars. Moreover, Applin et al. (2015) found that oxalate minerals are particularly susceptible to accumulation at the surface, as their experiments indicate they are as stable to UV radiation as sulfates and carbonates. Other significant degrading agents for organic materials on Mars are solar energetic particles and galactic cosmic rays which are more powerful than UV as its influence can extends much deeper (up to 9 m) in the subsurface than UV light (Pavlov et al., 2002; Dartnell et al., 2007b; a; Pavlov et al., 2012; Pavlov et al., 2022).

According to Fornaro et al. (2018), it is extremely difficult to develop a general mechanism to predict the behavior of Martian minerals. However, an accurate simulation of Martian environment considering all the key factors would contribute significantly to understanding the trends. Studies of organic matter association with different mineralogy have shown that some minerals aid in organic preservation while others speed up their degradation (Fornaro et al., 2018; Brucato and Fornaro, 2019). For instance, Poch et al. (2015) found that when adenine and glycine molecules were exposed to mid-UV irradiation in iron (III) smectite clay and non-tronite at very high molecule-mineral mass ratios, the effectiveness of the two organic molecules' photodecomposition was reduced by a factor of 5. This suggests not only physical shielding by non-tronite but also organic molecules stabilization by their interaction with the minerals allow recombination of dissociated organic molecule fragments (Poch et al., 2015). Furthermore, dos Santos et al. (2016) studied preservation of 25 amino acids in various minerals like olivine, enstatite, goethite, hematite, gypsum, jarosite, labradorite, augite, montmorillonite, non-tronite, and saponite, as well as basaltic lava. In their studies, clay minerals were found to exhibit amino acid-protective properties. The typical characteristics of clay minerals, such as their high surface area, which promotes molecular adsorption, their small pore sizes, which prevent radiation from penetrating, and their ideal interlayer sites for the accommodation of organic compounds, which create a shielded environment against external agents, were used to explain this behavior (dos Santos et al., 2016). Such minerals have been detected at several sites on Mars (Bishop et al., 2008; Ehlmann et al., 2009; Ehlmann et al., 2011a; Ehlmann et al., 2011b; Noe Dobrea et al., 2012; Carter et al., 2013; Ehlmann and Edwards, 2014; Vaniman et al., 2014; Adeli et al., 2015; Michalski et al., 2015) which has also been seen as a trace of past active hydrosphere. Millan et al. (2022b) discovered that highest diversity of organic molecules in Martian soil was associated with samples having highest proportion of clay and calcium sulfate minerals. These results comply with the recent findings that phyllosilicate clay minerals show better preservation potential for organic molecules, especially in active Fe-Mn redox environment (Gasda et al., 2022).

Detection of high abundance of thiophenic and other sulfur containing organic carbon in the Murray Formation mudstone (Confidence Hills) infer the possibility that sulfurization of organic matter during deposition or during early diagenesis would also have added their preservation (Eigenbrode et al., 2018). Observations from Earth's natural environment have indicated that sulfurization of organic matter increases their refractory properties by reducing reactive functional groups and adding small cross links between small unstable molecules (Aubrey et al., 2006; Kotler et al., 2008). Incidentally, studies of terrestrial organic-mineral interaction have demonstrated better preservation of organic molecules by entrapment within the crystal structure (Aubrey et al., 2006; Kotler et al., 2008).

Significant capacity to store organic molecules has also been seen in terrestrial hypersaline halite- and perchlorate subsurface deposits, which are common on Mars (Fernández-Remolar et al., 2013).

Applin et al. (2015), who reinterpreted the results collected by various pyrolysis-based instruments aboard Viking, Phoenix, and Curiosity, have recently reported the potential existence of oxalates on Mars in various regions (Biemann et al., 1977; Boynton et al., 2009; Mahaffy et al., 2012; Ming et al., 2014). Furthermore, a re-analysis of a Martian mudstone by Applin et al. (2015) using X-ray diffraction (XRD) of the Mars Science Laboratory (MSL) onboard Curiosity rover revealed that the data are compatible with the existence of refractory Ca, Fe, or Mg oxalates at low levels. It has been shown that oxalates can be produced from other organic compounds through both biological (Johnston and Vestal, 1993; Baran, 2014) and abiotic pathways (Durand and Nicaise, 1980; Peltzer et al., 1984). By interacting with aqueous cations, oxalate in solution forms weddellite, whewellite, and glushinskite, which are extremely resistant to chemical and physical degradation (Wilson et al., 1980; Chen et al., 2000; Frost and Weier, 2003; Frost, 2004). According to Cheng et al. (2016) oxalate formation may be a widespread process in the Martian surface condition and potentially play a key role in global carbon cycling and organic matter geochemistry.

Challenges in distinguishing between biogenic and abiogenic hydrocarbons

Various hydrocarbons can be formed abiotically in terrestrial and extraterrestrial environments by high temperature and pressure Fischer-Tropsch synthesis (McCollom, 2003). Aliphatic hydrocarbons and amino acids are among the byproducts of Fischer-Tropsch type synthesis, according to laboratory investigations; these substances may be oxidized at the surface to more stable molecules. Other abiotic processes may also result in the formation of organics, possibly even in the current Martian environment. Through photolytic or catalytic reactions triggered by light at ultraviolet and visible wavelengths, many organic compounds can be produced, for examples photolysis of water and CO_2 mixtures (Franz et al., 2020). Formic acid, formaldehyde, and oxalate, as well as methane and alcohol, have been produced in laboratory studies by these processes. It has also been demonstrated that ultraviolet irradiation of siderite in the presence of water produces formate, formaldehyde, and its derivatives (Franz et al., 2020). These organic compounds may be oxidized on the Martian surface to form oxalate minerals (Benner et al., 2000; Eigenbrode et al., 2018).

SAM stepped combustion experiment study on Yellowknife Bay Mudstone at Gale Crater, Mars by Stern et al. (2022) demonstrated that CO_2 evolved at < 550°C was enriched in ^{13}C ($\delta^{13}C$ = +1.5‰ ± 3.8‰), based on the mass balance calculation they suggested that 431 μg C/g of the released CO_2 represented indigenous organic and inorganic Martian carbon. At > 550°C released CO_2 was enriched in ^{12}C ($\delta^{13}C$ = −32.9 to −10.1‰) which represented 273 ± 30 μg C/g of mineral-bound refractory organic carbon. This amount of organic carbon is 40 times more than reported by Eigenbrode et al. (2018) and even higher than reported in Martian meteorites (Steele et al., 2016). Franz et al. (2020) found the isotopic composition to be compatible with macromolecular organic carbon derived from igneous sources, meteoritic infall, diagenetically-altered biomass, or a mix of these. Therefore, the source of high-temperature organic carbon cannot be conclusively identified.

Before 2003, all efforts to find statistically significant methane on Mars were unsuccessful (Maguire, 1977; Krasnopolsky et al., 1997; Lellouch et al., 2000). Subsequently four research groups claimed CH_4 detection (Formisano et al., 2004; Krasnopolsky et al., 2004; Mumma et al., 2009; Webster et al., 2018; Giuranna et al., 2019). Using Thermal Emission Spectrometer onboard Mars Global Surveyor, Fonti and Marzo, (2010) suggested presence of CH_4 on the Martian surface, however, later re-analysis of the data could not reproduce unambiguous evidence for CH_4 (Fonti et al., 2015). Many other researchers failed to detect CH_4 on Mars (Krasnopolsky, 2007; Villanueva et al., 2013; Aoki et al., 2018; Korablev et al., 2019). More recently House et al. (2022) reported anomalously ^{13}C depleted methane (−137‰ ± 8‰) in the drilled samples from multiple horizons of the Gale Crater of Mars using SAM instruments and skeptically suggested three possible scenarios for its production: 1) photolysis of biological methane, 2) photoreduction of CO_2, and 3) deposition of cosmic dust. Nevertheless, they concluded that identifying which of the three scenarios most likely represented the condition that existed on the 3 billion years old Martian surface is not possible based on the current knowledge.

Future Mars mission strategy to detect hydrocarbons

The prerequisites for a future Mars rover mission that would look for organic chemicals on the Martian surface and address their biogenic or abiotic origin must be understood by researchers based on the outcomes of previous Mars missions. So far organic matter search on Mars has achieved moderate success by demonstrating the release of simple organic molecules (i.e., aliphatic hydrocarbons) and some larger molecules (thiophenic and aromatic hydrocarbons) during the pyrolysis of drilled samples from the sedimentary outcrops (Freissinet et al., 2015; Eigenbrode et al., 2018). However, there is still no information about the source organic material present in sedimentary rocks. Also, it is unknown whether that organic matter is linked with a specific mineral phase. Given these uncertainties, combining previously used experimental techniques with new ones to analyze samples collected at a future landing site using both old and new approaches is the most meaningful strategy for continuing the *in situ* research of organic Mars. This will assure continuity throughout all landed Mars missions until samples are returned from Mars and new data can be compared to that from earlier missions.

The ExoMars rover mission, which will launch in the near future, is a component of the ExoMars project, which also includes the 2016 ExoMars orbiter mission. The ExoMars rover's high-level scientific objectives are: 1) to look for evidence of past and extant life on Mars, and 2) analyze geochemical and water conditions with increasing depth in the shallow subsurface of Mars. The most important capability of the ExoMars rover payload instrument is to drill and retrieve a 2 m deep core that will be crushed and analyzed for organic matter using Mars Organic Matter Analyzer (Goetz et al., 2016). According to experimental and modeling research (Kminek and Bada, 2006; Pavlov et al., 2012), Martian rocks at a depth of 2 m are expected to remain less affected by UV radiation and galactic cosmic radiation (Pavlov et al., 2002; Pavlov et al., 2012; Pavlov et al., 2022).

In the case of a complicated depositional scenario, where the deposition and denudation cycle takes place, determining whether the subsurface Martian samples were unaffected throughout its history is

extremely challenging. For example, given that generally Mars underwent low volcanic activity during the past two billion years, the deposition scenario in which sediment was rapidly deposited, and subsequently exposed for long-term denudation (Gale Crater) might also be a common process on the surface of Mars. However, it is conceivable that the rover could find organic-bearing strata that primarily underwent insignificant exposure to high particle radiation over the geological timescale and may have preserved near original signature of ancient organic deposit. Thus, drilling down to a depth of 2 m would be a huge step toward gaining access to such least altered organic molecule containing Martian rock samples. In order to search the ancient life on ancient Mars, the ExoMars rover requires this insight back in time.

Derivatization/thermochemolysis-GCMS, Laser Desorption and Ionization (LDI)-MS, and Pyrolysis-GCMS are the three operational modes of MOMA. MOMA-MS is not made to identify light molecules (e.g. CO_2, N_2, O_2, CH_4, and HCl) that are either a part of Martian atmosphere or change during pyrolysis, even though that mass threshold of MS can be modified during operations (Goetz et al., 2016). In MOMA, derivatization agent (DA) will be chosen during surface operations on Mars based on the kinds of organic residues that are anticipated to be found in the sample. Based on the success of SAM derivatization experiments MTBSTFA seems to be the most versatile DA and it is expected to be used in MOMA. Amino acid or carboxylic acid with chiral centers will be separated using dimethylformamide dimethylacetal (DMF-DMA) that preserves the asymmetrical center and is used with the enantio-selective GC column onboard MOMA to achieve chiral separation. For example, since biology on Earth specifically uses left-handed amino acids, one of the main goals of ExoMars mission will be to determine the homochirality or enantiomeric excess of potential Martian amino acids. Another selected DA is TMAH that will be helpful in volatilizing non-soluble refractory kerogen fraction. In addition, laser techniques such as MOMA-LDI and Raman spectroscopy can be used to avoid thermal decomposition of organic matter (Li et al., 2015).

Discussion and conclusion

From Viking to Mars 2020, rover missions searching for signatures of extinct life in Martian surface rocks and sediments have moderately advanced our knowledge of challenges lying ahead in finding organic molecules of biological origin. These challenges also include understanding the carbon cycling of Mars, identifying sites with better preservation potential for hydrocarbons, developing more refined methods for improved extraction of indigenous organic molecules with minimum contamination. The lack of a non-ambiguous detection of organic molecule by pyrolysis-GCMS method onboard Viking and Phoenix missions (Biemann et al., 1977; Mazur et al., 1978; Lauer et al., 2009) made researchers to reevaluate the design and methods of the corresponding units for the subsequent missions. These findings prompted the use of derivatization wet chemistry in the subsequent rover mission i.e., Curiosity and the future ExoMars mission to enhance the chances of organic molecule detection. This method uses derivatization agents such as MTBSTFA, DMF, and TMAH to improve the thermal stability of compounds and ameliorate volatility by substituting their active polar functional groups (Schummer et al., 2009). Using the derivatization/thermochemolysis method Curiosity detected several Cl, S, and, N containing aliphatic and aromatic compounds (Table 1) above the background level of the instrument (Glavin et al., 2013; Leshin et al., 2013; Ming et al., 2014; Freissinet et al., 2015; Eigenbrode et al., 2018; Millan et al., 2022a; Millan et al., 2022b). There are, however, some unanswered questions about the origin of the detected organic compounds: 1) How much of the detected organic compounds represent Martian indigenous organic compounds? 2) Which and how much of the detected organic compounds are derived from indigenous biogenic/abiogenic hydrocarbons?

MTBSTFA, DMF, and TMAH have been generally employed as derivatizing agents for organic compound studies on Earth with GCMS temperature ramping up between 50°C and 200°C at which the derivatization agents are found relatively stable (Molnar-Perl and Katona, 2000; Schummer et al., 2009). But the MTBSTFA, DMF, and TMAH derivatization experiments by Curiosity on Mars has been conducted at a higher temperature (generally >200°) at which the degradation product of these derivatization agents is not properly known. Recently, a lab based simple flash pyrolysis study of MTBSTFA and DMF demonstrated that both the derivatization agent degrade at higher temperature to produce several organic products including dimethylamino-acetonitrile, styrene, benzene, phenol, entylbenzene, 3-methylebenzonitrile, toluene, napthalene, methyl-naphthalene, naphtonitrile, pyrene, above ≥500°C (Table 2) (He, 2021). Thus, results from Martian missions employing derivatization agents can have contaminant complex organic compounds and need to be interpreted with caution. A systematic study of high temperature degradation products of MTBSTFA, DMF, and TMAH in Martian analog conditions can be helpful in resolving such issue and in refining our current understanding of complex hydrocarbons and their origin in Martian surface and subsurface samples.

The use of non-destructive techniques, such as the Raman and fluorescence spectroscopy currently employed by Perseverance, and the Raman and infrared spectroscopy plus the laser desorption techniques that will be employed by the future ExoMars mission, will minimize possible alteration and contamination of the original organic material due to the measurements themselves. This is essential to provide more insights into the nature of the organics detected in the Martian samples, while we wait for their return to Earth and analysis with more sensitive techniques.

Author contributions

AA: Conceptualization, writing an original draft, data Curation, reviewing and writing the final draft.

Acknowledgments

I would like thank my colleagues from NASA spaceward bound India initiative for their time-to-time updates on Mars related research that encouraged me to write this review paper. Also, I want to express my gratitude to the reviewers for their insightful comments, which helped this manuscript's quality tremendously. The BSIP contribution number for this MS is BSIP/RDCC/Publication no 47/2022-2023.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbey, W. J., Bhartia, R., Beegle, L. W., Deflores, L., Paez, V., Sijapati, K., et al. (2017). Deep UV Raman spectroscopy for planetary exploration: The search for *in situ* organics. *Icarus* 290, 201–214. doi:10.1016/j.icarus.2017.01.039

Adeli, S., Hauber, E., Le Deit, L., and Jaumann, R. (2015). Geologic evolution of the eastern Eridania basin: Implications for aqueous processes in the southern highlands of Mars. *J. Geophys. Res. Planets* 120, 1774–1799. doi:10.1002/2015je004898

Aoki, S., Richter, M., Dewitt, C., Boogert, A., Encrenaz, T., Sagawa, H., et al. (2018). Stringent upper limit of CH4 on Mars based on SOFIA/EXES observations. *Astronomy Astrophysics* 610, A78. doi:10.1051/0004-6361/201730903

Aponte, J. C., Dworkin, J. P., and Elsila, J. E. (2014). Assessing the origins of aliphatic amines in the Murchison meteorite from their compound-specific carbon isotopic ratios and enantiomeric composition. *Geochimica Cosmochimica Acta* 141, 331–345. doi:10.1016/j.gca.2014.06.035

Applin, D., Izawa, M., Cloutis, E., Goltz, D., and Johnson, J. (2015). Oxalate minerals on mars? *Earth Planet. Sci. Lett.* 420, 127–139. doi:10.1016/j.epsl.2015.03.034

Arvidson, R., Adams, D., Bonfiglio, G., Christensen, P., Cull, S., Golombek, M., et al. (2008). Mars Exploration Program 2007 Phoenix landing site selection and characteristics. *J. Geophys. Res. Planets* 113, E00A03. doi:10.1029/2007je003021

Atreya, S. K., Mahaffy, P. R., and Wong, A.-S. (2007). Methane and related trace species on Mars: Origin, loss, implications for life, and habitability. *Planet. Space Sci.* 55, 358–369. doi:10.1016/j.pss.2006.02.005

Aubrey, A., Cleaves, H. J., Chalmers, J. H., Skelley, A. M., Mathies, R. A., Grunthaner, F. J., et al. (2006). Sulfate minerals and organic compounds on Mars. *Geology* 34, 357–360. doi:10.1130/g22316.1

Bada, J. L., Glavin, D. P., Mcdonald, G. D., and Becker, L. (1998). A search for endogenous amino acids in Martian meteorite ALH84001. *Science* 279, 362–365. doi:10.1126/science.279.5349.362

Baker, V. R. (2006). Geomorphological evidence for water on Mars. *Elements* 2, 139–143. doi:10.2113/gselements.2.3.139

Baran, E. J. (2014). Review: Natural oxalates and their analogous synthetic complexes. *J. Coord. Chem.* 67, 3734–3768. doi:10.1080/00958972.2014.937340

Beaty, D. W., Grady, M. M., Mcsween, H. Y., Sefton-Nash, E., Carrier, B. L., Altieri, F., et al. (2019). The potential science and engineering value of samples delivered to Earth by Mars sample return: International MSR Objectives and Samples Team (iMOST). *Meteorit. Planet. Sci.* 54, S3–S152. doi:10.1111/maps.13242

Becker, L., Glavin, D. P., and Bada, J. L. (1997). Polycyclic aromatic hydrocarbons (PAHs) in Antarctic Martian meteorites, carbonaceous chondrites, and polar ice. *Geochimica Cosmochimica Acta* 61, 475–481. doi:10.1016/s0016-7037(96)00400-0

Becker, L., Popp, B., Rust, T., and Bada, J. L. (1999). The origin of organic matter in the Martian meteorite ALH84001. *Adv. Space Res.* 24, 477–488. doi:10.1016/s0273-1177(99)00090-3

Benner, S. A., Devine, K. G., Matveeva, L. N., and Powell, D. H. (2000). The missing organic molecules on Mars. *Proc. Natl. Acad. Sci.* 97, 2425–2430. doi:10.1073/pnas.040539497

Bhartia, R., Beegle, L. W., Deflores, L., Abbey, W., Razzell Hollis, J., Uckert, K., et al. (2021). Perseverance's scanning habitable environments with Raman and luminescence for organics and chemicals (SHERLOC) investigation. *Space Sci. Rev.* 217, 1–115.

Bhartia, R., Salas, E. C., Hug, W. F., Reid, R. D., Lane, A. L., Edwards, K. J., et al. (2010). Label-free bacterial imaging with deep-UV-laser-induced native fluorescence. *Appl. Environ. Microbiol.* 76, 7231–7237. doi:10.1128/aem.00943-10

Biemann, K., Oro, J., Toulmin, P., Iii, Orgel, L., Nier, A., Anderson, D., et al. (1977). The search for organic substances and inorganic volatile compounds in the surface of Mars. *J. Geophys. Res.* 82, 4641–4658. doi:10.1029/js082i028p04641

Bishop, J. L., Dobrea, E. Z. N., Mckeown, N. K., Parente, M., Ehlmann, B. L., Michalski, J. R., et al. (2008). Phyllosilicate diversity and past aqueous activity revealed at Mawrth Vallis, Mars. *Science* 321, 830–833. doi:10.1126/science.1159699

Borg, L. E., and Draper, D. S. (2003). A petrogenetic model for the origin and compositional variation of the Martian basaltic meteorites. *Meteorit. Planet. Sci.* 38, 1713–1731. doi:10.1111/j.1945-5100.2003.tb00011.x

Boynton, W., Ming, D., Kounaves, S., Young, S., Arvidson, R., Hecht, M., et al. (2009). Evidence for calcium carbonate at the Mars Phoenix landing site. *Science* 325, 61–64. doi:10.1126/science.1172768

Brucato, J. R., and Fornaro, T. (2019). "Role of mineral surfaces in prebiotic processes and space-like conditions," in *Biosignatures for Astrobiology* (Springer), 183–204.

Carr, M. H., and Head, J. W., III (2010). Geologic history of mars. *Earth Planet. Sci. Lett.* 294, 185–203. doi:10.1016/j.epsl.2009.06.042

Carrier, B. L., Beaty, D. W., Meyer, M. A., Blank, J., Chou, L., Dassarma, S., Des Marais, D., Eigenbrode, J., Grefenstette, N., and Lanza, N. (2020). Mars extant life: What's next? Conference report. *Astrobiology* 20, 785–814. doi:10.1089/ast.2020.2237

Carter, J., Poulet, F., Bibring, J. P., Mangold, N., and Murchie, S. (2013). Hydrous minerals on Mars as seen by the CRISM and OMEGA imaging spectrometers: Updated global view. *J. Geophys. Res. Planets* 118, 831–858. doi:10.1029/2012je004145

Chen, J., Blume, H.-P., and Beyer, L. (2000). Weathering of rocks induced by lichen colonization—A review. *Catena* 39, 121–146. doi:10.1016/s0341-8162(99)00085-5

Cheng, Z., Fernández-Remolar, D., Izawa, M., Applin, D., Chong Díaz, M., Fernandez-Sampedro, M., et al. (2016). Oxalate formation under the hyperarid conditions of the Atacama desert as a mineral marker to provide clues to the source of organic carbon on Mars. *J. Geophys. Res. Biogeosciences* 121, 1593–1604. doi:10.1002/2016jg003439

Chyba, C., and Sagan, C. (1992). Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: An inventory for the origins of life. *Nature* 355, 125–132. doi:10.1038/355125a0

Dartnell, L. R., Desorgher, L., Ward, J. M., and Coates, A. J. (2007a). Martian sub-surface ionising radiation: Biosignatures and geology. *Biogeosciences* 4, 545–558. doi:10.5194/bg-4-545-2007

Dartnell, L. R., Desorgher, L., Ward, J. M., and Coates, A. J. (2007b). Modelling the surface and subsurface Martian radiation environment: Implications for astrobiology. *Geophys. Res. Lett.* 34, L02207. doi:10.1029/2006gl027494

Davis, W. L., and Mckay, C. P. (1996). Origins of life: A comparison of theories and application to mars. *Orig. Life Evol. Biosphere* 26, 61–73. doi:10.1007/bf01808160

Dos Santos, R., Patel, M., Cuadros, J., and Martins, Z. (2016). Influence of mineralogy on the preservation of amino acids under simulated Mars conditions. *Icarus* 277, 342–353. doi:10.1016/j.icarus.2016.05.029

Durand, B., and Nicaise, G. (1980). *Kerogen: Insoluble organic matter from sedimentary rocks*. Paris: Editions technip, 35–53.

Ehlmann, B. L., and Edwards, C. S. (2014). Mineralogy of the martian surface. *Annu. Rev. Earth Planet. Sci.* 42, 291–315. doi:10.1146/annurev-earth-060313-055024

Ehlmann, B. L., Mustard, J. F., Clark, R. N., Swayze, G. A., and Murchie, S. L. (2011a). Evidence for low-grade metamorphism, hydrothermal alteration, and diagenesis on Mars from phyllosilicate mineral assemblages. *Clays Clay Minerals* 59, 359–377. doi:10.1346/ccmn.2011.0590402

Ehlmann, B. L., Mustard, J. F., Murchie, S. L., Bibring, J.-P., Meunier, A., Fraeman, A. A., et al. (2011b). Subsurface water and clay mineral formation during the early history of Mars. *Nature* 479, 53–60. doi:10.1038/nature10582

Ehlmann, B. L., Mustard, J. F., Swayze, G. A., Clark, R. N., Bishop, J. L., Poulet, F., et al. (2009). Identification of hydrated silicate minerals on Mars using MRO-CRISM: Geologic context near Nili Fossae and implications for aqueous alteration. *J. Geophys. Res. Planets* 114, E00D08. doi:10.1029/2009je003339

Ehrenfreund, P., Röling, W., Thiel, C., Quinn, R., Sephton, M., Stoker, C., et al. (2011). Astrobiology and habitability studies in preparation for future mars missions: Trends from investigating minerals, organics and biota. *Int. J. Astrobiol.* 10, 239–253. doi:10.1017/s1473550411000140

Eigenbrode, J. L., Summons, R. E., Steele, A., Freissinet, C., Millan, M., Navarro-González, R., et al. (2018). Organic matter preserved in 3-billion-year-old mudstones at Gale crater, Mars. *Science* 360, 1096–1101. doi:10.1126/science.aas9185

Elsila, J. E., Aponte, J. C., Blackmond, D. G., Burton, A. S., Dworkin, J. P., and Glavin, D. P. (2016). Meteoritic amino acids: Diversity in compositions reflects parent body histories. *ACS Central Sci.* 2, 370–379. doi:10.1021/acscentsci.6b00074

Elsila, J. E., Callahan, M. P., Glavin, D. P., Dworkin, J. P., and Brückner, H. (2011). Distribution and stable isotopic composition of amino acids from fungal peptaibiotics: Assessing the potential for meteoritic contamination. *Astrobiology* 11, 123–133. doi:10.1089/ast.2010.0505

Fairén, A. G., Davila, A. F., Lim, D., Bramall, N., Bonaccorsi, R., Zavaleta, J., et al. (2010). Astrobiology through the ages of mars: The study of terrestrial analogues to understand the habitability of mars. *Astrobiology* 10, 821–843. doi:10.1089/ast.2009.0440

Farley, K. A., Williford, K. H., Stack, K. M., Bhartia, R., Chen, A., De La Torre, M., et al. (2020). Mars 2020 mission overview. *Space Sci. Rev.* 216, 1–41. doi:10.1007/s11214-020-00762-y

Fernández-Remolar, D. C., Chong-Díaz, G., Ruíz-Bermejo, M., Harir, M., Schmitt-Kopplin, P., Tziotis, D., et al. (2013). Molecular preservation in halite- and perchlorate-rich hypersaline subsurface deposits in the Salar Grande basin (Atacama Desert, Chile): Implications for the search for molecular biomarkers on Mars. *J. Geophys. Res. Biogeosciences* 118, 922–939. doi:10.1002/jgrg.20059

Fonti, S., Mancarella, F., Liuzzi, G., Roush, T., Frouard, M. C., Murphy, J., et al. (2015). Revisiting the identification of methane on Mars using TES data. *Astronomy Astrophysics* 581, A136. doi:10.1051/0004-6361/201526235

Fonti, S., and Marzo, G. (2010). Mapping the methane on mars. *Astronomy Astrophysics* 512, A51. doi:10.1051/0004-6361/200913178

Formisano, V., Atreya, S., Encrenaz, T., Ignatiev, N., and Giuranna, M. (2004). Detection of methane in the atmosphere of Mars. *science* 306, 1758–1761. doi:10.1126/science.1101732

Fornaro, T., Steele, A., and Brucato, J. R. (2018). Catalytic/protective properties of martian minerals and implications for possible origin of life on Mars. *Life* 8, 56. doi:10.3390/life8040056

Franz, H., Mahaffy, P., Webster, C., Flesch, G., Raaen, E., Freissinet, C., et al. (2020). Indigenous and exogenous organics and surface–atmosphere cycling inferred from carbon and oxygen isotopes at Gale crater. *Nat. Astron.* 4, 526–532. doi:10.1038/s41550-019-0990-x

Freissinet, C., Glavin, D., Mahaffy, P. R., Miller, K., Eigenbrode, J., Summons, R., et al. (2015). Organic molecules in the sheepbed mudstone, gale crater, Mars. *J. Geophys. Res. Planets* 120, 495–514. doi:10.1002/2014je004737

Frost, R. L. (2004). Raman spectroscopy of natural oxalates. *Anal. Chim. Acta* 517, 207–214. doi:10.1016/j.aca.2004.04.036

Frost, R. L., and Weier, M. L. (2003). Raman spectroscopy of natural oxalates at 298 and 77 K. *J. Raman Spectrosc.* 34, 776–785. doi:10.1002/jrs.1052

Gasda, P. J., Comellas, J., Essunfeld, A., Das, D., Bryk, A. B., Dehouck, E., et al. (2022). Overview of the morphology and chemistry of diagenetic features in the clay-rich glen Torridon unit of Gale Crater, mars. *J. Geophys. Res. Planets* 127, e2021JE007097. doi:10.1029/2021je007097

Giuranna, M., Viscardy, S., Daerden, F., Neary, L., Etiope, G., Oehler, D., et al. (2019). Independent confirmation of a methane spike on Mars and a source region east of Gale Crater. *Nat. Geosci.* 12, 326–332. doi:10.1038/s41561-019-0331-9

Glavin, D. P., Elsila, J. E., Mclain, H. L., Aponte, J. C., Parker, E. T., Dworkin, J. P., et al. (2021). Extraterrestrial amino acids and L-enantiomeric excesses in the CM 2 carbonaceous chondrites Aguas Zarcas and Murchison. *Meteorit. Planet. Sci.* 56, 148–173. doi:10.1111/maps.13451

Glavin, D. P., Freissinet, C., Miller, K. E., Eigenbrode, J. L., Brunner, A. E., Buch, A., et al. (2013). Evidence for perchlorates and the origin of chlorinated hydrocarbons detected by SAM at the Rocknest aeolian deposit in Gale Crater. *J. Geophys. Res. Planets* 118, 1955–1973. doi:10.1002/jgre.20127

Goetz, W., Brinckerhoff, W., Arevalo, R., Freissinet, C., Getty, S., Glavin, D., et al. (2016). Moma: The challenge to search for organics and biosignatures on mars. *Int. J. Astrobiol.* 15, 239–250. doi:10.1017/s1473550416000227

Goudge, T. A., Milliken, R. E., Head, J. W., Mustard, J. F., and Fassett, C. I. (2017). Sedimentological evidence for a deltaic origin of the Western fan deposit in Jezero crater, Mars and implications for future exploration. *Earth Planet. Sci. Lett.* 458, 357–365. doi:10.1016/j.epsl.2016.10.056

Goudge, T. A., Mustard, J. F., Head, J. W., and Fassett, C. I. (2012). Constraints on the history of open-basin lakes on Mars from the composition and timing of volcanic resurfacing. *J. Geophys. Res. Planets* 117. doi:10.1029/2012je004115

Goudge, T. A., Mustard, J. F., Head, J. W., Fassett, C. I., and Wiseman, S. M. (2015). Assessing the mineralogy of the watershed and fan deposits of the Jezero crater paleolake system, Mars. *J. Geophys. Res. Planets* 120, 775–808. doi:10.1002/2014je004782

Greenwood, J. P., Itoh, S., Sakamoto, N., Vicenzi, E. P., and Yurimoto, H. (2008). Hydrogen isotope evidence for loss of water from Mars through time. *Geophys. Res. Lett.* 35, L05203. doi:10.1029/2007gl032721

Grotzinger, J. P., Gupta, S., Malin, M. C., Rubin, D. M., Schieber, J., Siebach, K., et al. (2015). Deposition, exhumation, and paleoclimate of an ancient lake deposit, Gale crater, Mars. *Science* 350, aac7575. doi:10.1126/science.aac7575

Grotzinger, J. P., Sumner, D. Y., Kah, L., Stack, K., Gupta, S., Edgar, L., et al. (2014). A habitable fluvio-lacustrine environment at Yellowknife Bay, Gale Crater, mars. *Science* 343, 1242777. doi:10.1126/science.1242777

He, Y. (2021). Search for organic compounds with MTBSTFA/DMF derivatization and TMAH thermochemolysis on Mars. *Univ. Paris-Saclay*.

Hecht, M. H., Kounaves, S. P., Quinn, R., West, S. J., Young, S. M., Ming, D. W., et al. (2009). Detection of perchlorate and the soluble chemistry of martian soil at the Phoenix lander site. *Science* 325, 64–67. doi:10.1126/science.1172466

Herd, C., Bosak, T., Stack, K., Sun, V., Gupta, S., Shuster, D., et al. (2022). *The notional plan for sample collections by the perseverance rover for Mars sample return*.

Hirschmann, M. M., and Withers, A. C. (2008). Ventilation of CO2 from a reduced mantle and consequences for the early Martian greenhouse. *Earth Planet. Sci. Lett.* 270, 147–155. doi:10.1016/j.epsl.2008.03.034

Hollis, J. R., Abbey, W., Beegle, L. W., Bhartia, R., Ehlmann, B. L., Miura, J., et al. (2021). A deep-ultraviolet Raman and Fluorescence spectral library of 62 minerals for the SHERLOC instrument onboard Mars 2020. *Planet. Space Sci.* 209, 105356. doi:10.1016/j.pss.2021.105356

Hollis, J. R., Moore, K. R., Sharma, S., Beegle, L., Grotzinger, J. P., Allwood, A., et al. (2022). The power of paired proximity science observations: Co-Located data from SHERLOC and PIXL on mars. *Icarus* 387, 115179. doi:10.1016/j.icarus.2022.115179

House, C. H., Wong, G. M., Webster, C. R., Flesch, G. J., Franz, H. B., Stern, J. C., et al. (2022). Depleted carbon isotope compositions observed at Gale crater, Mars. *Proc. Natl. Acad. Sci.* 119, e2115651119. doi:10.1073/pnas.2115651119

Huidobro, J., Aramendia, J., Arana, G., and Madariaga, J. M. (2022). Reviewing *in situ* analytical techniques used to research Martian geochemistry: From the Viking Project to the MMX future mission. *Anal. Chim. Acta* 1197, 339499. doi:10.1016/j.aca.2022.339499

Johnston, C. G., and Vestal, J. R. (1993). Biogeochemistry of oxalate in the Antarctic cryptoendolithic lichen-dominated community. *Microb. Ecol.* 25, 305–319. doi:10.1007/bf00171895

Jull, A., Beck, J., and Burr, G. (2000). Isotopic evidence for extraterrestrial organic material in the Martian meteorite, Nakhla. *Geochimica Cosmochimica Acta* 64, 3763–3772. doi:10.1016/s0016-7037(00)00458-0

Jull, A., Courtney, C., Jeffrey, D., and Beck, J. (1998). Isotopic evidence for a terrestrial source of organic compounds found in Martian meteorites Allan Hills 84001 and Elephant Moraine 79001. *Science* 279, 366–369. doi:10.1126/science.279.5349.366

Keil, R., and Mayer, L. (2014). *Mineral matrices and organic matter*.

Keppler, F., Vigano, I., Mcleod, A., Ott, U., Früchtl, M., and Röckmann, T. (2012). Ultraviolet-radiation-induced methane emissions from meteorites and the Martian atmosphere. *Nature* 486, 93–96. doi:10.1038/nature11203

Kminek, G., and Bada, J. L. (2006). The effect of ionizing radiation on the preservation of amino acids on Mars. *Earth Planet. Sci. Lett.* 245, 1–5. doi:10.1016/j.epsl.2006.03.008

Kminek, G., Meyer, M. A., Beaty, D. W., Carrier, B. L., Haltigin, T., and Hays, L. E. (2022). Mars sample return (MSR): Planning for returned sample science. *Astrobiology* 22, S-1–S-4. doi:10.1089/ast.2021.0198

Korablev, O., Vandaele, A. C., Montmessin, F., Fedorova, A. A., Trokhimovskiy, A., Forget, F., et al. (2019). No detection of methane on mars from early ExoMars trace gas orbiter observations. *Nature* 568, 517–520. doi:10.1038/s41586-019-1096-4

Kotler, J. M., Hinman, N. W., Yan, B., Stoner, D. L., and Scott, J. R. (2008). Glycine identification in natural jarosites using laser desorption fourier transform mass spectrometry: Implications for the search for life on mars. *Astrobiology* 8, 253–266. doi:10.1089/ast.2006.0102

Krasnopolsky, V. A. (2007). Comment on "methane on mars: A product of H2O photolysis in the presence of CO" by A. Bar-nun and V. Dimitrov. *Icarus* 188, 540–542. doi:10.1016/j.icarus.2006.12.019

Krasnopolsky, V. A., Maillard, J. P., and Owen, T. C. (2004). Detection of methane in the martian atmosphere: Evidence for life? *Icarus* 172, 537–547. doi:10.1016/j.icarus.2004.07.004

Krasnopolsky, V., Bjoraker, G., Mumma, M., and Jennings, D. (1997). High-resolution spectroscopy of mars at 3.7 and 8 μm: A sensitive search for H2O2, H2CO, HCl, and CH4, and detection of hdo. *J. Geophys. Res. Planets* 102, 6525–6534. doi:10.1029/96je03766

Lapen, T. J., Righter, M., Brandon, A., Debaille, V., Beard, B., Shafer, J., et al. (2010). A younger age for ALH84001 and its geochemical link to shergottite sources in Mars. *Science* 328, 347–351. doi:10.1126/science.1185395

Lauer, H., Ming, D., Sutter, B., Golden, D., Morris, R., and Boynton, W. (2009). "thermal and evolved gas analysis of magnesium perchlorate: Implications for perchlorates in soils at the mars Phoenix landing site perchlorates in soils at the mars Phoenix landing site," in *40th annual lunar and planetary science conference)*, 2196.

Lellouch, E., Encrenaz, T., De Graauw, T., Erard, S., Morris, P., Crovisier, J., et al. (2000). The 2.4– spectrum of Mars observed with the infrared space observatory. *Planet. Space Sci.* 48, 1393–1405. doi:10.1016/s0032-0633(00)00118-5

Lemmon, R. M. (1970). Chemical evolution. *Chem. Rev.* 70, 95–109. doi:10.1021/cr60263a003

Leshin, L., Mahaffy, P., Webster, C., Cabane, M., Coll, P., Conrad, P., et al. (2013). Volatile, isotope, and organic analysis of martian fines with the Mars Curiosity rover. *science* 341, 1238937. doi:10.1126/science.1238937

Lewis, J., Eigenbrode, J., Wong, G., Mcadam, A., Archer, P., Sutter, B., et al. (2021). Pyrolysis of oxalate, acetate, and perchlorate mixtures and the implications for organic salts on Mars. *J. Geophys. Res. Planets* 126, e2020JE006803. doi:10.1029/2020je006803

Li, X., Danell, R. M., Brinckerhoff, W. B., Pinnick, V. T., Van Amerom, F., Arevalo, R. D., Jr, et al. (2015). Detection of trace organics in Mars analog samples containing perchlorate by laser desorption/ionization mass spectrometry. *Astrobiology* 15, 104–110. doi:10.1089/ast.2014.1203

Maguire, W. C. (1977). Martian isotopic ratios and upper limits for possible minor constituents as derived from Mariner 9 infrared spectrometer data. *Icarus* 32, 85–97. doi:10.1016/0019-1035(77)90051-3

Mahaffy, P. (2008). Exploration of the habitability of Mars: Development of analytical protocols for measurement of organic carbon on the 2009 Mars Science Laboratory. *Space Sci. Rev.* 135, 255–268. doi:10.1007/s11214-007-9223-1

Mahaffy, P. R., Webster, C. R., Cabane, M., Conrad, P. G., Coll, P., Atreya, S. K., et al. (2012). The sample analysis at Mars investigation and instrument suite. *Space Sci. Rev.* 170, 401–478. doi:10.1007/s11214-012-9879-z

Malin, M. C. (1976). Age of martian channels. *J. Geophys. Res.* 81, 4825–4845. doi:10.1029/jb081i026p04825

Mangold, N., Poulet, F., Mustard, J., Bibring, J. P., Gondet, B., Langevin, Y., et al. (2007). Mineralogy of the Nili Fossae region with OMEGA/mars express data: 2. Aqueous alteration of the crust. *J. Geophys. Res. Planets* 112. doi:10.1029/2006je002835

Marais, D. D., and Walter, M. (1999). Astrobiology: Exploring the origins, evolution, and distribution of life in the universe. *Annu. Rev. Ecol. Syst.* 30, 397–420. doi:10.1146/annurev.ecolsys.30.1.397

Masursky, H., Boyce, J., Dial, A., Schaber, G., and Strobell, M. (1977). Classification and time of formation of Martian channels based on Viking data. *J. Geophys. Res.* 82, 4016–4038. doi:10.1029/js082i028p04016

Mazur, P., Barghoorn, E. S., Halvorson, H. O., Jukes, T. H., Kaplan, I. R., and Margulis, L. (1978). Biological implications of the viking mission to mars. *Space Sci. Rev.* 22, 3–34. doi:10.1007/bf00215812

Mccollom, T. M. (2003). Formation of meteorite hydrocarbons from thermal decomposition of siderite (FeCO3). *Geochimica Cosmochimica Acta* 67, 311–317. doi:10.1016/s0016-7037(02)00945-6

Mckay, D. S., Gibson, E. K., Jr, Thomas-Keprta, K. L., Vali, H., Romanek, C. S., Clemett, S. J., et al. (1996). Search for past life on mars: Possible relic biogenic activity in martian meteorite ALH84001. *Science* 273, 924–930. doi:10.1126/science.273.5277.924

Mcsween, H. Y., Jr (1994). What we have learned about Mars from SNC meteorites. *Meteoritics* 29, 757–779. doi:10.1111/j.1945-5100.1994.tb01092.x

Michalski, J. R., Cuadros, J., Bishop, J. L., Dyar, M. D., Dekov, V., and Fiore, S. (2015). Constraints on the crystal-chemistry of Fe/Mg-rich smectitic clays on Mars and links to global alteration trends. *Earth Planet. Sci. Lett.* 427, 215–225. doi:10.1016/j.epsl.2015.06.020

Michalski, J. R., and Niles, P. B. (2010). Deep crustal carbonate rocks exposed by meteor impact on Mars. *Nat. Geosci.* 3, 751–755. doi:10.1038/ngeo971

Millan, M., Teinturier, S., Malespin, C., Bonnet, J., Buch, A., Dworkin, J., et al. (2022a). Organic molecules revealed in Mars's Bagnold Dunes by Curiosity's derivatization experiment. *Nat. Astron.* 6, 129–140. doi:10.1038/s41550-021-01507-9

Millan, M., Williams, A. J., Mcadam, A. C., Eigenbrode, J. L., Steele, A., Freissinet, C., et al. (2022b). Sedimentary organics in glen Torridon, Gale Crater, mars: Results from the SAM instrument suite and supporting laboratory analyses. *J. Geophys. Res. Planets* 127, e2021JE007107. doi:10.1029/2021je007107

Ming, D., Morris, R., Niles, B., Lauer, H., Archer, P., Sutter, B., et al. (2009). "Combustion of organic molecules by the thermal decomposition of perchlorate salts: Implications for organics at the Mars Phoenix Scout Landing Site," in *40th lunar and planetary science conference)*.

Ming, D. W., Archer, P., Jr, Glavin, D., Eigenbrode, J., Franz, H., Sutter, B., et al. (2014). Volatile and organic compositions of sedimentary rocks in Yellowknife Bay, Gale Crater, mars. *science* 343, 1245267. doi:10.1126/science.1245267

Molnar-Perl, I., and Katona, Z. F. (2000). GC-MS of amino acids as their trimethylsilyl/t-butyldimethylsilyl derivatives: In model solutions III. *Chromatographia* 51, S228–S236. doi:10.1007/bf02492811

Morris, R. V., Ruff, S. W., Gellert, R., Ming, D. W., Arvidson, R. E., Clark, B. C., et al. (2010). Identification of carbonate-rich outcrops on Mars by the Spirit rover. *Science* 329, 421–424. doi:10.1126/science.1189667

Mumma, M. J., Villanueva, G. L., Novak, R. E., Hewagama, T., Bonev, B. P., Disanti, M. A., et al. (2009). Strong release of methane on Mars in northern summer 2003. *Science* 323, 1041–1045. doi:10.1126/science.1165243

Nasa (2012). Curiosity rover mission overview. [Online]. Available at: https://mars.nasa.gov/msl/mission/overview/(Accessed November, 2022).

Nasa (2020). Mars 2020 mission perseverance rover. [Online]. Available at: https://mars.nasa.gov/mars2020/spacecraft/instruments/(Accessed November, 2022).

Nasa (2008). Phoenix lander on mars. [Online]. Available at: https://mars.nasa.gov/resources/1929/phoenix-lander-on-mars/.[Accessed 2022.

Nasa (2010). Viking lander model. [Online]. Available at: https://mars.nasa.gov/resources/3355/viking-lander-model/(Accessed November, 2022).

Navarro-González, R., Vargas, E., De La Rosa, J., Raga, A. C., and Mckay, C. P. (2010). Reanalysis of the Viking results suggests perchlorate and organics at midlatitudes on Mars. *J. Geophys. Res. Planets* 115, E12010. doi:10.1029/2010je003599

Nisbet, E., and Sleep, N. (2001). The habitat and nature of early life. *Nature* 409, 1083–1091. doi:10.1038/35059210

Noe Dobrea, E., Wray, J., Calef, F., Iii, Parker, T., and Murchie, S. (2012). Hydrated minerals on Endeavour Crater's rim and interior, and surrounding plains: New insights from CRISM data. *Geophys. Res. Lett.* 39. doi:10.1029/2012gl053180

Oparin, A. (1965). The origin of life and the origin of enzymes. *Adv. Enzymol. Relat. Areas Mol. Biol.* 27, 347–380. doi:10.1002/9780470122723.ch7

Pavlov, A. A., Mclain, H. L., Glavin, D. P., Roussel, A., Dwork2in, J. P., Elsila, J. E., et al. (2022). Rapid radiolytic degradation of amino acids in the martian shallow subsurface: Implications for the search for extinct life. *Astrobiology* 22, 1099–1115. doi:10.1089/ast.2021.0166

Pavlov, A. K., Blinov, A. V., and Konstantinov, A. N. (2002). Sterilization of Martian surface by cosmic radiation. *Planet. Space Sci.* 50, 669–673. doi:10.1016/s0032-0633(01)00113-1

Pavlov, A., Vasilyev, G., Ostryakov, V., Pavlov, A., and Mahaffy, P. (2012). Degradation of the organic molecules in the shallow subsurface of Mars due to irradiation by cosmic rays. *Geophys. Res. Lett.* 39. doi:10.1029/2012gl052166

Peltzer, E., Bada, J., Schlesinger, G., and Miller, S. (1984). The chemical conditions on the parent body of the Murchison meteorite: Some conclusions based on amino, hydroxy and dicarboxylic acids. *Adv. Space Res.* 4, 69–74. doi:10.1016/0273-1177(84)90546-5

Poch, O., Jaber, M., Stalport, F., Nowak, S., Georgelin, T., Lambert, J.-F., et al. (2015). Effect of nontronite smectite clay on the chemical evolution of several organic molecules under simulated martian surface ultraviolet radiation conditions. *Astrobiology* 15, 221–237. doi:10.1089/ast.2014.1230

Scheller, E. L., Hollis, J. R., Cardarelli, E. L., Steele, A., Beegle, L. W., Bhartia, R., et al. (2022). Aqueous alteration processes in Jezero crater, Mars– implications for organic geochemistry. *Science* 378, 1105–1110. doi:10.1126/science.abo5204

Schuerger, A. C., Moores, J. E., Clausen, C. A., Barlow, N. G., and Britt, D. T. (2012). Methane from UV-irradiated carbonaceous chondrites under simulated Martian conditions. *J. Geophys. Res. Planets* 117. doi:10.1029/2011je004023

Schulze-Makuch, D., Irwin, L. N., and Guan, H. (2002). Search parameters for the remote detection of extraterrestrial life. *Planet. Space Sci.* 50, 675–683. doi:10.1016/s0032-0633(01)00121-0

Schummer, C., Delhomme, O., Appenzeller, B. M., Wennig, R., and Millet, M. (2009). Comparison of MTBSTFA and BSTFA in derivatization reactions of polar compounds prior to GC/MS analysis. *Talanta* 77, 1473–1482. doi:10.1016/j.talanta.2008.09.043

Sephton, M. A., and Hazen, R. M. (2013). On the origins of deep hydrocarbons. *Rev. Mineralogy Geochem.* 75, 449–465. doi:10.2138/rmg.2013.75.14

Sephton, M., Wright, I., Gilmour, I., De Leeuw, J., Grady, M., and Pillinger, C. (2002). High molecular weight organic matter in martian meteorites. *Planet. Space Sci.* 50, 711–716. doi:10.1016/s0032-0633(02)00053-3

Smith, P. H., Tamppari, L., Arvidson, R., Bass, D., Blaney, D., Boynton, W. V., et al. (2009). H2O at the Phoenix landing site. *Science* 325, 58–61. doi:10.1126/science.1172339

Steele, A., Fries, M., Amundsen, H., Mysen, B., Fogel, M., Schweizer, M., et al. (2007). Comprehensive imaging and Raman spectroscopy of carbonate globules from Martian meteorite ALH 84001 and a terrestrial analogue from Svalbard. *Meteorit. Planet. Sci.* 42, 1549–1566. doi:10.1111/j.1945-5100.2007.tb00590.x

Steele, A., Mccubbin, F., Fries, M., Kater, L., Boctor, N., Fogel, M., et al. (2012). A reduced organic carbon component in martian basalts. *Science* 337, 212–215. doi:10.1126/science.1220715

Steele, A., Mccubbin, F. M., and Fries, M. D. (2016). The provenance, formation, and implications of reduced carbon phases in Martian meteorites. *Meteorit. Planet. Sci.* 51, 2203–2225. doi:10.1111/maps.12670

Stern, J. C., Malespin, C. A., Eigenbrode, J. L., Webster, C. R., Flesch, G., Franz, H. B., et al. (2022). Organic carbon concentrations in 3.5-billion-year-old lacustrine mudstones of Mars. *Proc. Natl. Acad. Sci.* 119, e2201139119. doi:10.1073/pnas.2201139119

Stern, R., and Jedrzejas, M. J. (2008). Carbohydrate polymers at the center of life's origins: The importance of molecular processivity. *Chem. Rev.* 108, 5061–5085. doi:10.1021/cr078240l

Szopa, C., Freissinet, C., Glavin, D. P., Millan, M., Buch, A., Franz, H. B., et al. (2020). First detections of dichlorobenzene isomers and trichloromethylpropane from organic matter indigenous to mars mudstone in Gale Crater, mars: Results from the sample analysis at mars instrument onboard the curiosity rover. *Astrobiology* 20, 292–306. doi:10.1089/ast.2018.1908

Townsend, J., Klein, D., and Brooks, S. (2022). "Robotics verification and validation strategies for perseverance rover sampling and caching," in *IEEE aerospace conference (AERO)* (IEEE), 1–13.

Vaniman, D., Bish, D., Ming, D., Bristow, T., Morris, R., Blake, D., et al. (2014). Mineralogy of a mudstone at Yellowknife Bay, Gale Crater, mars. *science* 343, 1243480. doi:10.1126/science.1243480

Villanueva, G., Mumma, M., Novak, R., Radeva, Y., Käufl, H., Smette, A., et al. (2013). A sensitive search for organics (CH4, CH3OH, H2CO, C2H6, C2H2, C2H4), hydroperoxyl (HO2), nitrogen compounds (N2O, NH3, HCN) and chlorine species (HCl, CH3Cl) on Mars using ground-based high-resolution infrared spectroscopy. *Icarus* 223, 11–27. doi:10.1016/j.icarus.2012.11.013

Webster, C. R., Mahaffy, P. R., Atreya, S. K., Moores, J. E., Flesch, G. J., Malespin, C., et al. (2018). Background levels of methane in Mars' atmosphere show strong seasonal variations. *Science* 360, 1093–1096. doi:10.1126/science.aaq0131

Wilson, M., Jones, D., and Russell, J. (1980). Glushinskite, a naturally occurring magnesium oxalate. *Mineral. Mag.* 43, 837–840. doi:10.1180/minmag.1980.043.331.02

Yen, A., Kim, S., Hecht, M., Frant, M., and Murray, B. (2000). Evidence that the reactivity of the martian soil is due to superoxide ions. *Science* 289, 1909–1912. doi:10.1126/science.289.5486.1909

Yen, A. S., Murray, B., Rossman, G. R., and Grunthaner, F. J. (1999). Stability of hydroxylated minerals on mars: A study on the effects of exposure to ultraviolet radiation. *J. Geophys. Res. Planets* 104, 27031–27041. doi:10.1029/1999je001065

Zahnle, K., Freedman, R. S., and Catling, D. C. (2011). Is there methane on Mars? *Icarus* 212, 493–503. doi:10.1016/j.icarus.2010.11.027

Zent, A. P., and Mckay, C. P. (1994). The chemical reactivity of the martian soil and implications for future missions. *Icarus* 108, 146–157. doi:10.1006/icar.1994.1047

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Mini Review
PUBLISHED 13 February 2023
DOI 10.3389/fspas.2023.1107371

Check for updates

OPEN ACCESS

EDITED BY
Alberto Fairén,
Center for Astrobiology (CSIC), Spain

REVIEWED BY
Lars Behrendt,
Uppsala University, Sweden
Nicoletta La Rocca,
University of Padua, Italy

*CORRESPONDENCE
Patrick Jung,
patrick_jung90@web.de

SPECIALTY SECTION
This article was submitted to Astrobiology, a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 24 November 2022
ACCEPTED 05 January 2023
PUBLISHED 13 February 2023

CITATION
Jung P, Harion F, Wu S, Nürnberg DJ, Bellamoli F, Guillen A, Leira M and Lakatos M (2023), Dark blue-green: Cave-inhabiting cyanobacteria as a model for astrobiology.
Front. Astron. Space Sci. 10:1107371.
doi: 10.3389/fspas.2023.1107371

COPYRIGHT
© 2023 Jung, Harion, Wu, Nürnberg, Bellamoli, Guillen, Leira and Lakatos. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Dark blue-green: Cave-inhabiting cyanobacteria as a model for astrobiology

Patrick Jung[1]*, Felix Harion[1], Shujie Wu[2], Dennis J. Nürnberg[2,3], Francesco Bellamoli[4], Antonio Guillen[5], Manuel Leira[6] and Michael Lakatos[1]

[1]Integrative Biotechnology, University of Applied Sciences Kaiserslautern, Kaiserslautern, Germany, [2]Institute of Experimental Physics, Freie Universität Berlin, Berlin, Germany, [3]Dahlem Centre for Plant Sciences, Freie Universität Berlin, Berlin, Germany, [4]Department of Biotechnology, University of Verona, Verona, Italy, [5]Scientific coordinator of Astroland Agency, Logroño, Spain, [6]Parque Científico y Tecnológico de Cantabria, Santander, Spain

Subterranean environments on Earth serve as an analog for the study of microbes on other planets, which has become an active area of research. Although it might sound contradictory that photosynthetic cyanobacteria thrive in extreme low light environments, they are frequent inhabitants of caves on Earth. Throughout the phylum these cyanobacteria have developed unique adaptations that cannot only be used for biotechnological processes but also have implications for astrobiology. They can, for example, both accommodate for the low light conditions by producing specific pigments that allow photosynthesis in near-infrared (IR) radiation/far-red light, and they can synthesize bioplastic compounds and calcium carbonate sheaths which represent valuable resources during human colonization of other planets or rock bodies. This article will highlight the potential benefits of cave-inhabiting cyanobacteria and will present a suitable bioreactor technique for the utilization of these special microbes during future space missions.

KEYWORDS

emerse photobioreactor, polyhydroxybutyrate, biofilm bioreactor, far-red photosynthesis, calcification

1 Introduction

For human colonies to be installed on currently unhabitable planets preparation of the area of interest is crucial (Levchenko et al., 2021). Creating such an enclosed space where humans can operate under preferred conditions -also known as the biosphere- is likely not achievable on large scales, but will need to be achieved gradually. As an initial step, this includes a breathable atmosphere with a constant supply of oxygen, a moderate temperature range and protection from cosmic rays. At a later stage, resources are needed for the construction, maintenance of the colonies and food production. All of this must include a sustainable, circular low-waste concept, most preferably based on resources from the planet of interest (Ceylan, 2018; Santo, 2022). In the past, the concept of natural caves has been used as a test scenario for colonizing extra-terrestrial environments (Walden et al., 1998). Caves on Mars, for example, would not only be shielded from damaging ionizing and UV radiation, but also offer protection against small-scale meteorite impacts (e. g., Walden et al., 1998), and extreme weather conditions such as the dust storms that typically affect the Martian surface (Ryan and Sharman, 1981; Viúdez-Moreiras et al., 2019). These subterranean environments could protect astronauts during early human missions to other planets and could act as primary producers creating hospitable biospheres.

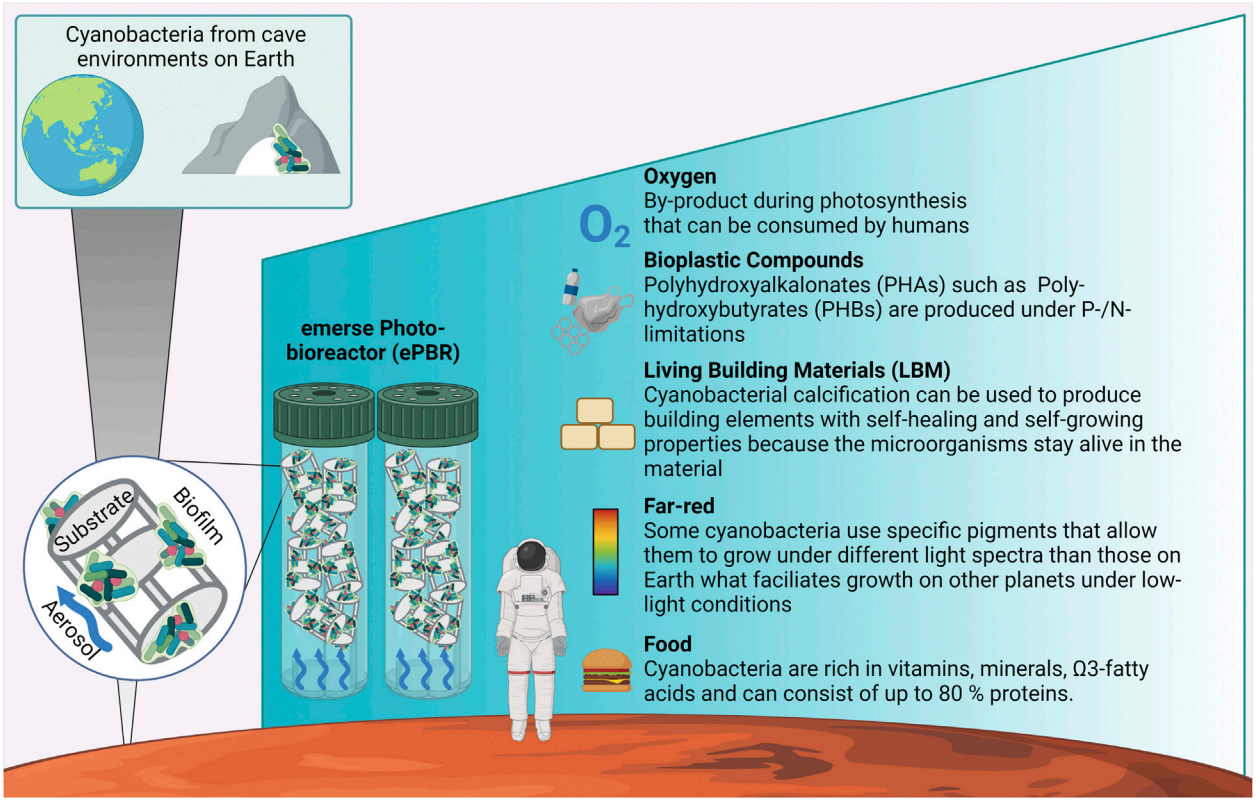


FIGURE 1
Potential of cave-inhabiting cyanobacteria as a model for astrobiology. *Cyanobacteria* from caves on Earth possess unique properties that make them great candidates for biotechnological applications during human colonization of other planets or the Earth's Moon. Using minimum resources they can be grown in emerse photo-bioreactors (ePBR) that produce an aerosol directed over surfaces of biocarriers on which the cyanobacterial biofilm establishes. This technique allows the manufacturing of important and yet in this context ignored products such as living building materials or bioplastic compounds.

While Mars contains large and interconnected cave systems across the planet (Cushing et al., 2007; Sauro et al., 2020), the Earth's Moon is well known for its craters and lunar tubes (Chappaz et al., 2017; Kaku et al., 2017). Such pit craters, formed from the collapsed ceilings of subsurface void spaces such as natural caves or lava tubes, have been detected on almost every rock body within the inner Solar System (Wyrick et al., 2004; Haruyama et al., 2009; Davey et al., 2013; Titus et al., 2021): so far, approximately 2,660 subsurface access points have been cataloged on Solar System bodies (Titus et al., 2021), and since the discovery of the first caves on Mars, the spacecraft NASA's robotic Mars Odyssey Orbiter has helped identify over 1,000 probable gaps on the Red Planet that have been compiled into the Mars Global Cave Candidate Catalog, also known as the MGC (Cushing, 2017). Recently it has been shown that some of the lunar pits and caves provide a temperature around 17°C with a stable and safe thermal environment for long-term exploration and habitation of the Moon (Horvath et al., 2022; Rodeghiero et al., 2022). In addition, the ultraviolet radiation environment and shielding in pit craters and cave skylights on Mars can now be reliably modeled and demonstrated the protective character of such subterranean environments (Viúdez-Moreiras et al., 2021). Studying caves on Earth could thus support the prediction of conditions in cavities on other planets or rock bodies (Sauro et al., 2019; Sauro et al., 2020).

Lava tunnels and cave-like structures have been observed with the use of Mars satellite imaging and special measuring tools. They have been generally considered as suitable environments to search for life and biomarkers (Léveillé and Datta, 2010). However, photosynthetic bacteria such as cyanobacteria are one of the most promising cave-inhabitants in this context. They have evolved on Earth 3.5 mya (Sanchez-Baracaldo and Cardona, 2020; Hickman-Lewis et al., 2022) and are adapted to a wide range of environments, ranging from both hot and cold deserts (Friedmann, 1980; Wynn-Williams, 2000; Lacap-Bugler et al., 2017) with strong radiation (Rastogi et al., 2014) to high salinity habitats (Oren, 2015) and symbioses with lichens (Spribille et al., 2022) or plants (Rai et al., 2000). In addition, they have been proposed as suitable, multi-purpose microorganisms that are supportive during human missions to other planets including the selection of model strains (e.g., Ramalho et al., 2022). Researchers are able to not only isolate and grow them under artificial conditions, but also use them for large-scale industrial production of natural goods (Mutale-Joan et al., 2022). They are regularly found in high abundances and diversity in caves around the globe (Miscoe et al., 2016; Behrendt et al., 2020; Popović et al., 2020), possibly due to the relative constant conditions found in these habitats. In the following we review the suitability of cave-inhabiting cyanobacteria for colonization of other planets with a special emphasis on biotechnological aspects as demonstrated in Figures 1, 2.

2 Current challenges and new solutions for the efficient cultivation of cyanobacteria

Before devising any new biotechnological application where cyanobacteria can be utilized -irrespectively whether on Earth or

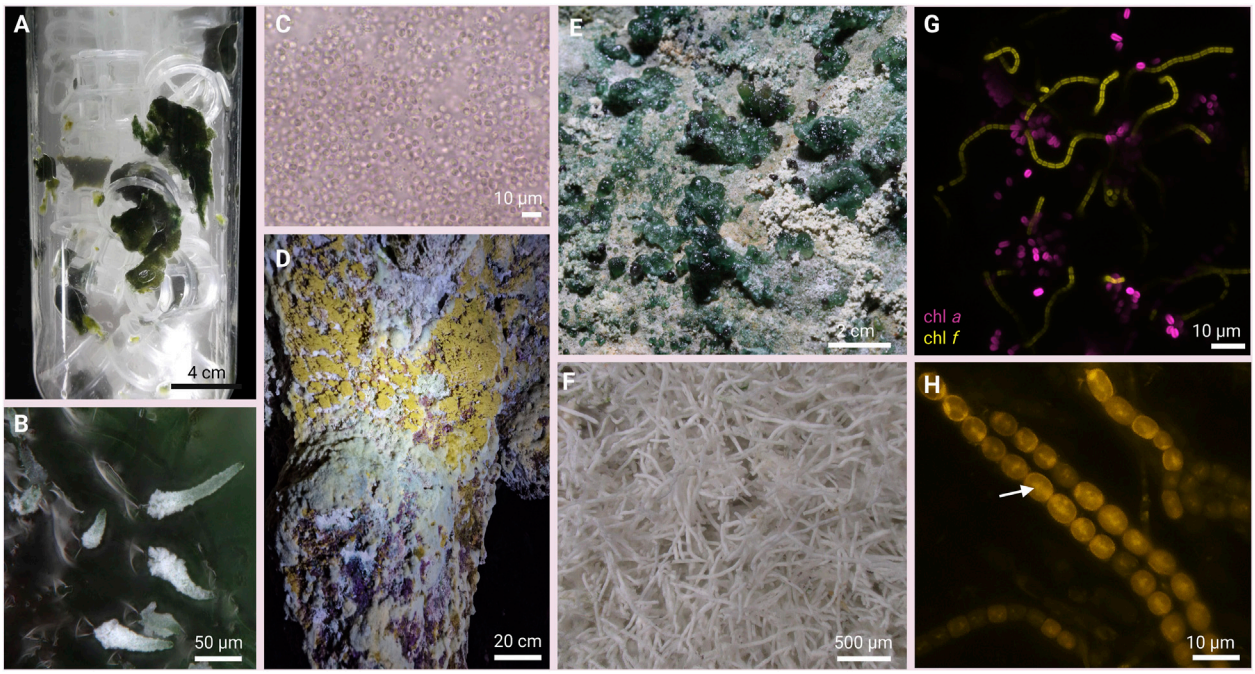


FIGURE 2
Cyanobacteria from the cave environment. **(A)** emerse photobioreactor (ePBR) filled with biocarriers as substrate on which cyanobacteria grow. The reactor is aerated with aerosol as the medium. **(B)** isolated *Scytonema* sp. with calcified tip filaments under laboratory conditions. **(C)** microscopy of *Gloeobacter* sp. **(D, E)** natural cyanobacteria dominated biofilms from caves. **(F)** calcified natural sheaths of *Geitleria* sp. **(G)** confocal laser scanning microscopy (CLSM) showing cyanobacteria with emission wavelengths characteristic for chl *d*/ *f* (yellow; 725–750 nm) and chl *a* and phycobilisomes (magenta; 650–700 nm). **(H)** Nile red staining and fluorescence of *Nostoc* sp. Isolated from the cave biofilms showing PHB-rich granules (arrow).

elsewhere- one needs to consider how cyanobacteria accumulate biomass and how these cultivation techniques need to be adapted. One of the major setbacks for cyanobacterial applications in industry as of today is the high amount of water used for submersed cultivations. During this process cyanobacterial biomass is, for example, cultivated in large open ponds filled with hundreds of liters of medium and stirred for several weeks under natural light conditions. To harvest biomass gained under these conditions, cells must be separated from the medium, resulting in poor biomass to water yields, usually of about <1% (w/v), and great amounts of wastewater that can hardly be recycled due to contaminations (Johnson et al., 2018). Moreover, most strains that are currently used in commercial approaches are aquatic cyanobacterial strains with a low acclimatization potential (e.g., artificial light and temperature regime in bioreactors) and from which only a few products can be derived, such as food supplements or pigments (Khalifa et al., 2021). However, there are also photobioreactors that were specifically designed to grow cyanobacteria as a life-support system on Mars (Verseux et al., 2021), but also these systems are based on submersed systems and were mostly tested for aquatic cyanobacterial strains.

A new biotechnological approach to cyanobacterial farming is the use of emerse -air exposed- biofilm photobioreactors (ePBRs; Figures 1, 2A), where only a little amount of aqueous medium is needed for aerosolization over cyanobacterial biofilms that grow immobilized on surfaces (Lakatos and Strieth, 2017; Scherer et al., 2022). This method is resource-friendly because the wetting of biofilms only with aerosols saves approximately 90% of water investment compared to submerse cultivation methods. Moreover, the energy consumption for dewatering in downstream processing can be up to 82 MJ per kg dry biomass (Chisti 2007; Jorquera et al., 2010; Ozkan et al., 2012). Consequently, the production of 1 kg biomass requires an energy input of 385 MJ kg^{-1} in a commercial tubular system and 9.18 MJ kg-1 in a raceway pond—both including mixing during cultivation (Chisti, 2007; Jorquera et al., 2010) while biofilm reactors only consume 4.7 MJ kg^{-1} because mixing and dewatering are no crucial process steps (Ozkan et al., 2012). In addition, in biofilm photobioreactors the cell density is up to 90 times higher in comparison to submerge technologies like open ponds or closed systems (Jorquera et al., 2010; Ozkan et al., 2012; Podola et al., 2017). Especially the ePBR-technology is suited for terrestrial cyanobacteria showing a much broader ecological spectrum in terms of light quantity and quality, pH, desiccation, temperature and other abiotic factors (Kuhne et al., 2014; Lakatos and Strieth 2017). Terrestrial cyanobacteria from e.g. biocrusts of arid deserts experience, for example, great temperature amplitudes as the soil heats up during the day and cools down at night and they need to be capable of using also short hydration events provided by, e.g., fog (Jung et al., 2019) while aquatic cyanobacteria constantly have a water supply and the water column has a constant temperature. Terrestrial cyanobacteria from caves could potentially not only photo-acclimatize towards the shifted light spectrum that we expect from other planets, but they could also be key to new products and materials created with the ePBR technology. Among these are for example induced calcification processes that result in so-called living building materials (LBMs) or bioplastic compounds among others (Figure 1).

3 Ecology and benefits of cave-inhabiting cyanobacteria for future applications

Cyanobacteria are the oldest living microorganisms capable of oxygenic photosynthesis on Earth and as a result, they are highly adapted to a wide range of environments (Chen et al., 2021). They are known to be pioneer organisms, capable of conquering new habitats and laying the foundation for whole ecosystems as primary producers and due to their ability of priming soils (Muñoz-Rojas et al., 2018; Chen et al., 2021). The Great Oxygenation Event during the Paleoproterozoic era demonstrated what photosynthetic microorganisms are capable of as this was the time interval when the Earth's atmosphere and the shallow ocean first experienced a rise in the amount of oxygen (Knoll and Nowak, 2017). This injection of toxic oxygen into an anaerobic biosphere probably caused the extinction of many existing anaerobic species present on Earth at that time (Hodgskiss et al., 2019).

It has been predicted that extra-terrestrial photosynthesis might be possible where an Earth-like atmosphere is present and water can be supplied (Lingam and Loeb, 2020). On planetary surfaces, several factors could impair cyanobacterial colonization (Verseux et al., 2016), although some highly resistant strains have already been discovered (Billi et al., 2022). Especially cyanobacteria from extreme terrestrial habitats on Earth have already been shown to be suitable candidates to not only survive travelling through Space (Napoli et al., 2022), they could also grow under Space- and Mars-like conditions (de Vera et al., 2019), and on Martian regolith (Baqué et al., 2013). In particular, *Chroococcidiopsis* sp. CCMEE 029 from the Negev desert has been extensively studied and reported to withstand years of desiccation and the exposure to space and Mars-like stimulations (Billi, 2009; Billi et al., 2019). Their ability to withstand such extremes is due to a great amplitude of mechanisms such as the formation of extra cellular polymeric substances (EPS) that provide physical protection and prevent desiccation (Rossi and De Philippis, 2015) and also a variety of pigments that increases their resistance against UV irradiation (Sinha and Häder, 2008).

In contrast, caves represent a protected habitat with reduced abiotic stressors such as high irradiance, UV/gamma radiation, desiccation, different contaminants and fast temperature variations, at the cost of limited light availability (Billi et al., 2022). On Earth, caves show high (cyano)bacterial diversity (Hauer et al., 2015; Popović et al., 2020; Prescott et al., 2022) and could represent suitable growth environments on other planetary bodies featuring relatable conditions (Wynne et al., 2021). These conditions are represented by a stable temperature regime, rough surface textures and constant as well as continuous humidity supplies provided by runoff- or seepage water (Figures 2D, E). Karst areas are especially prone to form caves across the globe since the main component of limestone is calcium carbonate that easily gets dissolved by water. Such cave systems can be colonized by naturally occurring phototrophic microbial communities (Roldán and Hernández-Mariné, 2009) or their establishment can be induced by artificial lightning in touristic show caves, also known as Lampenflora (Mulec, 2019).

A common unicellular species of cyanobacteria in natural caves is *Gloeobacter violaceus* (Mareš et al., 2013; Figure 2C), representing one of the oldest lineages of modern cyanobacteria. The genus *Gloeobacter* is characterized by the lack of thylakoid membranes and the positioning of the photosynthetic machinery in the cytoplasmic membrane which makes it an interesting model for biotechnology and research on the origin of photosynthesis (Engqvist et al., 2015).

Other low-light adapted cyanobacteria from cave environments possess specialized chlorophylls, such as chlorophyll *d* and chlorophyll *f*, capable of absorbing far-red light in addition to visible light. The presence of far-red cyanobacteria as a consequence of the spectral shift towards longer wavelengths has been recently reported for cave systems (Behrendt et al., 2020). The ability to produce such pigments could allow the growth of cyanobacteria on planets and rock bodies exposed to different light spectra than Earth (Billi et al., 2022).

In addition, many filamentous cyanobacteria can be found in caves, some of which are well known to fix atmospheric nitrogen and to form calcium carbonate sheaths around their cells. Among the most prominent examples for such calcifying cyanobacteria are *Geitleria* spp. (Kilgore et al., 2018; Figure 2F) and *Scytonema julianum* (Couté and Bury, 1988; Figure 2B). Especially cells of the latter are encased with thick, well-developed calcified sheaths with external diameters of 11–25 µm, which develop through the sequential precipitation of amorphous $CaCO_3$, acicular calcite crystals, triradiate calcite crystals, and dendrite calcite or aragonite crystals (Figure 2F; Jones and Peng, 2014) with a high degree of mineral variability (Hoffmann, 1992). Employing such cyanobacteria in the biomineralization of carbon dioxide by calcium carbonate precipitation has been suggested to offer self-sustaining strategies for point-source carbon capture and sequestration in biotechnological processes (Jansson and Northen, 2010; Phillips et al., 2013).

Interestingly, a high proportion of cave inhabiting cyanobacterial taxa have recently been shown to produce biocompatible and biodegradable bioplastics e.g., polyhydroxyalkanoates such as polyhydroxybutyrate (PHA / PHB; Figure 2H) with various application potential (Djebaili et al., 2022). *Cyanobacteria* have minimal nutrient requirements for growth and accumulate PHAs and PHBs through oxygenic photosynthesis under nutrient-limited conditions, such as nitrogen and/or phosphate starvation, where cells enter a quiescent state known as chlorosis (Koch et al., 2020a).

Next to the production of oxygen, dietary supplement, dyes, and food, the mentioned aspects such as far-red adaptations, PHB production, calcification, and biotechnological advantages make cave-inhabiting cyanobacteria suitable and multifunctional candidates to study the requirements for life on other planets.

4 Far-red light photosynthetic adaptations in *Cyanobacteria*

In limestone caves light forms a spectral gradient where, from the entrance to greater depths of the cavity, the visible component gradually becomes limiting while the far-red (FR, above 700 nm) portion of the spectrum is enriched (Behrendt et al., 2020). This is comparable to the Earth's Moon, where features such as lava tube-related sinkhole chains present temperatures and radiance levels (Horvath et al., 2022; Rodeghiero et al., 2022). That could be compatible with cyanobacterial life. Due to the fact, that materials such as regolith and basalts show an increased reflectivity in the red/IR region (Anbazhagan and Arivazhagan, 2009; Gaier et al., 2012; Xu et al., 2021), a FR enrichment analogous to Earth caves could be expected. Similar features (Sauro et al., 2020) and materials (Mandon et al., 2022) can be found on Mars, whose caves might also present a

more hospitable environment than the surface (Léveillé and Datta, 2010; Irwin and Schulze-Makuch, 2020). Further away from the Solar System, for habitable zone planetary conditions, the light spectrum emitted by M-stars (red dwarfs), albeit red-shifted compared to the Sun, could sustain oxygenic photosynthesis in cyanobacteria (Claudi et al., 2020), as could K-stars (Covone et al., 2021).

Cyanobacteria can modulate their photosynthetic activity according to light spectral distribution *via* several different processes designated chromatic acclimation (CA) (Sanfilippo et al., 2019). Some species, in response to FR light enriched exposure, show a CA called Far-Red Light Photoacclimation (FaRLiP). This process induces profound changes in the composition of photosystems, antennae and chlorophyll content, by synthesizing chl *f* and chl *d* (Figure 2G) in addition to chl *a* (Gan et al., 2014). Even if in FR acclimated cells chl *f* constitutes only about 12% of the total chlorophyll content, it is not only an accessory pigment, but it is instead present in both photosystems and it is directly involved in charge separation (Nürnberg et al., 2018). In contrast, in most species of the genus *Acaryochloris* where chl *d* represents the primary photopigment, photosynthetic use of FR light is a constitutive and more restrictive adaptation than FaRLiP (Wolf and Blankenship, 2019; Viola et al., 2022).

FR light adaptations allow cyanobacterial growth in widespread ecological niches, including caves, lakes, beach rock, deserts, hot spring mats, and subtropical forests (Antonaru et al., 2020). Cave light gradients can be considered, on a much bigger scale, comparable to vertical micro-gradients observed in biofilms, desert crusts or endolithic habitats, where the upper layers absorb the visible component, while FR light penetrates into deeper layers. Both cave- and micro gradient-induced FR light enrichment lead to a substantial increase in the growth of FR-adapted cyanobacteria (Behrendt et al., 2012; Behrendt et al., 2020; Kühl et al., 2020).

Far-red adaptations could then give cyanobacteria a crucial advantage in space biology applications, on the surface but especially in caves. For instance, direct light emitted by M-stars allows the growth of FaRLiP and chl *d* dependent cyanobacteria with little advantage over non-FR adapted strains (Ritchie et al., 2018; Claudi et al., 2020). In fact, while in those conditions *Acaryochloris* potential productivity estimate has been considered the highest among other commonly (or widely) used oxygenic phototrophs (e.g., 0.178 vs. 0.155 g C m^{-2} h^{-1} in *Synechococcus* sp. PCC 7942; Ritchie et al., 2018), a major drawback of FaRLiP strains is that they show reduced growth rates if they are cultivated exclusively under far-red light (Claudi et al., 2020). For *Leptolyngbya* sp. JSC-1, for example, it was shown that its doubling time is four times lower than in white (visible) light (Nien et al., 2022). Another FaRLiP strain, *Halomicronema hongdechloris* utilizes both chl *a* and chl *f* for photosynthesis under FR low light conditions (chl content per cell: 1.54 mg chl g^{-1} cell wet weight), nevertheless, and grows more rapidly than it does under WL (white light) low light conditions (chl content per cell: 1.0 mg chl g^{-1} cell wet weight; Li et al., 2014). Additionally, it has been shown that growth rates of some isolates can reach up to 80% by mimicking the field far-red light intensities of 6–8 µmol photons m^{-2} s^{-1} in comparison to those under the above described white light conditions, suggesting that chl *f* and chl *d* contributed effectively to cyanobacterial growth (Zhang et al., 2019). However, being in the presence of an already red-shifted spectrum, FR light adaptations could be vital for growing in further FR enriched environments as how expoplanet caves could be. Another prospect for space applications could be to genetically engineer non-FR-adapted cyanobacterial strains in order to allow the synthesis of red-shifted chlorophylls and extend the available light spectrum (Luan et al., 2020; Liu et al., 2021). Potential background strains could be chosen for improvements in biomass production (Verseux et al., 2016; Liu et al., 2021; Ramalho et al., 2022), mineral resources extraction (Olsson-Francis and Cockell, 2010), or resistance to extreme conditions (Billi et al., 2022). Preliminary studies to partially incorporate FaRLiP components have already been performed in cyanobacteria (Shen et al., 2019; Trinugroho et al., 2020; Tros et al., 2020), albeit so far, a full integration of the complete gene set has yet to be achieved. Such strains could be promising candidates for the cultivation in ePBRs on other planets as they effectively use the underlying natural light spectrum.

5 Cyanobacterial bioplastic production

Biodegradable and biocompatible materials such as PHAs/PHBs rapidly gained attention over the last decades (Anastas and Kirchhoff, 2002; Lenz and Marchessault, 2005). PHBs are microbially produced polymers (Figure 2H), which act as storage units for both energy and carbon (Williams et al., 1999; Steinbüchel and Lütke-Eversloh, 2003). Microbial PHB production depends on external triggers, such as nitrogen or phosphate starvation as well as an oversupply of carbon sources (Steinbüchel and Lütke-Eversloh, 2003). Industrial bio-based production solely relies on heterotrophic microorganisms, which require a carbon feedstock, mostly polysaccharides (Chavez et al., 2022). This production bears some disadvantages: firstly, the loss of these feedstocks as food-source and secondly, the high costs of the feedstocks.

Material properties of PHBs are comparable to petroleum-based polymers such as polypropylene or polyethylene in terms of elasticity module and tensile strength (Chee et al., 2018). PHBs are biodegradable and can be degraded enzymatically by soil microorganisms, which makes them even more appealing regarding plastic pollution (Madison and Huisman, 1999).

With cyanobacteria being able to produce a broad range of PHBs from CO_2 and light, feedstock costs are low, making production partly attractive (Lee et al., 2021). Setbacks such as longer cultivation times have yet to be overcome (Das and Maiti, 2021; Lee et al., 2021). A case study about caves in Italy showed the capabilities of cave-inhabiting cyanobacteria to produce PHBs (Djebaili et al., 2022). Dark periods are somehow crucial for PHB production (e.g., Ansari and Fatma 2016), but why the ability for PHB biosynthesis in cave cyanobacteria is widespread remains unknown yet.

Compared to petroleum-based manufacturing processes, bio-based production requires fewer facilities: bioreactors with a subsequent extraction step are sufficient, which means that fewer materials would have to be deployed to other planets (Harding et al., 2007). Recent studies focused on blending PHBs with different materials or using chemical engineering to diversify material properties in order to further extend the application range (Chee et al., 2018; Turco et al., 2021). As of today, pilot scale projects are on the rise, using cyanobacteria firstly to reduce the carbon footprint of processes (power plants) and to produce PHBs, as started in 2011 by the Austrian producer of electricity EVN. Commercially available PHBs are mostly produced with bacteria; however, genetic engineering of cyanobacteria also makes them more and more appealing for industrial application (Koch et al., 2020b).

Based on these findings PHBs could be produced extra-terrestrially on-site as a platform compound with different post-productional blendings or chemical alterations to suit various applications in the biomedical, building or packaging sector. If combined with 3D printing technologies, product manufacturing could be fully automated and remotely supervised to install facilities before human colonization of other planets (Frone et al., 2019; Chiulan et al., 2021). The biodegradability of PHBs also possesses one more crucial advantage, the potential to establish a zero-waste circular economy that decreases the resource demands (Talan et al., 2022).

6 Cyanobacterial calcification and applications

Calcification, the process of biomineralization of CO_2 by calcium carbonate ($CaCO_3$) precipitation, is a widely occurring phenomenon in ecosystems and plays a crucial role in the biogeochemical carbon cycle (Gattuso and Buddemeier, 2000; Ridgwell and Zeebe, 2005). The first descriptions of entanglement between microorganisms and calcification were presented many years ago for the process of stromatolite formation (Logan et al., 1964).

Mostly, calcification is known for its role in deep-sea carbon storage, carried out by marine microorganisms, especially cyanobacteria and coccolithophores (Westbroek et al., 1984; Planavsky et al., 2009; Milner et al., 2016; Feng et al., 2017; Kalokora et al., 2020). However, calcification is not exclusively taking place in aquatic environments. It can also be observed in terrestrial habitats e.g. by aerophytic cyanobacteria (Hodač et al., 2015; Kamran et al., 2020). Here, a special emphasis should be given to cave-inhabiting cyanobacteria, since numerous studies found them capable of calcification when associated with a Ca^{2+} rich source, such as in karst caves (Figures 2B, F; Bourrelly and Dupuy, 1973; Arino et al., 1997). One theory states the potential for photon entrapment by distinct crystal structures, such as aragonite, which was shown for stone corals (*Acropora* sp.) and referred to as luminescent light collection (Neumann-Micheau and Tributsch, 2018). This theory could also be extrapolated to cave inhabiting cyanobacteria under low light conditions.

Although the mechanisms of this process remains largely unknown, biotechnological applications for calcification are on the rise. In terms of carbon-capturing processes, microbially produced $CaCO_3$ sheaths represent an inorganic carbon sink (Jansson and Northen, 2010). The usage of cyanobacteria exhibits a double effect on the carbon utilization capacity, because firstly, they are fixing CO_2 producing organic molecules during growth and photosynthesis and secondly capture inorganic carbon (CO_2 / HCO_3^- / CO_3^{2-}) during calcification. Biotechnology has experimented with calcifying microorganisms in the field of sustainable and eco-friendly building materials (De Muynck et al., 2010; Wang et al., 2012). A recent study by Heveran and colleagues presented a proof of concept where calcifying *Synechococcus* sp. PCC 7002 can be used to build actual concrete like-building blocks, which they referred to as living building materials (LBMs) (Williams et al., 2019; Heveran et al., 2020). While the use of cyanobacteria for manufacturing of concrete type building materials may seem far away, first steps have been taken for an industrial approach realization. The company Prometheus Materials (Colorado, United States), for example already developed an algae-based construction material with a low carbon footprint. Also, Minus Materials (Colorado, United States) aims for what they refer to as biogenic limestone to replace concrete as building material. The industrial application in this field rapidly increases leaving room for more exploration, even in Space.

When thinking about installing permanent infrastructure for human residence on another planet / Moon, building materials from local resources is a must-have to decrease the deployment of materials from Earth by expensive space launches. If calcium and CO_2 were present in considerable amounts at the destined exo-planet, one could think about using cyanobacteria for manufacturing LBMs according to the published protocols of Heveran et al. (2020). Regarding the presence of lunar regolith, it might be possible to use these sands in a mixture with grown cyanobacterial biomass. According to NASA's data, lunar regolith is mainly composed of silicon (45.4%) but also bears a considerable amount of calcium (11.8%) (Taylor, 1975). Notably, a study also analyzed the distribution of plagioclase (a calcium-rich feldspar) on the Moon, and showed a high abundance of this material (Ohtake et al., 2009; Baasch et al., 2021). Moon mining approaches, combined with the fabrication of LBMs and 3D printing technologies, could provide a low cost-solution for building infrastructure on Earth's Moon and other planets (Cesaretti et al., 2014; Vítek et al., 2020).

Author contributions

PJ prepared the first draft of the manuscript and the figures. All authors contributed figures, revised and edited the manuscript and approved the final version.

Funding

The project is funded by the German Research Foundation (DFG) with the grant number JU 3228/1-1 to PJ and NU 421/1 to DJN. The joint project "Cyanolessinia" of FB is funded by the University of Verona (Italy) with the grant number JPVR19S9PT. ML is funded by the Ministry of Science and Health Rhineland-Palatinate (PhytoBioTech, 724-0116#2021/004-1501 15,405) and by the Federal Ministry of Education and Research (W2V-Strategy2Value, 03WIR4502A & Technology2Value 03WIR4504B). Moreover, ML and FH are supported by EU-HORIZON (Waste2BioComp ID: 101058654).

Acknowledgments

AG wants to thank Antonio Ordóñez, founder of Virtual Biodiversity (BV). All authors want to thank Astroland Agency and their astrobiology project within the framework of which this work was conducted.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anastas, P. T., and Kirchhoff, M. M. (2002). Origins, current status, and future challenges of green chemistry. *Chem. Res.* 35, 686–694. doi:10.1021/ar010065m

Anbazhagan, S., and Arivazhagan, S. (2009). Reflectance spectra of analog basalts; implications for remote sensing of lunar geology. *Planet. Space Sci.* 5712, 1346–1358. doi:10.1016/j.pss.2009.06.020

Ansari, S., and Fatma, T. (2016). Cyanobacterial polyhydroxybutyrate (PHB): Screening, optimization and characterization. *PloS one* 11 (6), e0158168. doi:10.1371/journal.pone.0158168

Antonaru, L. A., Cardona, T., Larkum, A. W., and Nürnberg, D. J. (2020). Global distribution of a chlorophyll f cyanobacterial marker. *ISME J.* 14 (9), 2275–2287. doi:10.1038/s41396-020-0670-y

Arino, X., Hernandez-Marine, M., and Saiz Jimenez, C. (1997). Colonization of Roman tombs by calcifying cyanobacteria. *Physiologica* 36, 366–373. doi:10.2216/i0031-8884-36-5-366.1

Baasch, J., Windisch, L., Koch, F., Linke, S., Stoll, E., and Schilde, C. (2021). Regolith as substitute mold material for aluminum casting on the Moon. *Acta Astronaut.* 182, 1–12. doi:10.1016/j.actaastro.2021.01.045

Baqué, M., de Vera, J. P., Rettberg, P., and Billi, D. (2013). The BOSS and BIOMEX space experiments on the EXPOSE-R2 mission: Endurance of the desert cyanobacterium Chroococcidiopsis under simulated space vacuum, Martian atmosphere, UVC radiation and temperature extremes. *Acta Astronaut.* 91, 180–186. doi:10.1016/j.actaastro.2013.05.015

Behrendt, L., Larkum, A. W., Trampe, E., Norman, A., Sørensen, S. J., and Kühl, M. (2012). Microbial diversity of biofilm communities in microniches associated with the didemnid ascidian Lissoclinum patella. *ISME J.* 66, 1222–1237. doi:10.1038/ismej.2011.181

Behrendt, L., Trampe, E. L., Nord, N. B., Nguyen, J., Kühl, M., Lonco, D., et al. (2020). Life in the dark: Far-red absorbing cyanobacteria extend photic zones deep into terrestrial caves. *Environ. Microbiol.* 22 (3), 952–963. doi:10.1111/1462-2920.14774

Billi, D., Napoli, A., Mosca, C., Fagliarone, C., De Carolis, R., Balbi, A., et al. (2022). Identification of far-red light acclimation in an endolithic Chroococcidiopsis strain and associated genomic features: Implications for oxygenic photosynthesis on exoplanets. *Front. Microbiol.* 13, 933404. doi:10.3389/fmicb.2022.933404

Billi, D. (2009). Subcellular integrities in Chroococcidiopsis sp. CCMEE 029 survivors after prolonged desiccation revealed by molecular probes and genome stability assays. *Extremophiles* 13, 49–57. doi:10.1007/s00792-008-0196-0

Billi, D., Verseux, C., Fagliarone, C., Napoli, A., Baqué, M., and de Vera, J.-P. (2019). A desert cyanobacterium under simulated mars-like conditions in low earth orbit: Implications for the habitability of Mars. *Astrobiology* 19 (2), 158–169. doi:10.1089/ast.2017.1807

Bourrelly, P., and Dupuy, P. (1973). Quelques stations françaises de Geitleria calcarea, Cyanophycée cavernicole. *Schweiz. z. Hydrol.* 35, 136–140. doi:10.1007/bf02502067

Cesaretti, G., Dini, E., Kestelier, X. D., Colla, V., and Pambaguian, L. (2014). Building components for an outpost on the Lunar soil by means of a novel 3D printing technology. *Acta Astronaut.* 93, 430–450. doi:10.1016/j.actaastro.2013.07.034

Ceylan, S. (2018). Space, architecture, and science fiction: An architectural interpretation of space colonization. *Int. J. Constr. Environ.* 9 (2), 1–17. doi:10.18848/2154-8587/cgp/v09i02/1-17

Chappaz, L., Sood, R., Melosh, H. J., Howell, K. C., Blair, D. M., Milbury, C., et al. (2017). Evidence of large empty lava tubes on the Moon using GRAIL gravity. *Geophys. Res. Lett.* 44 (1), 105–112. doi:10.1002/2016gl071588

Chavez, B. A., Raghavan, V., and Tartakovsky, B. (2022). A comparative analysis of biopolymer production by microbial and bioelectrochemical technologies. *RSC Adv.* 12, 16105–16118. doi:10.1039/d1ra08796g

Chee, J., Yeo, C., Muiruri, J. K., Thitsartarn, W., and Li, Z. (2018). Recent advances in the development of biodegradable PHB-based toughening materials: Approaches, advantages and applications. *Mat. Sci. Eng. C* 92, 1092–1116. doi:10.1016/j.msec.2017.11.006

Chen, M.-Y., Teng, W. K., Zhao, L., Hu, C. X., Zhou, Y. K., Han, B. P., et al. (2021). Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. *ISME J.* 15, 211–227. doi:10.1038/s41396-020-00775-z

Chisti, Yusuf (2007). Biodiesel from microalgae. *Biotechnol. Adv.* 25 (3), 294–306. doi:10.1016/j.biotechadv.2007.02.001

Chiulan, I., Frone, A. N., Brandabur, C., Mihaela, D., and Id, P. (2021). Recent advances in 3D printing of aliphatic polyesters. *Bioengineering* 5, 2–18. doi:10.3390/bioengineering5010002

Claudi, R., Alei, E., Battistuzzi, M., Cocola, L., Erculiani, M. S., Pozzer, A. C., et al. (2020). Super-earths, M dwarfs, and photosynthetic organisms: Habitability in the lab. *Life* 1, 10. doi:10.3390/life11010010

Couté, A., and Bury, E. (1988). Ultrastructure d'une cyanophycée aérienne calcifiée cavernicole: *Scytonema julianum* (frank) richter (hormogonophycideae, nostocales, scytonemataceae). *Hydrobiologia* 160 (3), 219–239. doi:10.1007/bf00007137

Covone, G., Ienco, R. M., Cacciapuoti, L., and Inno, L. (2021). Efficiency of the oxygenic photosynthesis on Earth-like planets in the habitable zone. *Mon. Notices R. Astronomical Soc.* 5053, 3329–3335. doi:10.1093/mnras/stab1357

Cushing, G. E. (2017). "Mars global cave candidate catalog (MGC3)," in Astrobiology Science Conference 2017 (LPI Contrib. No. 1965).

Cushing, G. E., Titus, T. N., Wynne, J. J., and Christensen, P. R. (2007). THEMIS observes possible cave skylights on Mars. *Geophys. Res. Lett.* 34 (17), L17201. doi:10.1029/2007gl030709

Das, M., and Maiti, S. K. (2021). Recent progress and challenges in cyanobacterial autotrophic production of polyhydroxybutyrate (PHB), a bioplastic. *J. Environ. Chem. Eng.* 9 (4), 105379. doi:10.1016/j.jece.2021.105379

Davey, S. C., Ernst, R. E., Samson, C., and Grosfils, E. B. (2013). Hierarchical clustering of pit crater chains on Venus. *Can. J. Earth Sci.* 50 (1), 109–126. doi:10.1139/cjes-2012-0054

De Muynck, W., De Belie, N., and Verstraete, W. (2010). Microbial carbonate precipitation in construction materials: A review. *Ecol. Eng.* 36, 118–136. doi:10.1016/j.ecoleng.2009.02.006

De Vera, J. P., Alawi, M., Backhaus, T., Baqué, M., Billi, D., Böttger, U., et al. (2019). Limits of life and the habitability of Mars: The ESA space experiment BIOMEX on the ISS. *Astrobiology* 19 (2), 145–157. doi:10.1089/ast.2018.1897

Djebaili, R., Mignini, A., Vaccarelli, I., Pellegrini, M., Spera, D. M., Del Gallo, M., et al. (2022). Polyhydroxybutyrate-producing cyanobacteria from lampenflora: The case study of the "Stiffe" caves in Italy. *Front. Microbiol.* 13, 933398. doi:10.3389/fmicb.2022.933398

Engqvist, M. K., McIsaac, R. S., Dollinger, P., Flytzanis, N. C., Abrams, M., Schor, S., et al. (2015). Directed evolution of Gloeobacter violaceus rhodopsin spectral properties. *J. Mol. Biol.* 427 (1), 205–220. doi:10.1016/j.jmb.2014.06.015

Feng, Y., Roleda, M. Y., Armstrong, E., Boyd, P. W., and Hurd, C. L. (2017). Environmental controls on the growth, photosynthetic and calcification rates of a Southern Hemisphere strain of the coccolithophore Emiliania huxleyi. *Limnol. Oceanogr.* 62, 519–540. doi:10.1002/lno.10442

Friedmann, E. I. (1980). "Endolithic microbial life in hot and cold deserts," in *Limits of life* (Dordrecht: Springer), 33–45.

Frone, A. N., Batalu, D., Chiulan, I., Oprea, M., Gabor, A. R., Nicolae, C. A., et al. (2019). Morpho-structural, thermal and mechanical properties of PLA/PHB/Cellulose biodegradable nanocomposites obtained by compression molding, extrusion, and 3d printing. *Nanomaterials* 10 (1), 51. doi:10.3390/nano10010051

Gaier, J. R., Ellis, S., and Hanks, N. (2012). Thermal optical properties of lunar dust simulants. *J. Thermophys. heat Transf.* 264, 573–580. doi:10.2514/1.t3838

Gan, F., Zhang, S., Rockwell, N. C., Martin, S. S., Lagarias, J. C., and Bryant, D. A. (2014). Extensive remodeling of a cyanobacterial photosynthetic apparatus in far-red light. *Science* 345, 62021312–62021317. doi:10.1126/science.1256963

Gattuso, J., and Buddemeier, R. W. (2000). Calcification and CO 2. *Ocean. Biogeochem.* 407, 311–313. doi:10.1038/35030280

Harding, K. G., Dennis, J. S., Blottnitz, H. Von, and Harrison, S. T. L. (2007). Environmental analysis of plastic production processes: Comparing petroleum-based polypropylene and polyethylene with biologically-based poly-β-hydroxybutyric acid using life cycle analysis. *J. Bioeconomics* 130, 57–66. doi:10.1016/j.jbiotec.2007.02.012

Haruyama, J., Hioki, K., Shirao, M., Morota, T., Hiesinger, H., van der Bogert, C. H., et al. (2009). Possible lunar lava tube skylight observed by SELENE cameras. *Geophys. Res. Lett.* 36 (21), L21206. doi:10.1029/2009gl040635

Hauer, T., Mühlsteinová, R., Bohunická, M., KaštovskÝ, J., and Mareš, J. (2015). Diversity of cyanobacteria on rock surfaces. *Biodivers. conservation* 244, 759–779. doi:10.1007/s10531-015-0890-z

Heveran, C. M., Williams, S. L., Qiu, J., Artier, J., Hubler, M. H., Cook, S. M., and Srubar, W. V., III (2020). Biomineralization and successive regeneration of engineered living building materials. *Matter* 2 (2), 481–494. doi:10.1016/j.matt.2019.11.016

Hickman-Lewis, K., Cavalazzi, B., Giannoukos, K., D'Amico, L., Vrbaski, S., Saccomano, G., et al. (2022). Advanced two- and three-dimensional insights into Earth's oldest stromatolites (ca. 3.5 Ga): Prospects for the search for life on Mars. *Geology* 51 (1), 33–38. doi:10.1130/G50390.1

Hodač, L., Brinkmann, N., Mohr, K. I., Arp, G., Hallmann, C., Ramm, J., et al. (2015). Diversity of microscopic green algae (Chlorophyta) in calcifying biofilms of two karstic streams in GermanyVariability in the crystal morphology of calcified terrestrial Scytonema populations (Cyanobacteria, Cyanophyceae). *Geomicrobiol. JournalGeomicrobiology J.* 3210 (3-41), 27559–29064. doi:10.1080/01490459209377904

Hodgskiss, M. S., Crockford, P. W., Peng, Y., Wing, B. A., and Horner, T. J. (2019). A productivity collapse to end Earth's Great Oxidation. *Proc. Natl. Acad. Sci.* 116 (35), 17207–17212. doi:10.1073/pnas.1900325116

Hoffmann, L. (1992). Variability in the crystal morphology of calcified terrestrial Scytonema populations (Cyanobacteria, Cyanophyceae). *Geomicrobiol. J.* 10 (1), 59–64.

Horvath, T., Hayne, P. O., and Paige, D. A. (2022). Thermal and illumination environments of lunar pits and caves: Models and observations from the diviner lunar radiometer experiment. *Geophys. Res. Lett.* 49, 1–9. doi:10.1029/2022gl099710

Irwin, L. N., and Schulze-Makuch, D. (2020). The astrobiology of alien worlds: Known and unknown forms of life. *Universe* 9, 130. doi:10.3390/universe6090130

Jansson, C., and Northen, T. (2010). Calcifying cyanobacteria—The potential of biomineralization for carbon capture and storage. *Curr. Opin. Biotechnol.* 21 (3), 365–371. doi:10.1016/j.copbio.2010.03.017

Johnson, T. J., Anderson, G. A., Gibbons, W. R., Gu, L., and Zhou, R. (2018). Photobioreactor cultivation strategies for microalgae and cyanobacteria. *Am. Inst. Chem. Eng.* 34, 811–827. doi:10.1002/btpr.2628

Jones, B., and Peng, X. (2014). Multiphase calcification associated with the atmophytic cyanobacterium Scytonema julianum. *Sediment. Geol.* 313, 91–104. doi:10.1016/j.sedgeo.2014.09.002

Jorquera, O., Kiperstok, A., Sales, E. A., Embiruçu, M., and Ghirardi, M. L. (2010). Comparative energy life-cycle analyses of microalgal biomass production in open ponds and photobioreactors. *Bioresour. Technol.* 101 (4), 1406–1413. doi:10.1016/j.biortech.2009.09.038

Jung, P., Schermer, M., Briegel-Williams, L., Baumann, K., Leinweber, P., Karsten, U., et al. (2019). Water availability shapes edaphic and lithic cyanobacterial communities in the Atacama Desert. *J. Phycol.* 55 (6), 1306–1318. doi:10.1111/jpy.12908

Kaku, T., Haruyama, J., Miyake, W., Kumamoto, A., Ishiyama, K., Nishibori, T., et al. (2017). Detection of intact lava tubes at marius hills on the moon by SELENE (kaguya) lunar radar sounder. *Geophys. Res. Lett.* 44 (20), 155–161. doi:10.1002/2017gl074998

Kalokora, O. J., Buriyo, A. S., Asplund, M. E., Gullström, M., Mtolera, M. S., and Björk, M. (2020). An experimental assessment of algal calcification as a potential source of atmospheric CO2. *PloS one* 15 (4), e0231971. doi:10.1371/journal.pone.0231971

Kamran, A., Sauter, K., Reimer, A., Wacker, T., Reitner, J., and Hoppert, M. (2020). Cyanobacterial mats in calcite-precipitating serpentinite-hosted alkaline springs of the Voltri Massif, Italy. *Microorganisms* 9 (1), 62. doi:10.3390/microorganisms9010062

Khalifa, S. A., Shedid, E. S., Saied, E. M., Jassbi, A. R., Jamebozorgi, F. H., Rateb, M. E., et al. (2021). Cyanobacteria—from the oceans to the potential biotechnological and biomedical applications. *Mar. Drugs* 19 (5), 241. doi:10.3390/md19050241

Kilgore, C., Johansen, J. R., Mai, T., Hauer, T., Casamata, D. A., and Sheil, C. (2018). Molecular characterization of Geitleria appalachiana sp. nov.(Nostocales, Cyanobacteria) and formation of Geitleriaceae fam. nov, *Fottea* 18(2):150–163. doi:10.5507/fot.2018.002

Knoll, A. H., and Nowak, M. A. (2017). The timetable of evolution. *Sci. Adv.* 3. Jg., Nr. 5, S., e1603076. doi:10.1126/sciadv.1603076

Koch, M., Berendzen, K. W., and Forchhammer, K. (2020b). On the role and production of polyhydroxybutyrate (PHB) in the cyanobacterium synechocystis sp. PCC 6803. *Life* 10, 47. doi:10.3390/life10040047

Koch, M., Bruckmoser, J., Scholl, J., Hauf, W., Rieger, B., and Forchhammer, K. (2020a). Maximizing PHB content in synechocystis sp. PCC 6803: A new metabolic engineering strategy based on the regulator PirC. *Microb. Cell Factories* 19, 231. doi:10.1186/s12934-020-01491-1

Kühl, M., Trampe, E., Mosshammer, M., Johnson, M., Larkum, A. W., Frigaard, N.-U., et al. (2020). Substantial near-infrared radiation-driven photosynthesis of chlorophyll f-containing cyanobacteria in a natural habitat. *Elife* 9, e50871. doi:10.7554/elife.50871

Kuhne, S., Strieth, D., Lakatos, M., Muffler, K., and Ulber, R. (2014). A new photobioreactor concept enabling the production of desiccation induced biotechnological products using terrestrial cyanobacteria. *J. Biotechnol.* 192, 28–33. doi:10.1016/j.jbiotec.2014.10.002

Lacap-Bugler, D. C., Lee, K. K., Archer, S., Gillman, L. N., Lau, M. C., Leuzinger, S., et al. (2017). Global diversity of desert hypolithic cyanobacteria. *Front. Microbiol.* 8, 867. doi:10.3389/fmicb.2017.00867

Lakatos, M., and Strieth, D. (2017). Terrestrial microalgae: Novel concepts for biotechnology and applications. *Prog. Bot.* 79, 269–312.

Lee, J., Park, H. J., Moon, M., Lee, J. S., and Min, K. (2021). Recent progress and challenges in microbial polyhydroxybutyrate (PHB) production from CO2 as a sustainable feedstock: A state-of-the-art review. *Bioresour. Technol.* 339, 125616. doi:10.1016/j.biortech.2021.125616

Lenz, R. W., and Marchessault, R. H. (2005). Bacterial polyesters: Biosynthesis, biodegradable plastics and biotechnology. *Biomacromolecules* 6 (1), 1–8. doi:10.1021/bm049700c

Levchenko, I., Xu, S., Mazouffre, S., Keidar, M., and Bazaka, K. (2021). "Mars colonization: Beyond getting there," in *Terraforming Mars*. Editors M. Beech, J. Seckbach, and R. Gordon (Wiley). doi:10.1002/9781119761990

Léveillé, R. J., and Datta, S. (2010). Lava tubes and basaltic caves as astrobiological targets on earth and Mars: A review. *Planet. Space Sci.* 584, 592–598. doi:10.1016/j.pss.2009.06.004

Li, Y., Lin, Y., Loughlin, P. C., and Chen, M. (2014). Optimization and effects of different culture conditions on growth of *Halomicronema hongdechloris*–a filamentous cyanobacterium containing chlorophyll f. *Front. plant Sci.* 5, 67. doi:10.3389/fpls.2014.00067

Lingam, M., and Loeb, A. (2020). Photosynthesis on exoplanets and exomoons from reflected light. *Int. J. Astrobiol.* 19, 210–219. doi:10.1017/s1473550419000247

Liu, D., Liberton, M., Hendry, J. I., Aminian-Dehkordi, J., Maranas, C. D., and Pakrasi, H. B. (2021). Engineering biology approaches for food and nutrient production by cyanobacteria. *Curr. Opin. Biotechnol.* 67, 1–6. doi:10.1016/j.copbio.2020.09.011

Logan, B. W., Rezak, R., and Ginsburg, R. N. (1964). Classification and environmental significance of algal stromatolites. *J. Geol.* 72, 68–83. doi:10.1086/626965

Luan, G., Zhang, S., and Lu, X. (2020). Engineering cyanobacteria chassis cells toward more efficient photosynthesis. *Curr. Opin. Biotechnol.* 62, 1–6. doi:10.1016/j.copbio.2019.07.004

Madison, L. L., and Huisman, G. W. (1999). Metabolic engineering of poly (3-hydroxyalkanoates): From DNA to plastic. *Microbiol. Mol. Biol. Rev.* 63 (1), 21–53. doi:10.1128/mmbr.63.1.21-53.1999

Mandon, L., Beck, P., Quantin-Nataf, C., Dehouck, E., Thollot, P., Loizeau, D., et al. (2022). Roma: A database of rock reflectance spectra for martian *in situ* exploration. *Earth Space Sci.* 91, e2021EA001871. doi:10.1029/2021EA001871

Mareš, J., Hrouzek, P., Kaňa, R., Ventura, S., Strunecký, O., and Komárek, J. (2013). The primitive thylakoid-less cyanobacterium Gloeobacter is a common rock-dwelling organism. *PLoS One* 8 (6), e66323. doi:10.1371/journal.pone.0066323

Milner, S., Langer, G., Grelaud, M., and Ziveri, P. (2016). Ocean warming modulates the effects of acidification on Emiliania huxleyi calcification and sinking. *Limnol. Oceanogr.* 61, 1322–1336. doi:10.1002/lno.10292

Miscoe, L. H., Johansen, J. R., Kociolek, J. P., Lowe, R. L., Vaccarino, M. A., Pietrasiak, N., et al. (2016). The diatom flora and cyanobacteria from caves on Kauai, Hawaii. *Acta Bot. Hung.* 58, 3–4.

Mulec, J. (2019). "Lampenflora," in *Encyclopedia of caves* (Academic Press), 635–641.

Muñoz-Rojas, M., Chilton, A., Liyanage, G. S., Erickson, T. E., Merritt, D. J., Neilan, B. A., et al. (2018). Effects of indigenous soil cyanobacteria on seed germination and seedling growth of arid species used in restoration. *Plant Soil* 429 (1), 91–100. doi:10.1007/s11104-018-3607-8

Mutale-Joan, C., Sbabou, L., and Hicham, E. A. (2022). Microalgae and cyanobacteria: How exploiting these microbial resources can address the underlying challenges related to food sources and sustainable agriculture: A review. *J. Plant Growth Regul.* 42, 1–20. doi:10.1007/s00344-021-10534-9

Napoli, A., Micheletti, D., Pindo, M., Larger, S., Cestaro, A., de Vera, J. P., et al. (2022). Absence of increased genomic variants in the cyanobacterium Chroococcidiopsis exposed to Mars-like conditions outside the space station. *Sci. Rep.* 12 (1), 8437–8439. doi:10.1038/s41598-022-12631-5

Neumann-Micheau, N., and Tributsch, H. (2018). Luminescence light collection technology in the aragonite of stone corals. *Bioinspiration Biomimetics* 13 (6), 066006. doi:10.1088/1748-3190/aae1bf

Nien, T. S., Bryant, D. A., and Ho, M. Y. (2022). Use of quartz sand columns to study far-red light photoacclimation (FaRLiP) in cyanobacteria. *Appl. Environ. Microbiol.* 88 (13), e0056222–22. doi:10.1128/aem.00562-22

Nürnberg, D. J., Morton, J., Santabarbara, S., Telfer, A., Joliot, P., Antonaru, L. A., et al. (2018). Photochemistry beyond the red limit in chlorophyll f–containing photosystems. *Science* 3606394, 1210–1213. doi:10.1126/science.aar8313

Ohtake, M., Matsunaga, T., Haruyama, J., Yokota, Y., Morota, T., Honda, C., et al. (2009). The global distribution of pure anorthosite on the MoonCyanobacteria in hypersaline environments: Biodiversity and physiological properties. *NatureBiodiversity Conservation* 46124 (72614), 236781–240798. doi:10.1007/s10531-015-0882-z

Olsson-Francis, K., and Cockell, C. S. (2010). Use of cyanobacteria for *in-situ* resource use in space applications. *Planet. Space Sci.* 5810, 1279–1285. doi:10.1016/j.pss.2010.05.005

Oren, A. (2015). Cyanobacteria in hypersaline environments: Biodiversity and physiological properties. *Biodivers. Conserv.* 24, 781–798.

Ozkan, A., Kinney, K., Katz, L., and Berberoglu, H. (2012). Reduction of water and energy requirement of algae cultivation using an algae biofilm photobioreactor. *Bioresour. Technol.* 114, 542–548. doi:10.1016/j.biortech.2012.03.055

Phillips, A. J., Gerlach, R., Lauchnor, E., Mitchell, A. C., Cunningham, A. B., and Spangler, L. (2013). Engineered applications of ureolytic biomineralization: A review. *Biofouling* 29 (6), 715–733. doi:10.1080/08927014.2013.796550

Planavsky, N., Reid, R. P., Lyons, T. W., Myshrall, K. L., and Visscher, P. T. (2009). Formation and diagenesis of modern marine calcified cyanobacteria. *Geobiology* 7, 566–576. doi:10.1111/j.1472-4669.2009.00216.x

Podola, B., Li, T., and Melkonian, M. (2017). Porous substrate bioreactors: A paradigm shift in microalgal biotechnology? *Trends Biotechnol.* 35 (2), 121–132. doi:10.1016/j.tibtech.2016.06.004

Popović, S., Krizmanić, J., Vidaković, D., Karadžić, V., Milovanović, Ž., Pećić, M., et al. (2020). Biofilms in caves: Easy method for the assessment of dominant phototrophic groups/taxa *in situ*. *Environ. Monit. Assess.* 192. Jg., S., 1–17. doi:10.1007/s10661-020-08686-4

Prescott, R. D., Zamkovaya, T., Donachie, S. P., Northup, D. E., Medley, J. J., Monsalve, N., et al. (2022). Islands within islands: Bacterial phylogenetic structure and consortia in Hawaiian lava caves and fumaroles. *Front. Microbiol.* 13, 934708. doi:10.3389/fmicb.2022.934708

Rai, A. N., Söderbäck, E., and Bergman, B. (2000). Tansley review No. 116. *New Phytol.* 147 (3), 449–481. doi:10.1046/j.1469-8137.2000.00720.x

Ramalho, T. P., Chopin, G., Pérez-Carrascal, O. M., Tromas, N., and Verseux, C. (2022). Selection of anabaena sp. PCC 7938 as a cyanobacterium model for biological ISRU on Mars. *Appl. Environ. Microbiol.* 88 (15), e0059422–22. doi:10.1128/aem.00594-22

Rastogi, R. P., Sinha, R. P., Moh, S. H., Lee, T. K., Kottuparambil, S., Kim, Y. J., and Han, T. (2014). Ultraviolet radiation and cyanobacteria. *J. Photochem. Photobiol. B Biol.* 141, 154–169. doi:10.1016/j.jphotobiol.2014.09.020

Ridgwell, A., and Zeebe, R. E. (2005). The role of the global carbonate cycle in the regulation and evolution of the Earth system. *Earth Planet. Sci. Lett.* 234, 299–315. doi:10.1016/j.epsl.2005.03.006

Ritchie, R. J., Larkum, A. W., and Ribas, I. (2018). Could photosynthesis function on Proxima Centauri b? *Int. J. Astrobiol.* 172, 147–176. doi:10.1017/s1473550417000167

Rodeghiero, G., Pernechele, C., Munari, M., Pozzobon, R., Pajola, M., Di Antonio, I., et al. (2022). Radiance values inside lunar caves and lava tubes. *Space Telesc. Instrum. 2022 Opt. Infrared, Millim. Wave* 12180, 1198–1210. SPIE.

Roldán, M., and Hernández-Mariné, M. (2009). Exploring the secrets of the three-dimensional architecture of phototrophic biofilms in caves. *Int. J. Speleology* 38 (1), 41–53. doi:10.5038/1827-806x.38.1.5

Rossi, F., and De Philippis, R. (2015). Role of cyanobacterial exopolysaccharides in phototrophic biofilms and in complex microbial mats. *Life* 5 (2), 1218–1238. doi:10.3390/life5021218

Ryan, J. A., and Sharman, R. D. (1981). Two major dust storms, one Mars year apart: Comparison from Viking data. *J. Geophys. Res. Oceans* 86 (C4), 3247–3254. doi:10.1029/jc086ic04p03247

Sánchez-Baracaldo, P., and Cardona, T. (2020). On the origin of oxygenic photosynthesis and Cyanobacteria. *New Phytol.* 225 (4), 1440–1446. doi:10.1111/nph.16249

Sanfilippo, J. E., Garczarek, L., Partensky, F., and Kehoe, D. M. (2019). Chromatic acclimation in cyanobacteria: A diverse and widespread process for optimizing photosynthesis. *Annu. Rev. Microbiol.* 73, 407–433. doi:10.1146/annurev-micro-020518-115738

Santo, L. (2022). Space sustainability, advanced materials and micro/nanotechnologies for future life in outer Space. *Emergent Mater* 5, 237–240. doi:10.1007/s42247-022-00373-z

Sauro, F., Pozzobon, R., Massironi, M., De Berardinis, P., Santagata, T., and De Waele, J. (2020). Lava tubes on earth, moon and Mars: A review on their size and morphology revealed by comparative planetology. *Earth-Science Rev.* 209, 103288. doi:10.1016/j.earscirev.2020.103288

Sauro, F., Pozzobon, R., Santagata, T., Tomasi, I., Tonello, M., Martı´nez-Frı´as, J., et al. (2019). "Volcanic caves of lanzarote: A natural laboratory for understanding volcano-speleogenetic processes and planetary caves," in *Lanzarote and chinijo islands geopark: From earth to space* (Springer), 125–142.

Scherer, K., Soerjawinata, W., Schaefer, S., Kockler, I., Ulber, R., Lakatos, M., et al. (2022). Influence of wettability and surface design on the adhesion of terrestrial cyanobacteria to additive manufactured biocarriers. *Bioprocess Biosyst. Eng.* 45 (5), 931–941. doi:10.1007/s00449-022-02712-0

Shen, G., Canniffe, D. P., Ho, M.-Y., Kurashov, V., Est, A. v. d., Golbeck, J. H., et al. (2019). Characterization of chlorophyll f synthase heterologously produced in Synechococcus sp. PCC 7002. *Photosynth. Res.* 1401, 77–92. doi:10.1007/s11120-018-00610-9

Sinha, R. P., and Häder, D. P. (2008). UV-protectants in cyanobacteria. *Plant Sci.* 174 (3), 278–289. doi:10.1016/j.plantsci.2007.12.004

Spribille, T., Resl, P., Stanton, D. E., and Tagirdzhanova, G. (2022). Evolutionary biology of lichen symbioses. *New Phytol.* 234 (5), 1566–1582. doi:10.1111/nph.18048

Steinbüchel, A., and Lütke-Eversloh, T. (2003). Metabolic engineering and pathway construction for biotechnological production of relevant polyhydroxyalkanoates in microorganisms. *Biochem. Eng. J.* 16 (2), 81–96. doi:10.1016/s1369-703x(03)00036-6

Talan, A., Pokhrel, S., Tyagi, R. D., and Drogui, P. (2022). Biorefinery strategies for microbial bioplastics production: Sustainable pathway towards Circular Bioeconomy. *Bioresour. Technol. Rep.* 17, 100875. doi:10.1016/j.biteb.2021.100875

Taylor, S. R. (1975). *Lunar science: a post-Apollo view*. New York, NY: Pergamon Press.

Titus, T. N., Wynne, J. J., Malaska, M. J., Agha-Mohammadi, A.-a., Buhler, P. B., Alexander, E. C., et al. (2021). A roadmap for planetary caves science and exploration. *Nat. Astron.* 56, 524–525. doi:10.1038/s41550-021-01385-1

Trinugroho, J. P., Bečková, M., Shao, S., Yu, J., Zhao, Z., Murray, J. W., et al. (2020). Chlorophyll f synthesis by a super-rogue photosystem II complex. *Nat. Plants* 6 (3), 238–244. doi:10.1038/s41477-020-0616-4

Tros, M., Bersanini, L., Shen, G., Ho, M.-Y., Stokkum, I. H. v., Bryant, D. A., et al. (2020). Harvesting far-red light: Functional integration of chlorophyll f into Photosystem I complexes of Synechococcus sp. PCC 7002. *Biochimica Biophysica Acta (BBA)-Bioenergetics* 18618, 148206. doi:10.1016/j.bbabio.2020.148206

Turco, R., Santagata, G., Corrado, I., Pezzella, C., and Di Serio, M. (2021). *In vivo* and post-synthesis strategies to enhance the properties of PHB-based materials: A review. *Front. Bioeng. Biotechnol.* 8, 619266. doi:10.3389/fbioe.2020.619266

Verseux, C., Baqué, M., Lehto, K., Vera, J.-P. P. d., Rothschild, L. J., and Billi, D. (2016). Sustainable life support on Mars–the potential roles of cyanobacteria. *Int. J. Astrobiol.* 151, 65–92. doi:10.1017/s147355041500021x

Verseux, C., Heinicke, C., Ramalho, T. P., Determann, J., Duckhorn, M., Smagin, M., et al. (2021). A low-pressure, N2/CO2 atmosphere is suitable for Cyanobacterium-based life-support systems on Mars. *Front. Microbiol.* 67, 611798. doi:10.3389/fmicb.2021.611798

Viola, S., Roseby, W., Santabarbara, S., Nürnberg, D., Assunção, R., Dau, H., et al. (2022). Impact of energy limitations on function and resilience in long-wavelength Photosystem II. *Elife* 11, e79890. doi:10.7554/elife.79890

Vítek, P., Ascaso, C., Artieda, O., Casero, M. C., and Wierzchos, J. (2020). Raman imaging of microbial colonization in rock — Some analytical aspects. *Anal. Bioanal. Chem.*, 3717–3726. doi:10.1007/s00216-020-02622-8

Viúdez-Moreiras, D., Newman, C. E., De la Torre, M., Martínez, G., Guzewich, S., Lemmon, M., et al. (2019). Effects of the MY34/2018 global dust storm as measured by MSL REMS in Gale Crater. *J. Geophys. Res. Planets* 124 (7), 1899–1912. doi:10.1029/2019je005985

Walden, B. E., Billings, T. L., York, C. L., Gillett, S. L., and Herbert, M. V. (1998). Utility of lava tubes on other worlds. In Using *in situ* Resources for Construction of Planetary Outposts (p. 16).

Wang, J., Van Tittelboom, K., De Belie, N., and Verstraete, W. (2012). Use of silica gel or polyurethane immobilized bacteria for self-healing concrete. *Constr. Build. Mater* 26, 532–540. doi:10.1016/j.conbuildmat.2011.06.054

Westbroek, P., De Jong, E. W., Van der Wal, P., Borman, A. H., De Vrind, J. P. M., Kok, D., and Parker, S. B. (1984). Mechanism of calcification in the marine alga Emiliania huxleyi. *Philosophical Trans. R. Soc. Lond. B, Biol. Sci.* 304 (1121), 435–444.

Williams, S. F., Martin, D. P., Horowitz, D. M., and Peoples, O. P. (1999). PHA applications: Addressing the price performance issue. *Int. J. Biol. Macromol.* 25 (1-3), 111–121. doi:10.1016/s0141-8130(99)00022-7

Williams, S. L., Artier, J., Qiu, J., Heveran, C., Nagarajan, A., Hubler, M., et al. (2019): Regenerative hydrogel-based living microbial mortars: Investigation of viability and strength in successive material generations. 238–243. Pages/ Academic Journal of Civil Engineering, Vol 37 No 2 Special Issue - ICBBM 2019. doi:10.26168/ICBBM2019.34

Wolf, B. M., and Blankenship, R. E. (2019). Far-red light acclimation in diverse oxygenic photosynthetic organisms. *Photosynth. Res.* 1423, 349–359. doi:10.1007/s11120-019-00653-6

Wynn-Williams, D. D. (2000). "Cyanobacteria in deserts—Life at the limit?," in *The ecology of cyanobacteria* (Dordrecht: Springer), 341–366.

Wynne, J. J., Titus, T. N., Agha-Mohammadi, A.-a., Azua-Bustos, A., Boston, P. J., León, P., et al. (2021). Fundamental science and engineering questions in planetary cave exploration. *J. Geophys. Res. Planets*, e2022JE007194. doi:10.1029/2022JE007194

Wyrick, D., Ferrill, D. A., Morris, A. P., Colton, S. L., and Sims, D. W. (2004). Distribution, morphology, and origins of Martian pit crater chains. *J. Geophys. Res. Planets* 109 (E6), E06005. doi:10.1029/2004je002240

Xu, J., Wang, M., Lin, H., Wang, R., Feng, Q., and Xu, X. (2021). Effect of lunar complex illumination on *in situ* measurements obtained using visible and near-infrared imaging spectrometer of Chang'E-4. *Remote Sens.* 1312, 2359. doi:10.3390/rs13122359

Zhang, Z. C., Li, Z. K., Yin, Y. C., Li, Y., Jia, Y., Chen, M., et al. (2019). Widespread occurrence and unexpected diversity of red-shifted chlorophyll producing cyanobacteria in humid subtropical forest ecosystems. *Environ. Microbiol.* 21 (4), 1497–1510. doi:10.1111/1462-2920.14582

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Original Research
PUBLISHED 17 February 2023
DOI 10.3389/fspas.2023.1094282

Check for updates

OPEN ACCESS

EDITED BY
Olga V. Khabarova,
Tel Aviv University, Israel

REVIEWED BY
Terry Zixu Liu,
University of California, Los Angeles, United States
Luis Preisser,
Space Research Institute, Austria

*CORRESPONDENCE
Mirela Voiculescu,
Mirela.Voiculescu@ugal.ro

SPECIALTY SECTION
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 09 November 2022
ACCEPTED 02 February 2023
PUBLISHED 17 February 2023

CITATION
Echim M, Voiculescu M, Munteanu C, Teodorescu E, Voitcu G, Negrea C, Condurache-Bota S and Dănilă EB (2023), On the phenomenology of magnetosheath jets with insight from theory, modelling, numerical simulations and observations by Cluster spacecraft. *Front. Astron. Space Sci.* 10:1094282. doi: 10.3389/fspas.2023.1094282

COPYRIGHT
© 2023 Echim, Voiculescu, Munteanu, Teodorescu, Voitcu, Negrea, Condurache-Bota and Dănilă. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

On the phenomenology of magnetosheath jets with insight from theory, modelling, numerical simulations and observations by Cluster spacecraft

Marius Echim[1,2,3], Mirela Voiculescu[4]*, Costel Munteanu[1], Eliza Teodorescu[1], Gabriel Voitcu[1], Cătălin Negrea[1], Simona Condurache-Bota[4] and Emilian Bujor Dănilă[4]

[1]Institute of Space Science, Măgurele, Romania, [2]Royal Belgian Institute for Space Aeronomy, Brussels, Belgium, [3]Belgian Solar Terrestrial Center of Excellence, Brussels, Belgium, [4]Department of Chemistry, Physics and Environment, European Center of Excellence in Environment, Dunărea de Jos University of Galaţi, Galaţi, Romania

Introduction: During recent years magnetosheath plasma structures called "jets" are identified in spacecraft data as localized regions in the magnetosheath where the dynamic pressure is enhanced compared to the background. Although the nomenclature and detection algorithms vary from author to author, magnetosheath jets are part of a larger class of phenomena which can be globally called magnetosheath irregularities. In this review we focus on elements of jets phenomenology less discussed in the literature, though sustained by theoretical models for solar wind magnetosphere interaction, numerical studies based on Vlasov equilibrium models or kinetic numerical simulations.

Methods: The self-consistency of magnetosheath jets and the preservation of their physical identity (shape and physical properties), implicitly assumed in many recent experimental studies, is discussed in modelling and simulations studies and results as a consequence of kinetic processes at the edges of the jets. These studies provide evidence for the fundamental role played by a polarization electric field sustaining the forward motion of the jet with respect to the background plasma. Another natural consequence is the backward motion of surrounding magnetosheath plasma at the edges of jets. The conservation of magnetic moment of ions leads to a decrease of jets forward speed when it moves into increasing magnetic field. Our review is complemented by an analysis of magnetosheath data recorded by Cluster in 2007 and 2008. We applied an algorithm to detect jets based on searching localized enhancements of the dynamic pressure.

Results: This algorithm identifies a number of 960 magnetosheath jets (354 events in 2007 versus 606 events in 2008). A statistical analysis of jet plasma properties reveals an asymmetric distribution of the number of jets as well as a dawn-dusk asymmetry of jets temperature and density. The perturbative effects of jets on the background magnetosheath density/temperature are stronger in the dusk/dawn flank. We also found evidence for deceleration and perpendicular heating of jets with decreasing distance to the Earth. The braking of jets is correlated with the

variation of the magnetic field intensity: the stronger the magnetic field gradient, the more efficient is the jet breaking.

KEYWORDS

magnetosheath, plasma jets, magnetosphere, Cluster, solar wind-magnetosphere coupling, dawn-dusk asymmetries, adiabatic breaking

Introduction

Dynamical properties of magnetosheath plasma irregularities/jets from observations, theory and modelling

Models of solar wind—magnetosphere interaction postulate the key role of solar wind and magnetosheath irregularities, characterized by an excess of their dynamic pressure and/or momentum with respect to the background magnetosheath plasma state, in transferring the mass, momentum and energy from the solar wind to the magnetosphere (Lemaire, 1977; Lemaire and Roth, 1978; Lemaire and Roth, 1981). In these early models it is argued that such plasma irregularities propagate as self-sustained structures through the entire magnetosheath and finally collide with the magnetopause. Due to their excess of momentum/dynamic pressure compared to the magnetosheath average, they can move across the magnetopause and enter into the magnetosphere (Lemaire, 1977). In these models the magnetopause is viewed as a three dimensional surface defined as the locus of total (dynamic + magnetic) pressure equilibrium between the magnetosheath and magnetospheric plasma, and where the magnetic field experiences rather rapid variations.

Magnetosheath irregularities are investigated theoretically and with *in-situ* experimental data. They are described in the literature under various names, like magnetosheath *plasma blobs or plasmoids* (Lemaire, 1985; Echim and Lemaire, 2000; Karlsson et al., 2012; Karlsson et al., 2015), magnetosheath *transient flux enhancements* (Nemecek et al., 1998), *plasma clouds* or *plasma transfer events* (Lundin et al., 2003), magnetosheath *pressure pulses* (Archer et al., 2012), *plasma jets* (Echim and Lemaire, 2005; Savin et al., 2008; Hietala et al., 2009; Amata et al., 2011), *plasma filaments* (Lyatsky et al., 2016a; Lyatsky et al., 2016b) (see also Table 1 in Plaschke et al., 2018).

Some authors use the term "jet" for magnetosheath structures characterized by an excess of dynamic pressure, $p_{dyn} = \rho V^2$ (e.g., Plaschke et al., 2013), while the term "plasmoid" is used for magnetosheath irregularities with an excess of density, ρ (e.g., Karlsson et al., 2012), or momentum, $p = \rho V$ (Lemaire, 1977; Lemaire, 1985). Note also that the term "plasmoid" was first used by Bostick (1956) to define self-organized plasma elements created in laboratory experiments and characterized by i) a measurable translation (bulk) speed, ii) a transverse electric field, iii) a measurable magnetic moment, and iv) a measurable size (see also Roth, 1995). The term "*fast plasmoids*" is used for plasmoids characterized by a "pure" excess of velocity, V, with respect to the background magnetosheath (Karlsson et al., 2012, 2015; Goncharov et al., 2020). Plashke et al. (2018) and Goncharov et al. (2020) attempt a classification of various magnetosheath irregularities based on their dynamical properties (see also Preisser et al., 2020). In the following we will use the term "magnetosheath jet" (MSHJ) for the magnetosheath plasma structures characterized by an excess of the dynamics pressure compared to a background value. Studies of magnetosheath jets benefit from high resolution, multi-point measurements provided by spacecraft like Cluster, THEMIS and MMS during the minimum and/or maximum phase of the solar cycle (Shue et al., 2009; Archer et al., 2012; Plaschke et al., 2013; Plaschke et al., 2016; Plaschke et al., 2020; Plaschke et al., 2018; Karlsson et al., 2018; Raptis et al., 2019; Raptis et al., 2020; Vuorinen et al., 2019; Goncharov et al., 2020; Escoubet et al., 2020; Dmitriev et al., 2021; Koller et al., 2022, see Plaschke et al., 2018 for a review). High resolution data allow for a better sampling of plasma parameters and a better statistics for jet detection.

Dynamical properties of magnetosheath irregularities are derived from statistical analyses of data provided by various spacecraft, such as Cluster (e.g., Karlsson et al., 2012; Karlsson et al., 2015), THEMIS (e.g., Shue et al., 2009; Archer et al., 2012; Plaschke et al., 2016; Plaschke et al., 2018; Vuorinen et al., 2019), MMS or a combination between these (e.g., Raptis et al., 2019; Escoubet et al., 2020; Goncharov et al., 2020; Plaschke et al., 2020; Raptis et al., 2020). A review of jet observations is given by Plaschke et al. (2018). In terms of scales, it is found that the structures characterized by an excess of density have spatial scales of the order of one to several Earth radii (Karlsson et al., 2012). The structures characterized by an excess of dynamic pressures span a larger range, skewed however towards smaller scales, with a peak at roughly one R_E (Plaschke et al., 2013; Plaschke et al., 2016), revised later to 0.1 R_E (Plaschke et al., 2020). A tendency of jets to expand while moving away from the bow-shock is reported by Goncharov et al. (2020) based on MMS data.

In terms of their speed, one study based on THEMIS data reports than 98% jets are super-Alfvénic; however, this finding is based on the analysis of the velocity component in the anti-sunward direction (Plaschke et al., 2013). A superposed epoch analysis of the angle between magnetic field and ion velocity vectors as a function of normalized times suggests that the angles are changing only by about 10° (Plaschke et al., 2020). Liu et al. (2020) show that approximately 13% of high speed magnetosheath jets have a bow wave ahead of them with Mach number typically larger than 1.1, and that for such jets the electron energy flux is enhanced on average by a factor of 2 as compared to both those without bow waves and the ambient magnetosheath. A statistical analysis of jets detected by MMS suggest their speed tends to decrease with the distance from the bow-shock (Goncharov et al., 2020).

Jets are omnipresent in the magnetosheath. Vuorinen et al. (2019) estimate, in a study based on 3 years of THEMIS observations, that "jets larger than >2.0 R_E hit the magnetopause around 9.4 times per hour during quasi-radial IMF, around 4.1 times per hour during oblique IMF and around 0.85 times per hour during

high cone angle IMF." According to the same study, smaller sized jets (0.5–2.0 R_E) impact almost continuously the magnetopause, at a rate between 3.3 and 0.31 jets per minute, depending on the IMF orientation.

In our view, all variants of magnetosheath jets/plasmoids/ion flux enhancements reported in the literature from various spacecraft observations pertain to the same general class of magnetosheath phenomena, which can be named "multiscale dynamical irregularities of the background magnetosheath plasma state." It is a paradigm which covers the broad range of spatio-temporal structures formed in the magnetosheath and contributing to the observed variability; these structures manifest themselves as local perturbations of plasma parameters and/or field, spanning a large range of spatio-temporal scales. They represent an experimental confirmation of the early ideas advocated by Lemaire (1977), Lemaire (1985), Lemaire and Roth (1978), and Lemaire and Roth (1981) who emphasized the key role of isolated magnetosheath plasma structures for the physics of solar wind—magnetosphere interaction. The magnetosheath irregularities are most likely three-dimensional plasma structures, fulfilling the "plasmoid" definition given by Bostick (1956), see also Roth (1995). Indeed, the points i) and iv) of Bostick's definition introduced above are confirmed by observations in the magnetosheath (see, e.g., Plaschke et al., 2017; Karlsson et al., 2018).

Theoretical arguments, models and numerical simulations for jets origin and dynamics

The formation of magnetosheath jets irregularities is considered the effect of several possible physical mechanisms acting in the solar wind and/or at the terrestrial bow shock. There is no consensus on the dominant one(s). Lemaire (1977) and Lemaire (1985) suggests the magnetosheath irregularities impacting the terrestrial magnetopause are linked to the inherent multiscale and omnipresent non-homogeneity of the solar wind. In a study based on Cluster data; Karlsson et al. (2015) finds evidence for solar wind irregularities with properties very similar to magnetosheath jets and argues that "diamagnetic magnetosheath plasmoids originate from the solar wind plasmoids that cross the bow shock and convect downtail in the magnetosheath;" Parkhomov et al. (2021) considers that magnetosheath plasmoids are closely related to a class of solar wind irregularities which they called "diamagnetic structures (DS).'

Nemecek et al. (1998) argues that an interplanetary magnetic field (IMF) discontinuity interacting with the bow shock is at the origin of the plasma flux enhancements observed in the magnetosheath. In Savin et al. (2008) an argument is given that the high kinetic energy jets observed by INTERBALL are formed due to local processes mainly linked to a transition from a metastable plasma state with super-Alfvénic velocity to a stable state with Alfvénic or sub-Alfvénic flows. The induced "magnetic stress balance" leads to the observed high kinetic energy fluxes. Hietala et al. (2009) suggest that the local bow shock ripples occurring when the IMF is radial and the solar wind Mach number is large ($M_A > 10$) could be the source of super magnetosonic magnetosheath jets. This mechanism is also evidenced in hybrid simulations (Preisser et al., 2020). Simulations also suggest antiparallel reconnection observed at the quasi-parallel bow-shock can lead to the formation of diamagnetic irregularities/plasmoids (Preisser et al., 2020). Escoubet et al. (2020) consider the solar wind nanodust clouds (Lai and Russell, 2018) could be added in the list of possible sources for magnetosheath jets. A recent study (Raptis et al., 2020) discusses the shock reformation as an additional mechanism possibly contributing to the formation of magnetosheath jets. Voros et al. (2019) discuss small scale magnetosheath dynamical (coherent) structures as a local manifestation of the non-linear turbulent processes developing in the magnetosheath (for a recent review on magnetosheath turbulence see Echim et al., 2021). Larger scale structures observed in the vicinity of the magnetopause, close to the equatorial plane, can be linked to the formation of Kelvin-Helmholtz vortices (Hasegawa et al., 2004). The non-linear stage of the Kelvin-Helmholtz instability (Nykiri. 2013), may lead to detachment of rolled-up vortices which may interact with the magnetopause and lead to transport into the magnetosphere (Nykiri and Otto, 2001). In the close vicinity of the magnetopause such vortices can be detected as local irregularities of the magnetosheath plasma.

Lemaire (1985) discusses the physical properties of magnetosheath plasmoids from a theoretical (kinetic) point of view. It is argued that magnetosheath plasmoids dynamics in a background plasma and field is characterized by several key aspects like: a) a high plasma dielectric constant, b) the decoupling from the background plasma and field due to electric fields confined in the plasmoid's boundary layers, c) the formation of a self-polarization electric field allowing the forward motion of the plasmoid, d) the breaking of the plasmoid when moving in increased magnetic field due to a conversion of the bulk forward motion into gyration, similar to the physics of mirroring particles. This process, called *adiabatic breaking*, was probed in laboratory (e.g., Burgess and Scholer, 2013) and by numerical simulations (Voitcu and Echim, 2016).

Echim et al. (2005) describe a Vlasov equilibrium model for a *plasma slab*, defined as a localized plasma structure in motion with respect to the background plasma and field and which is characterized by an excess of the dynamic pressure compared to the background. The Vlasov model emphasizes dynamical features of the plasma slab which seem similar to properties of magnetosheath jets reported from observations by, for instance, Plaschke et al. (2017) or Karlsson et al. (2018). Indeed, counterstreaming of the background plasma at the edges of the magnetosheath slab/jet, as well as the formation of asymmetries at jet's edges are observed both in observations and model.

Interestingly, the vast majority of statistical analysis of jets properties based on experimental data implicitly assume the jets are self-sustained plasma structures which preserve their coherence, or self-identity, during their propagation in the magnetosheath, from the bow-shock to the magnetopause (e.g., Archer and Horbury, 2013; Dmitriev and Suvorova, 2015; Karlsson et al., 2015; Raptis et al., 2020). The physical mechanisms allowing for the self-existence and organization of such structures are less discussed in the observational reports. Nevertheless, the theoretical model proposed by Lemaire (1985) to describe the dynamics of magnetosheath plasmoids argues that kinetic processes at plasmoid's boundary layers contribute to an electric self-polarization (Schmidt, 1960) which allows the plasma irregularity to preserve its shape and dynamics.

Echim and Lemaire (2005) compute local Vlasov equilibrium solutions tailored for the boundary layers of plasma jets and demonstrate qualitatively and quantitatively that the jet is decoupled from the background plasma and field by a combination of Schmidt-like polarization electric field and weak-double-layers. The weak parallel electric fields/perpendicular polarization electric field ensures jet's decoupling in the direction parallel/perpendicular to the background magnetic field. At the origin of these electric field components stay plasma bulk velocity shears with a non-vanishing component in the direction parallel and perpendicular to the background magnetic field.

Three-dimensional kinetic simulations of jets injected into non-uniform distribution of the magnetic field (Voitcu and Echim, 2016; Voitcu and Echim, 2017) confirm the electric self-polarization of jets is a key physical process. These simulations evidence the self-formation of space charge layers at the edges of the moving jet sustaining a Schmidt-like polarization electric field. The jet maintains its self-coherence during the transport through the region of non-uniform field with a strong gradient, similar to the magnetosheath region in the vicinity of the magnetopause. A diffusion of jets mass is observed in the direction parallel to the magnetic field. However, in these numerical simulations the contribution of the background plasma is neglected and therefore the weak double-layers predicted theoretically cannot be simulated and their confining/decoupling effect is not simulated.

Data from magnetospheric spacecraft collected during the last decades confirm that dynamical irregularities populate the entire magnetosheath, for a broad range of solar wind input conditions. They carry a significant amount of mass and momentum towards the magnetopause, at virtually all local times and latitudes. A large fraction of these irregularities reaches the magnetopause and interacts with it locally. The emerging picture is that of quasi-permanent localized perturbations of the magnetopause driven by magnetosheath dynamical irregularities, as advocated by Lemaire (1977) and Lemaire and Roth (1978).

Magnetospheric and ionospheric effects of magnetosheath jets

The effects of magnetosheath irregularities/jets on the dynamics and electrodynamics of the terrestrial magnetosphere and ionosphere are under study. Based on first order kinetic plasma physics arguments, Lemaire 1977 and Lemaire 1985 argues that i) the magnetosheath irregularities are effectively decoupled from the background plasma and field, ii) they traverse the magnetopause and iii) they entry into the magnetosphere where they become local plasma perturbations. It is also postulated that intruding magnetosheath plasma jets are braked and eventually stopped in the magnetosphere by adiabatic and/or irreversible processes (e.g., coupling with the conducting ionosphere, see also Echim and Lemaire, 2000; Echim and Lemaire, 2002; Wing et al., 2014). The conditions for magnetopause traversal are investigated with 3D electromagnetic Particle-in-Cell numerical simulations (Voitcu and Echim, 2016; Voitcu and Echim, 2017; Voitcu and Echim, 2018). Magnetospheric observations of fully penetrated magnetosheath plasma elements are reported by studies based on data from magnetospheric missions like Prognoz-7 (Lundin and Aparicio, 1982; Lundin and Dubinin, 1984; Lundin, 1988), INTERBALL (Vaisberg et al., 1998), Cluster (Lundin et al., 2003; Gunell et al., 2012; Lyatsky et al., 2016a; Lyatsky et al., 2016b), THEMIS (Dmitriev and Suvorova, 2015).

Lundin et al. (2003) report Cluster observations of 134 magnetosheath jets detected, between January—March 2001, inside the magnetosphere, at high latitudes (they call these structures *plasma transfer events—PTE*). The authors discuss local magnetospheric effects like plasma drifts and energization, electric fields and currents occurring in response to the interaction of the background magnetospheric plasma with the penetrating magnetosheath jets/PTEs. "Evanescent" PTEs are magnetosheath plasma structures which no longer propagates in the magnetosphere but are stagnant and form static pressure irregularities. Lundin et al. (2003) also argue that the most likely mechanism leading to the occurrence of PTEs observed by Cluster is the transport of magnetosheath irregularities with an excess of dynamic pressure at the magnetopause and their impulsive penetration inside the magnetosphere. A more recent analysis of Cluster data recorded between 2007 and 2008 close to the equatorial regions, within a maximum cone angle range of 30°, finds evidence for more than 200 magnetosheath-like plasma structures, called *plasma filaments*, inside the magnetosphere (Lyatsky et al., 2016a; Lyatsky et al., 2016b). Generally, these plasma filaments show a strong stable earthward component of the plasma bulk velocity. This study also argues that a majority of the observed filaments are detached from the magnetopause. In a study of 642 large scale magnetosheath jets detected by THEMIS, Dmitriev and Suvorova (2015) find that more than 60% of jets move across the magnetopause, into the magnetosphere. The penetrating jets are generally characterized by velocities higher than 200 km/s and a large beta.

Recently, Parkhomov et al. (2021) and Parkhomov et al. (2022) analyze the magnetospheric effects of solar wind irregularities which they call "diamagnetic structures" (DS) based on the anticorrelation between the solar wind plasma density, which increases, and the magnetic field intensity, which decreases. The authors consider that DS move from the solar wind in the magnetosheath and in the magnetosphere and show evidence for global magnetic field perturbations linked to the interaction of DSs with the Earth's magnetosphere. One prominent feature is the occurrence of bursts of irregular magnetic field pulsations PI1-2 recorded on ground for a broad range of local times. Parkhomov et al., 2021; Parkhomov et al., 2022) argue that the pulsations observed by ground observatories have properties similar to the pulsations detected *in-situ* in the vicinity of magnetosheath jets (Katsavriasi et al., 2021). It is argued that the propagation of the pulsations to the ground, along closed magnetic field lines, proves the irregularities, which are the source of pulsations, moved impulsively from the magnetosheath into the magnetosphere, across the magnetopause (Echim and Lemaire, 2000).

Inside the magnetosphere, the propagating jets lead to the formation of sheared plasma flows layers at the interface with the background plasma (see, e.g., Lundin et al., 2003; Gunnell et al., 2012). It was shown that the coupling of sheared flow plasma layers with the conducting ionosphere leads to auroral effects sustained by a magnetospheric generator formed in the sheared flow region (Roth., 1995; Echim et al., 2007; Génot et al., 2021). Indeed, such

generators sustain field aligned potential drops leading to accelerated precipitating electrons and auroral arcs in regions like the Low Latitude Boundary Layer, the Plasma Sheet Boundary Layer or the polar cap (Echim et al., 2008; Echim et al., 2009; Balogh et al., 1997; Johnson and Wing, 2015, see also the review by Borovsky et al., 2020). The sheared flows formed at the interface between magnetosheath jets propagating inside the magnetosphere and the background magnetospheric plasma can be at the origin of throat auroras reported in the polar caps by Han et al. (2017), Han et al. (2018), Wang et al. (2018), and Raptis et al. (2019).

Other magnetospheric effects of magnetosheath jets are discussed by Hietala et al. (2009) who argue that high speed jets provide a "source for magnetopause waves during steady solar wind conditions," which can have an impact on the magnetosphere-ionosphere coupling. Plaschke et al. (2016) present a series of possible consequences of high speed magnetosheath jets such as: ionospheric flow enhancements, magnetic field variations observed on the ground, local magnetopause reconnection, inner magnetospheric and boundary surface waves, drop outs and other variations in radiation belt electron populations. Archer et al. (2019) evidence magnetopause perturbative effects of jets in terms of eigenmodes propagating tangentially to the magnetopause. Liu et al. (2020) argue that a magnetosheath jet compresses the ambient plasma such that a bow wave or shock can form ahead of the jet, which accelerates particles, and thus, contributes to magnetosheath heating and particle acceleration in the extended environment of the Earth's bow shock.

An analysis of magnetosheath jets properties detected by Cluster 3 in 2007 and 2008

In order to probe some of the theoretical predictions less investigated in the past and outlined in the previous section, like, e.g., the formation of space charge layers at the edges of magnetosheath jets, enabling their self-consistent forward motion in the background magnetosheath, the adiabatic breaking and the correlation between local jet perpendicular component of the bulk velocity and the local magnetic field intensity, the correlation between the local jet perpendicular temperature and the local magnetic field intensity, we performed an analysis of magnetosheath data collected by Cluster 3 (Escoubet et al., 1997) between 2007 and 2008. We applied a procedure to identify magnetosheath jets which follows the algorithm proposed by Archer and Horbury (2013) based on searching for local enhancements of the magnetosheath ion dynamic pressure with respect to a background reference value estimated from a running window average. In the following we discuss the observed dynamical properties of jets and how the possible links with the phenomenology presented in the previous section.

There is no general consensus on which methodology to be applied for detecting magnetosheath jets (MSHJ) from *in-situ* spacecraft observations. Different data selection criteria focus on different plasma parameters, such as dynamical pressure, density, bulk velocity. An earlier study identified "ion flux enhancements" (Nemecek et al., 1998), precursors of what we call today magnetosheath jets, from combined observations of plasma density and plasma velocity. Savin et al. (2008) consider a jet is defined by an increase of the ion kinetic energy well above the corresponding solar wind value. Hietala et al. (2009) used a combination of criteria to define the magnetosheath jets: the ion velocity takes values larger than 500 km/s, the jet shows a dominant sunward magnetic field component B_X and the angle between the flow direction and the magnetic field is less than 20°. Archer et al. (2012) defined jets as "magnetosheath pressure pulses" based on dynamic pressure enhancements above a background magnetosheath value. Karlsson et al. (2012) identified jets as "plasmoids" when the plasma density was larger than the background level. Plaschke et al. (2013) used the term "high speed jets" for magnetosheath structures satisfying several conditions, like an anti-sunward dynamic pressure of the jet be higher than half of the solar wind (SW) dynamic pressure.

Recent reports use the term "magnetosheath jets" (Raptis et al., 2020) for events characterized by a magnetosheath dynamic pressure higher than the corresponding 20 min average of SW dynamic pressure. In Plaschke et al. (2020), Liu et al. (2020) and Escoubet et al. (2020), "magnetosheath jets" are defined as events for which the dynamic pressure in the anti-sunward direction, in the GSE system, is larger than half of the SW dynamic pressure, while their duration is defined as the time interval for which the magnetosheath dynamic pressure is larger than a quarter of the SW corresponding magnitude.

Cluster 3 data selection procedure

In this study we analyze magnetosheath data provided by Cluster 3. The time intervals when the spacecraft 3 was in the magnetosheath are extracted from the FP7 STORM database built for an analysis of magnetosheath turbulence (http://www.storm-fp7.eu). Cluster's highly eccentric orbit allows for a seasonal sweeping of the magnetosheath between January and May of each year. All magnetosheath time intervals selected for this study were verified by a visual inspection of data summary plots provided by the Cluster UK database (http://www.cluster.rl.ac.uk) to ensure there is no mixing with the foreshock and/or solar wind.

The procedure we adopt here to identify the magnetosheath jets follows the method proposed by Archer and Horbury (2013) based on searching an excess of the dynamic pressure with respect to a background value, estimated from a running average with a time window of fixed length. The jet detection algorithm proceeds in two steps. First, the signal is divided in contiguous windows of equal time length, Δt. An automatic survey is performed to select those time windows within which at least one local value of the ion dynamic pressure, p_{dyn}, is N times larger than the background value. The latter is estimated from an average of p_{dyn} computed over all values included in the respective time window. We tested several variants of this approach, for various values adopted for N (N = 1.3, N = 1.5, N = 2), and for Δt (Δt = 10, 15 or 20 min). After verification of all results, we adopted a combination of parameters with N_s = 1.3 and Δt_s = 20 min. For larger values of N_s some smaller amplitude jets were missed; for smaller values of Δt_s the procedure required more time and some of the medium sized jets were missed. In the second step, the data for each time interval are inspected visually in order to select the time of start and time of end for each individual jet. In Figure 1 we illustrate an example of Cluster-3 which shows two jets identified in 22 January 2007. The two stars indicate the maximum of the dynamic pressure for each of the two jets. In the analyses included in this study for various plasma parameters

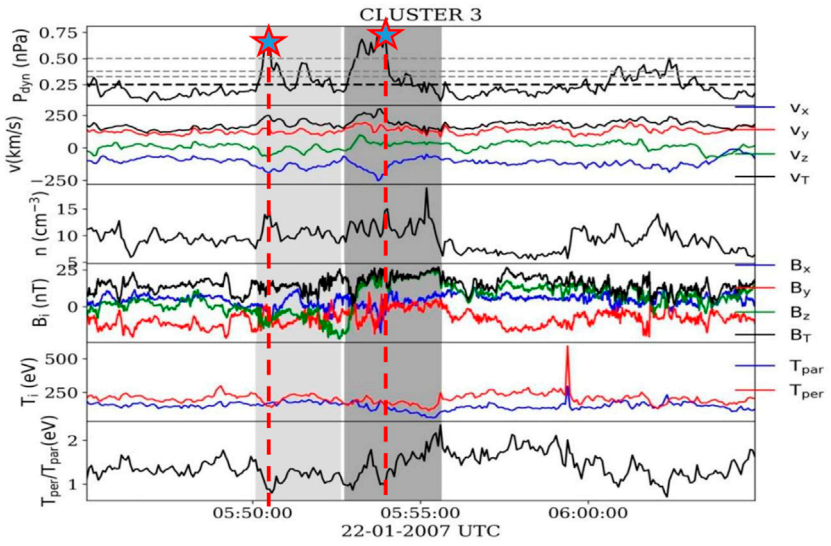


FIGURE 1
An example of Cluster 3 data illustrating two jets detected with the procedure described in the text. The panels show, from top to bottom, the dynamical pressure, p_{dyn}, the three components of the ion bulk velocity and the total speed, the number density, n, the three components of the magnetic field and the magnetic field intensity, the parallel and perpendicular temperature, the temperature anisotropy. The shaded areas indicate the two jets detected at 05:50:00 UT and 05:53:10 UT, respectively. The thin dashed lines represent the average value of the dynamic pressure computed over the 20 min of data showed in the figure and multiplied by N = 1.3, N = 1.5, N = 1.7, N = 2.The thick dashed line indicate the average value of the dynamic pressure. The two stars mark the maximum of the dynamic pressure and the vertical red dashed lines mark the time when this maximum is detected.

(Figures 4–7, 9–13) we consider one value of each variable per jet, precisely the one corresponding to the peak value of the dynamic pressure, exemplified by the two stars in Figure 1.

The method was applied on data from Cluster Ion Spectrometer (CIS, Reme, 1997) on board Cluster 3 and resulted in the detection of 960 events of which 354 jets in 2007 and 606 jets in 2008. Figure 2 illustrates the dynamical characteristics of the entire ensemble of jets; in this figure we show only one measurement per jet, namely, the measurement taken when the jet dynamic pressure reaches its maximum. One notes a significant variability of jets properties. The density, the antisunward velocity, the total magnetic field and the ion temperature at maximum dynamic pressure span a large range of values. A discussion on these variations is provided in the section below.

For each jet we also compute the position of a model magnetopause (using the approach by Shue et al., 1997) and of a model bow-shock (using the model Farris and Russel, 1994). The two models are initialized with solar wind data provided by OMNI database (https://omniweb.gsfc.nasa.gov/html/HROdocum.html). Thus, for each jet, we estimate model curves for:

a) the magnetopause (Shue et al., 1997):

$$r = r_0\left(\frac{2}{1+\cos\theta}\right)^{\alpha} \tag{1}$$

where $\alpha = 0.5$ is the tail flaring, and r_0 is the standoff distance and depends on the dynamical pressure, D_p, and the Oz component of the IMF, B_z (taken from the OMNI data) as:

$$r_0 = \begin{cases} (11.4 + 0.013B_z)\left(D_p\right)^{-\frac{1}{6.6}} & for\ B_z \geq 0 \\ (11.4 + 0.14B_z)\left(D_p\right)^{-\frac{1}{6.6}} & for\ B_z < 0 \end{cases} \tag{2}$$

b) for the bow shock (Farris and Russell, 1994):

$$R_{BS} = R_{BS_0}\frac{1+\varepsilon}{1+\varepsilon\cos\theta} \tag{3}$$

where R_{BS} is the bow-shock (BS) radial distance, θ is the solar zenith angle in aberrated coordinates, $\varepsilon = 0.81$ is the eccentricity, and R_{BS_0} is the bow-shock standoff distance taken form OMNI data.

For each jet we also calculate an estimation for the distance to the model magnetopause (denoted D_{MP}) as well as to the model bow shock (denoted D_{BS}). The two distances are estimated in the direction normal to the model magnetopause (MP) and bow-shock (BS), from jet's position, as illustrated in Figure 3A. Other approaches considered mapping to the bow-shock along magnetosheath streamlines (Karlson et al., 2018). More details on the procedure applied to determine the normal direction to the model bow shock and model magnetopause can be found in Teodorescu et al. (2021). Note that the data needed to compute D_{MP} and D_{BS} were available for a subset of 850 jets, illustrated in Figure 3 where we show their positions in the $(X,R^{YZ})_{GSE}$ plane, where $R^{YZ} = \sqrt{Y^2 + Z^2}$. Note also that 10% of jets are found outside the limits of the model magnetosheath (i.e., located either downstream the model magnetopause, inside the magnetosphere, or upstream the model bow shock). This means that the model boundaries (MP and BS) do not capture the true position of the respective boundaries for some time intervals; nevertheless, the jets are detected inside the "real" observed magnetosheath.

The model magnetopause and bow shock profiles shown in Figure 3B correspond to the innermost model magnetopause and outermost model bow shock of the ensemble of all model magnetopauses and bow-shocks computed for the entire set of jets. For a compact and symmetric view, we plot the distances to

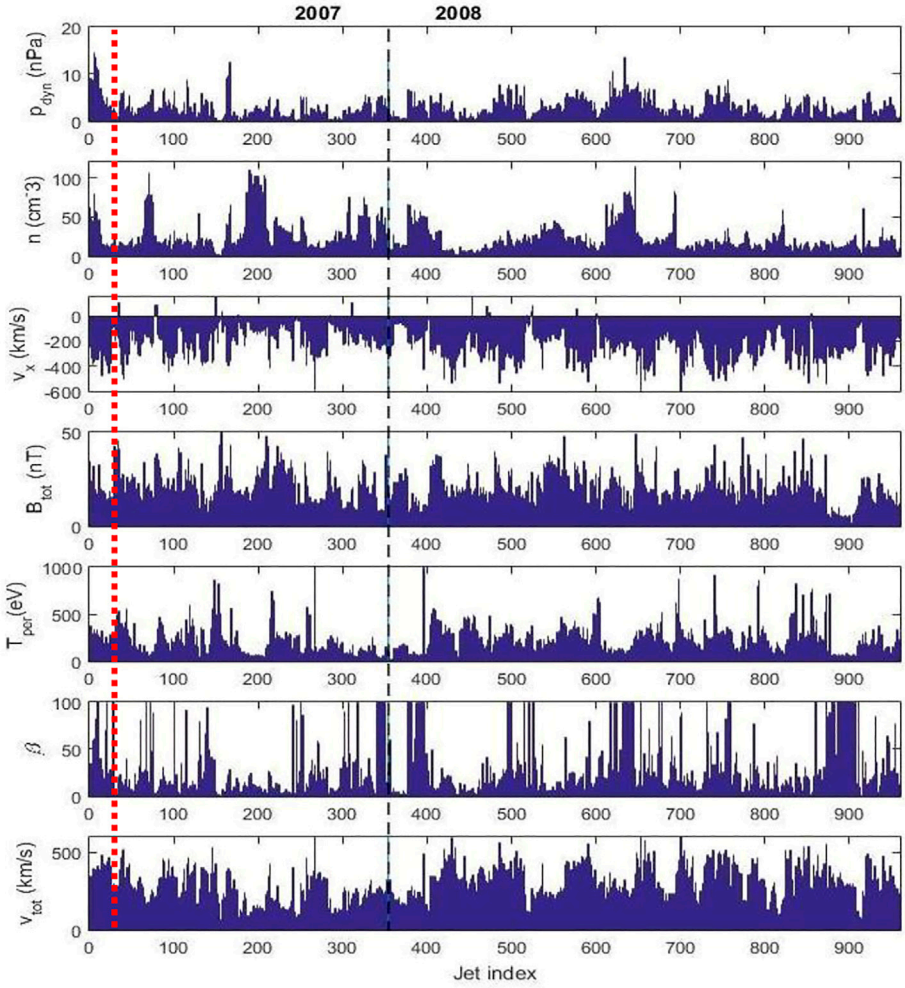


FIGURE 2
A summary of jets' properties included in this study. The panels show from top to bottom: the dynamic pressure, p_{dyn}, the plasma density, n, the component of the jet's velocity in *GSE x* direction, v_x, the magnetic field intensity, B_{tot}, the perpendicular ion temperature, T_{perp}, the total plasma beta, β, and the bulk speed, v_{tot}. Only one measurement per jet is shown, the one at the moment when the jet's dynamic pressure is maximum. The red dashed red line indicates the two jets presented in detail in Figure 1.

the magnetopause (highlighted by the gray shaded area) in a "mirrored" view, by artificially changing the values of the coordinate R^{YZ} from positive to negative. We defined five bins for the estimated distances: [0.,1.] R_E, [1.,2.] R_E, [2.,3.] R_E, [3.,5.] R_E and [5.,8.] R_E, where R_E is the Earth's radius; these bins are illustrated with different colors, as indicated in the figure.

Asymmetries of jets properties observed by Cluster in 2007 and 2008

During the first 4 months of 2007 and 2008 the Earth's magnetosphere was impacted by a sequence of recurrent high-speed streams (HSS) and Corotating Interaction Regions (CIRs), whose dynamical characteristics (maximum speed, maximum density) were different in 2007 compared to 2008 (see, e.g., Munteanu et al., 2019). The differences in driver's properties for 2007 compared to 2008 motivated us to analyze separately each of the 2 years. Koller et al. (2022) recently pointed out, from an analysis of THEMIS data, that the number of jets increases by 50% during HSS and CIR events A thorough analysis of the correlation between the interplanetary parameters and the properties of jets included in our database is the subject of another forthcoming study.

A set of dynamical properties (density, bulk velocity, magnetic field, temperature and dynamic pressure) of jets observed by Cluster 3 in 2007 and 2008 is illustrated in Figures 4, 6 as a function of jet position projected in the (X_{GSE}, Y_{GSE}) plane. In addition to the dynamical properties of jets, we plot in Figures 5, 7 the perturbation of the background magnetosheath produced by the jets, for the same set of physical variables as in Figures 4, 6. The perturbation of a physical parameter Q is estimated in Figures 5, 7 from the ration, $\frac{Q^{jet}-\langle Q\rangle}{\langle Q\rangle}$. Q^{jet} is the value of the variable in the jet (at the moment when the maximum of the dynamic pressure is observed); <Q> is the background value evaluated as the average between two estimations of the background magnetosheath state, one prior to jet detection, Q_1, and the other one post jet detection, Q_2. Q_1 is computed as an

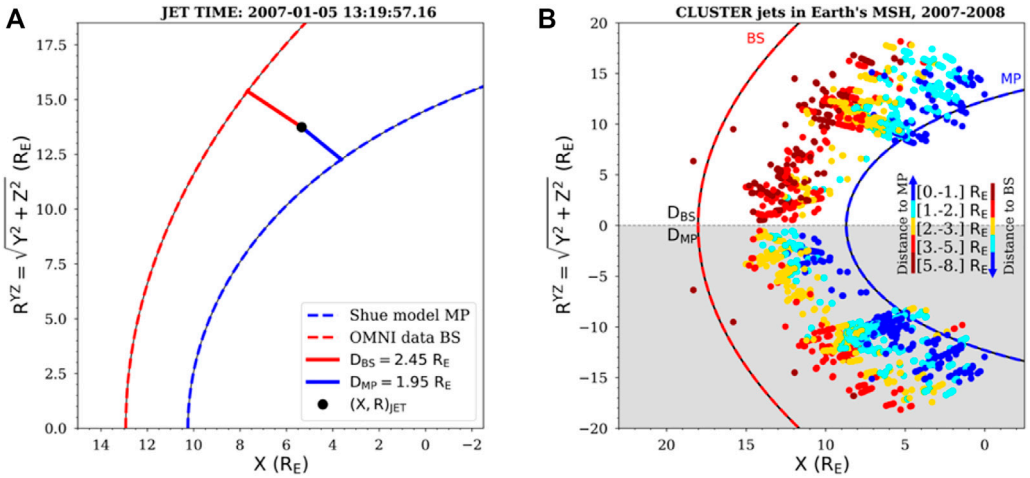


FIGURE 3
(A) Example of the estimated distance from the jet to the magnetopause (blue-curve) and to the bow shock (red-curve), **(B)** Overview of all jets locations in the (X, R^{YZ}) GSE plane: Positive R plane shows distances with respect to the bow shock; negative R plane (gray shaded area) shows distances with respect to the magnetopause. The blue and red curves indicate the innermost magnetopause and the outermost bow shock within the entire set.

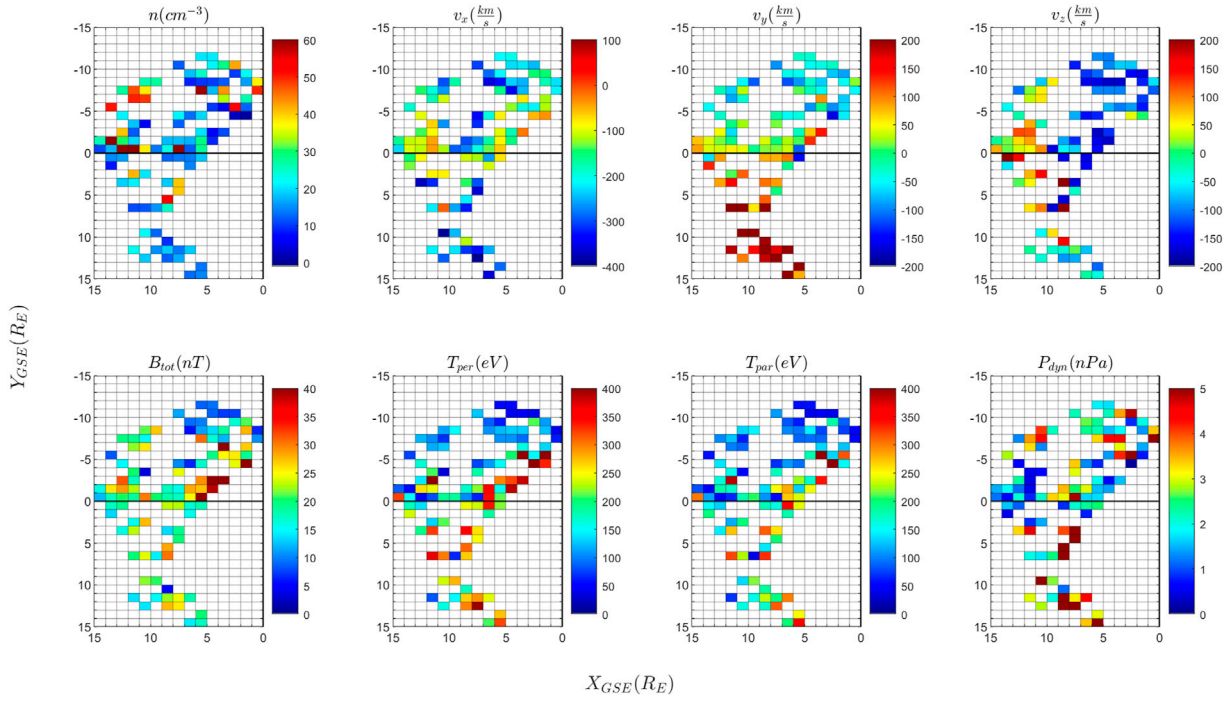


FIGURE 4
Summary of dynamical properties of magnetosheath jets observed by Cluster 3 between January and April 2007. The panels show the plasma density (n), the three components of the plasma bulk velocity (v_x,v_y,v_z), the magnetic field intensity (B_{tot}), the parallel and perpendicular temperature (T_{per}, T_{par}), the dynamic pressure (P_{dyn}). The value of the parameters is color coded as shown in the tables next to the panels; the values in each spatial bin are estimated as the median of the respective parameter for the ensemble of all jets detected in the respective bin. Only one value per jet is considered, namely, the one measured at the time when the dynamic pressure is maximum (indicated with stars in Figure 1).

average over 1 min of data preceding the detection of the jet; Q_2 is computed from an average of data over 30 s after the jet event.

The results are shown in Figures 4–7 on a uniform spatial grid formed of squared bins of one Earth radius size. In each spatial bin we plot, color coded, the median value of the respective variable estimated for the ensemble of jets (Figures 4, 6) and of the background magnetosheath perturbation (Figures 5, 7) detected in the respective bin. Only one value per jet is considered, namely,

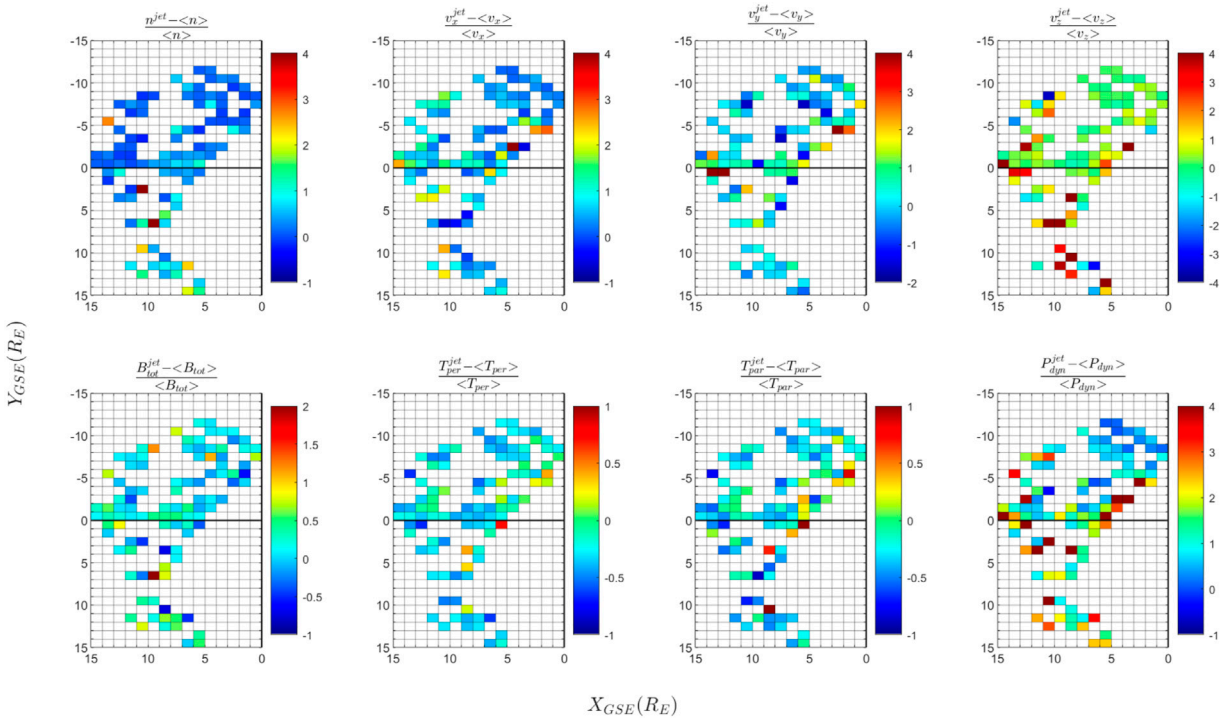


FIGURE 5
Relative perturbation of the background magnetosheath state produced by jets, for data recorded in 2007. We illustrate the same set of plasma variables as in Figure 4. The value of the relative perturbation is color coded as shown in the tables next to the panels. The perturbation of the variable Q is estimated as the ration $\frac{Q^{jet}-\langle Q\rangle}{\langle Q\rangle}$, where Q^{jet} is the value assigned to the jet and <Q> is the background value. <Q> is evaluated as the average between two time intervals, Δt1 equal to 1 min preceding the detection of the jet, and Δt2 equal to 30 s after the last point included in jet.

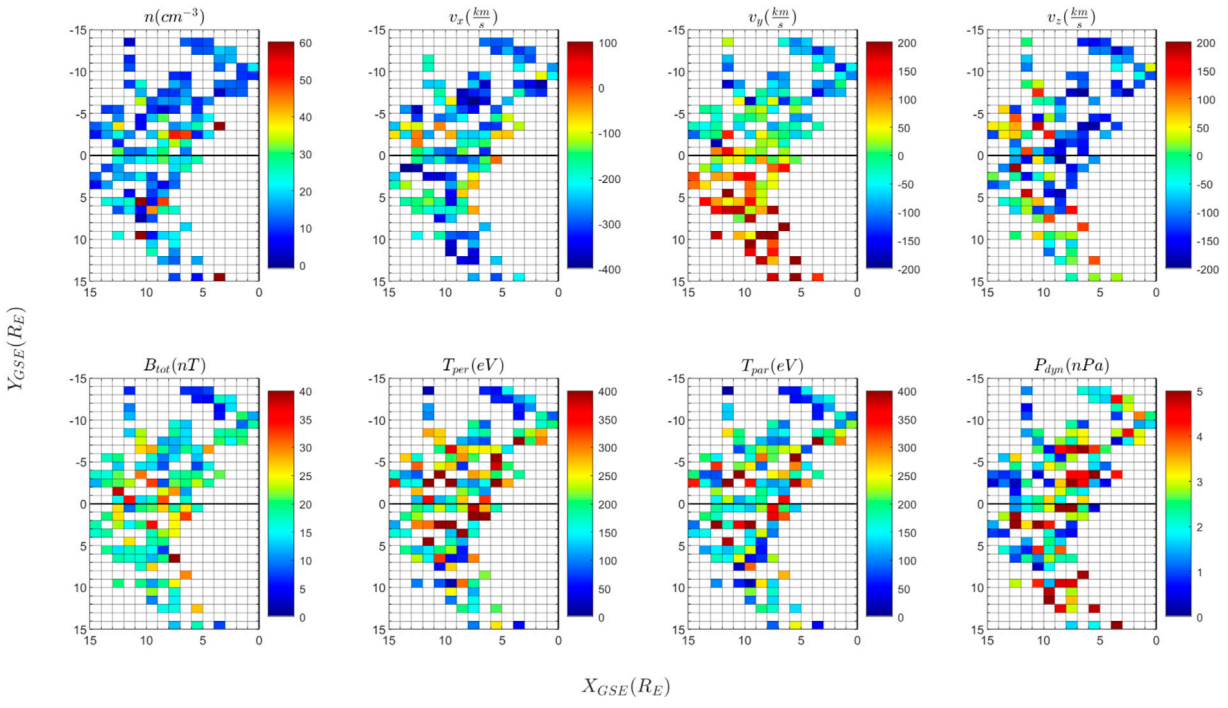


FIGURE 6
Summary of dynamical properties of magnetosheath jets observed by Cluster 3 in January–April 2008. Same format as Figure 4.

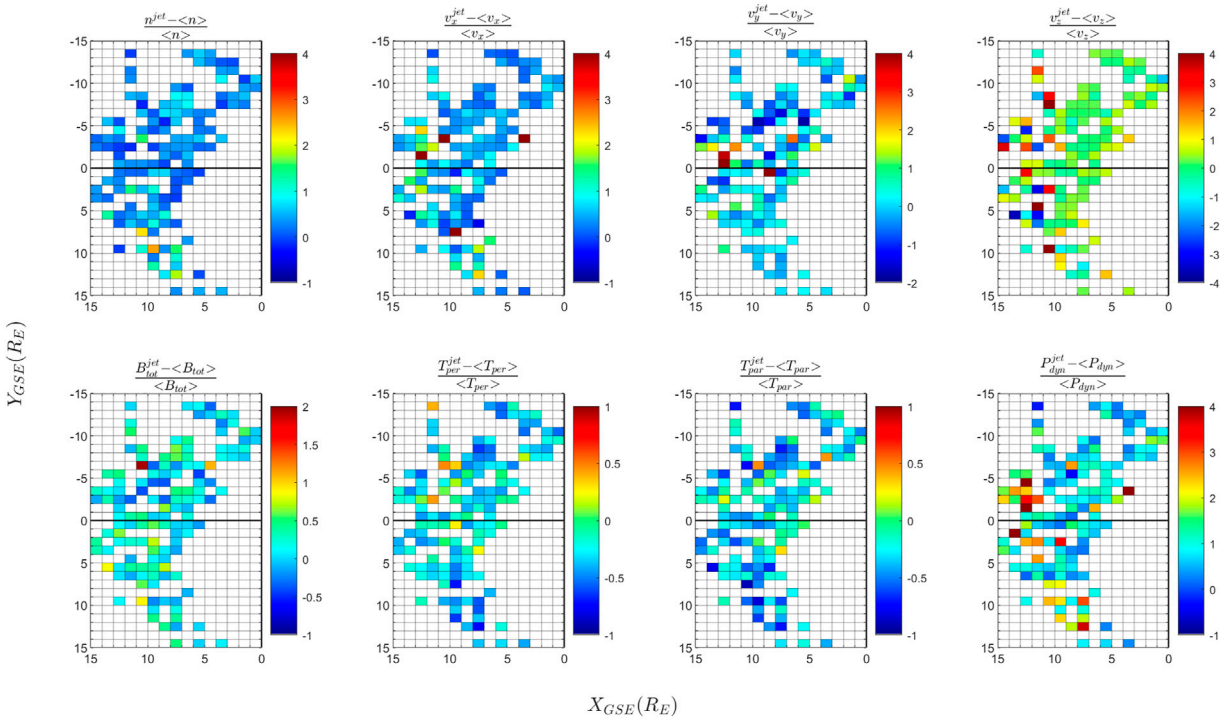


FIGURE 7
Relative perturbation of the background magnetosheath state produced by jets, illustrated for each physical quantity included in Figure 6, for data recorded in 2008. Same format as Figure 5.

the value recorded at the maximum of the dynamic pressure; the perturbation value is estimated as a time average, as described above.

The number of jets detected by Cluster-3 in the dawn flank is larger than at dusk. This asymmetry persists for the 2 years, 2007 and 2008. Nevertheless, the total number of observed jets is larger in 2008 than in 2007. The predominance of jets in the dawn flank is partly due to Cluster orbit, which intersects for slightly longer times the dawn magnetosheath flank during the time interval targeted by this study. However, in a recent study, Vuorinen et al. (2019) argue that a dawn-dusk asymmetry of the number of detected jets (with more jets detected in the dawn flank) should mostly be observed for oblique IMF. This effect is linked, according to Vuorinen et al. (2019), to a predominance of the quasi-parallel geometry in the dawn flank for oblique IMF. Radial and perpendicular orientation of the IMF seem to inhibit this relative asymmetry.

The distribution of the cone angle for our entire jets dataset is shown in Figure 8. One notes that in 2007 the distribution is skewed towards values larger than 60°. This effect is less present in 2008. Indeed, the cone angles in 2008 are more "ordered," their distribution is closer to the normal. This is probably due to the recurrent interaction with the high-speed streams and the corotating interaction regions observed in 2008 and which are characterized by a more "ordered" structure of the interplanetary magnetic field. In a recent study, Munteanu et al. (2019) showed that two systems of CIRs and HSS whose origin stayed stable on the Sun (same group of coronal holes) for several months in 2008, impacted the Earth, including for the time period investigated in this study (see, also, Negrea et al., 2021). While in 2007 the Earth was also impacted by HSS and CIRs their origin was more dynamic and their properties were therefore much more variable compared to 2008. A discussion of the driver properties for the jets included in this study is the topic of another paper, in preparation.

Figures 4, 6 also reveal asymmetries and spatial trends of jets plasma variables. The anti-sunward component of jet plasma bulk velocity, v_X, decreases when jets get closer to the magnetopause, for both years in the subsolar magnetosheath, for $|Y_{GSE}| < 5R_E$. One notes, however, that v_x takes larger values in 2008 than in 2007. Note also that, on average, higher speeds are recorded during jets observed in 2008 than in 2007. The spatial distribution of the other two components of jet velocity, v_y and v_z, follows the global circulation of magnetosheath plasma in the front side and flank sectors, with v_y being negative/positive in the dawn/dusk flanks. The jet plasma density shows an asymmetric dawn-dusk distribution. The density takes larger values in the dawn flank in 2007, while for 2008 this asymmetry is less pronounced. A dawn dusk asymmetry of the temperature is also observed for jets detected in 2007; the jets temperature takes larger values in the dawn flank than at dusk. However, this asymmetry is reduced for jets observed by Cluster 3 in 2008.

The properties of the background magnetosheath show similar asymmetries, although the pattern is slightly different compared to jets. Figures 5, 7 illustrate the relative perturbation produced by jets on the background magnetosheath state. We conclude these perturbations give a measure of the jet impact on the

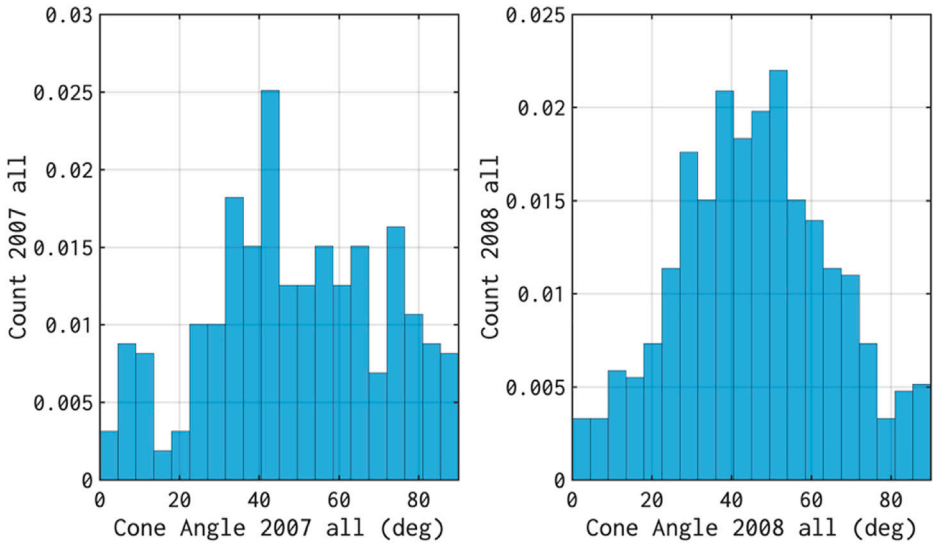


FIGURE 8
Normalized histograms of the IMF cone angle computed from OMNI data for time intervals when the jets were detected; left/right panels show results for 2007/2008.

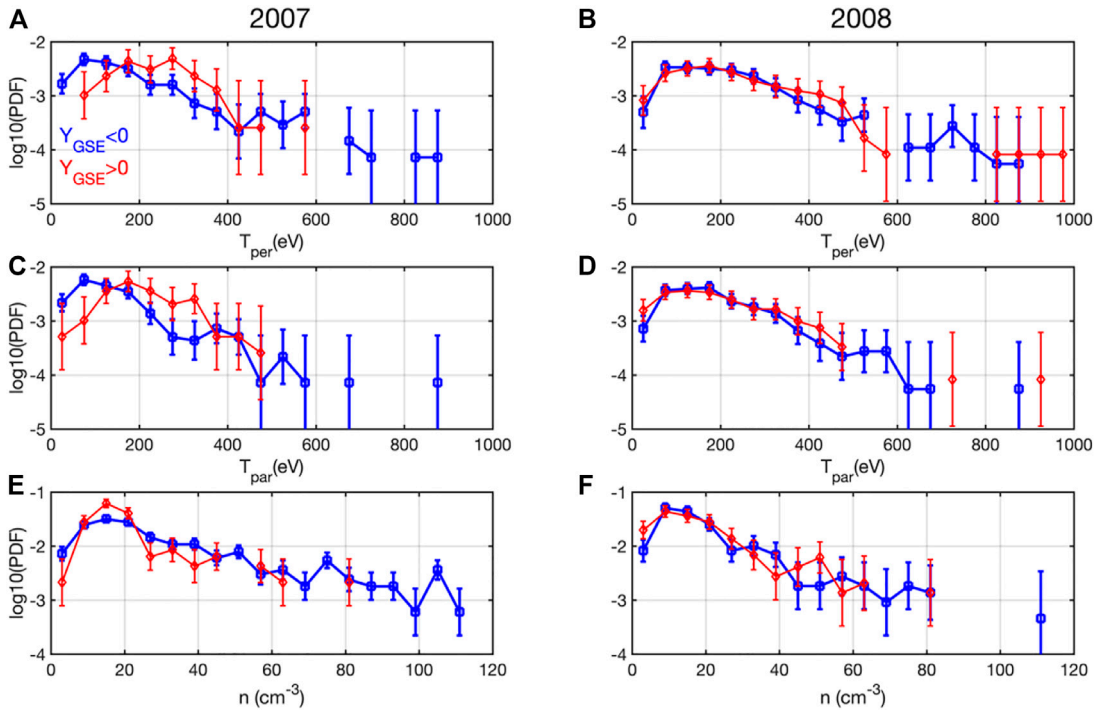


FIGURE 9
Histogram of the ion perpendicular [upper panels, **(A, B)**], parallel [lower panels, **(C, D)**] temperature and plasma density [lower panels, **(E, F)**] for 2007 (left) and 2008 (right). Blue/red profiles correspond to dawn (Y < 0)/dusk (Y > 0) flanks, respectively.

magnetosheath and also on the intrinsic jet asymmetries, not linked to the background magnetosheath asymmetric trends. We note the recurrence of increased density perturbations by jets is higher in the dusk than in the dawn, this tendency being more evident for data recorded in 2007. Following the plasmoid definition by Karlson et al. (2012), this asymmetry could mean that plasmoids are more likely encountered in the dusk sector. We also note that the density of jets tends to be larger in the dawn flank, however, the jet perturbation of the background magnetosheath density tends to be stronger in the dusk flank.

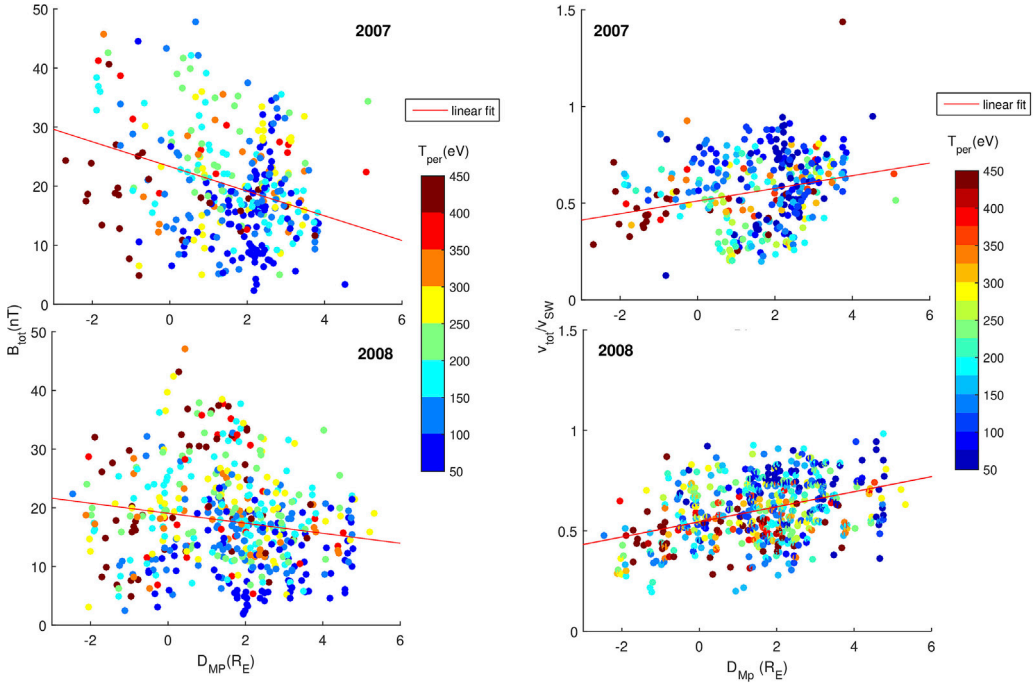


FIGURE 10
Distribution of ratio between jets speed and solar wind speed (v_{tot}/V_{SW}, right column) and magnetic field (B_{tot}, left column) as a function of the distance to the magnetopause ($DIST_{MP}$), for 2007 (top row), and 2008 (bottom row). The symbols identify the position of each individual jet, the color assigned to each symbol is a measure of the perpendicular temperature. $Dist_{MP}$ is the distance from the jet position to the magnetopause, estimated using the Shue et al. (1997) model (see text for details); negative values indicate the jet is detected downstream the model magnetopause. The red lines trace a linear regression over the ensemble of data in each set.

The strongest perturbation of the background magnetosheath temperature is observed in spatial bins located mostly in the dawn flank. However, one cannot assign a clear general dawn-dusk asymmetry for temperature perturbations. One notes that in 2007 the strongest perturbations of the temperature are observed at shorter distances from the Earth; this trend is not retrieved for 2008. The perturbation of the magnetosheath dynamic pressure by jets is asymmetric: the jet perturbations are stronger in the dusk flank (more visible in 2007 than in 2008).

A quantitative estimation of the asymmetries observed for jets temperature and density is provided by the probability density functions (normalized histograms) shown in Figure 9. We define three *ad hoc* temperature ranges: a) "cold," from 50 eV to 200 eV; b) "warm," from 200 eV to 400 eV and c) "hot," for more than 400 eV. The probabilities calculated for data measured in 2007 in the "cold" range are higher in the dawn than in the dusk flank; the probabilities calculated for the "warm" range are higher for the dusk flank. In the "hot" range the events for temperatures higher than 600 eV are detected only in the dawn flank but their statistic is poor. This trend in temperature asymmetry is not retrieved for 2008 data, except for a contracted "warm" range, between 375 eV and 475 eV, where the probabilities in the dusk flank are higher.

We also define three *ad hoc* density ranges for the probabilities computed in Figure 9: a) between 5 cm^{-3} and 20 cm^{-3}, b) between 25 cm^{-3} and 42 cm^{-3} and c) for densities higher than 50 cm^{-3}. In 2007 the probabilities for the lowest density range are slightly higher in the dusk than in the dawn flank. The probabilities in the intermediate density range are higher in the dawn than in the dusk. For densities higher than 50 cm^{-3} we do not see a significant dawn-dusk asymmetry. This trend is not retrieved in 2008, except for a contracted range between 45 and 55 cm^{-3} where the probabilities in the dusk are higher than in the dawn.

Thus, the jet properties recorded in 2007 suggest the jet inflow was denser and colder in the dawn than in the dusk flank. In 2008 this asymmetry is only partially retrieved, with jets being a bit warmer in the dusk side where they also show a slightly increased probability for higher densities. Interestingly, this asymmetry is partially consistent with the properties of the cold population of the cold dense plasma sheet, as reported by Wing et al. (2005), or Wing et al. (2014). Nevertheless, the formation of the cold dense plasma sheet is considered a signature of prolonged Northward IMF driving. Different mechanisms are believed to be involved in its formation (e.g., Kelvin-Helmholtz vortices in the flanks, Hasegawa et al., 2004). From an analysis of THEMIS magnetosheath data recorded between 2009 and 2015, Dimmock et al. (2016) find that the magnetosheath ion density shows a dawn-dusk asymmetry, with larger values at the dawn-side. The asymmetry detected for jets density is consistent with the trend found by Dimmock et al. (2016) and also the trends found in our background magnetosheath datasets.

Possible signatures for jet adiabatic breaking

The maps shown in Figure 10 illustrate the distribution of jet speed (normalized to the solar wind speed) and magnetic field

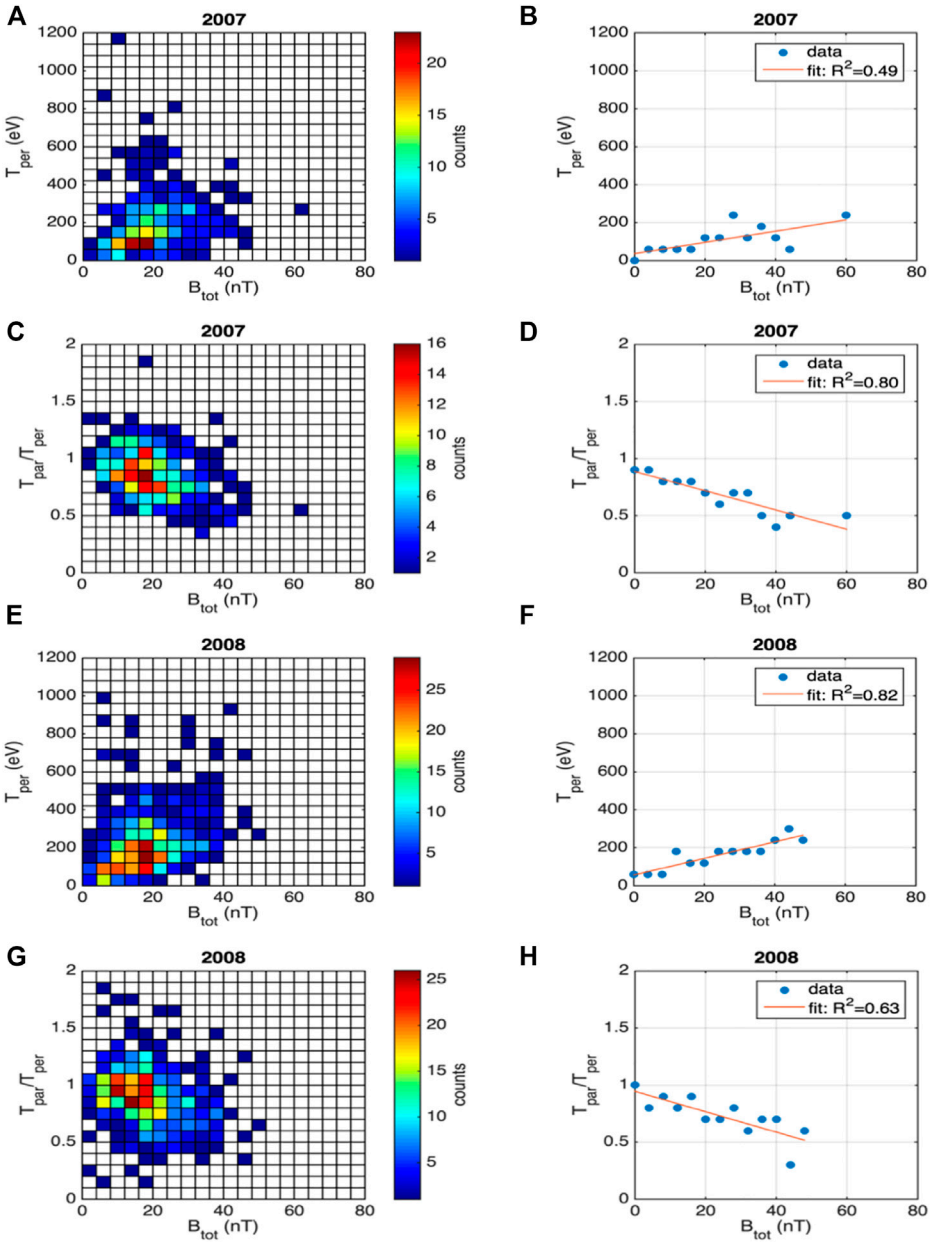


FIGURE 11
Left column: Histograms computed in the two-dimensional space defined by jet's perpendicular ion temperature, T_{per}, and magnetic field intensity, B_{tot}, **(A)** for 2007, and **(E)** for 2008 and by the jet's ion temperature anisotropy (T_{par}/T_{per}) and magnetic field in the jet **(C, G)**. Only one value per jet is considered, the one at maximum of the dynamic pressure (see Figure 1). Panels in the right column **(B, D, F, H)** show linear fits of the respective quantities as a function of the local value of the jet magnetic field intensity.

intensity as a function of the distance from the magnetopause and the perpendicular temperature of ions. The data suggest that the closer to the magnetopause is the jet, the more likely is that the jet speed decreases while the jet perpendicular temperature increases. This effect can be linked to an adiabatic breaking of the jet while it moves towards the magnetopause, in an increasing magnetic field.

Indeed, when the gradient of the magnetic field is smooth and the Alfven conditions are satisfied, the magnetic moment of ions is conserved leading to an increase of the kinetic perpendicular velocity, thus of the perpendicular temperature, while the jet ions move into an increasing magnetic field. On the other hand, collective effects linked to the electric polarization of the jet itself (Lemaire, 1985; Borovsky, 2021, see also Voitcu and Echim, 2016) lead to a conservation of the sum between the bulk forward energy and the thermal energy:

$$\frac{m_e + m_i}{2} V_{perp}^2 \nabla(x) + (\overline{\mu_i} + \overline{\mu_e}) B(x) = constant \qquad (4)$$

where $V_{perp}^2 \nabla(x)$ is the component of the jet bulk velocity in the direction of the magnetic field gradient, $\overline{\mu_i}$ and $\overline{\mu_e}$ are the magnetic

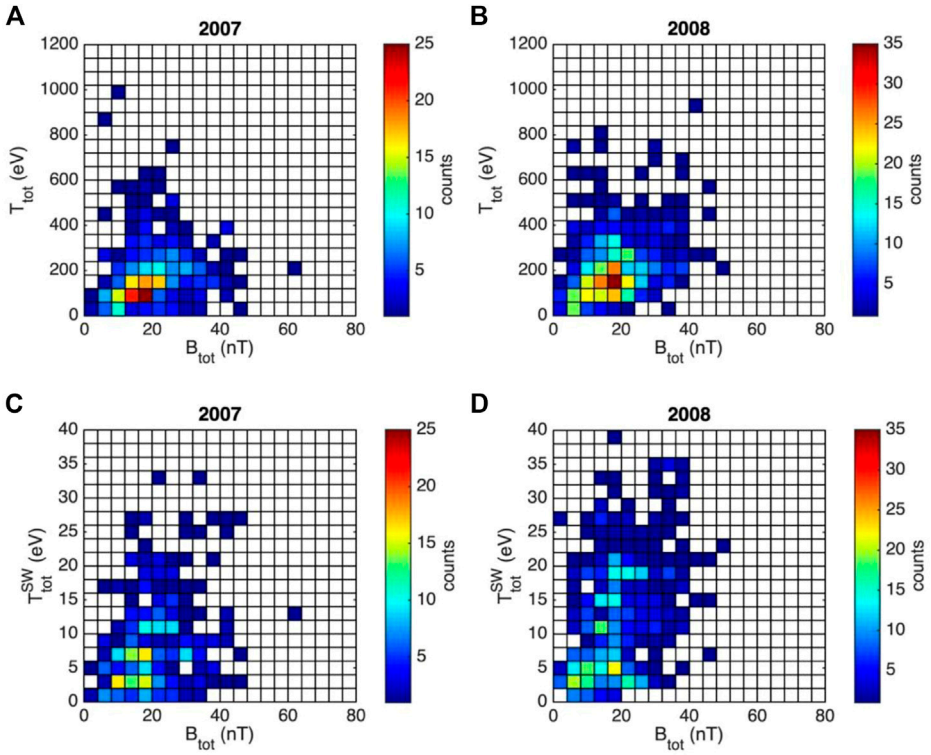


FIGURE 12
Upper row: 2D histograms in the space defined by the jet ion total temperature computed as $T = \frac{2}{3}T_{per} + \frac{1}{3}T_{par}$ and the jet magnetic field **(A)** for 2007 and **(B)** for 2008. Bottom row: 2D histograms defined in the space defined by the solar wind ion total temperature measured by ACE and the jet magnetic field **(C)** for 2007, and **(D)** for 2008. Only one value per jet is considered, the one at maximum of the dynamic pressure (see Figure 1).

moments derived for the thermal kinetic perpendicular speed of ions and electrons (see Lemaire, 1985; Voitcu and Echim, 2016).

To further investigate a possible signature for the perpendicular heating of jets, we compute the variation of the jet ion temperature anisotropy, the ratio between parallel and perpendicular temperature, $\frac{T_{par}}{T_{per}}$, as a function of the jet magnetic field intensity measured at the jet's location. The results are shown in Figure 11 together with the variation of the perpendicular ion temperature with the local magnetic field intensity. The data suggest that the perpendicular temperature/anisotropy increases/decreases with B_{tot}.

The observed correlation between the perpendicular temperature of jet ions with the jet magnetic field intensity, suggested by the results shown in Figure 11, can be a local effect, possibly linked to the adiabatic breaking of jets while advancing in an increasing magnetic field, or it can be imprinted in the solar wind and preserved during jet propagation in the magnetosheath. The latter scenario was investigated by assessing the correlation between the solar wind ion temperature and the solar wind magnetic field intensity, at the moment of jet detection. However, ACE, the solar wind monitoring spacecraft, provides measurement of the total ion temperature and not of the temperature anisotropy. Thus, we evaluated the correlation between the solar wind ion total temperature and the solar wind magnetic field intensity, for those time intervals when Cluster detected jets in the magnetosheath. In order to be consistent with the analysis performed for solar wind ion temperature data, we recomputed the correlation between the jet total ion temperature versus jet total magnetic field. These results are presented in Figure 12.

The two-dimensional histogram of jet total ion temperature versus jet total magnetic field intensity suggests that the number of observations of jets with higher ion temperatures increases with the magnetic field intensity. However, the histogram computed for solar wind ion temperature shows a larger spreading of data and the trend observed for jets is not retrieved. We consider that the increase of ion temperature with the magnetic field is a result of a local process in the magnetosheath and not inherited from the solar wind origin of jets.

A measure of the jet adiabatic breaking is given by ΔV, the amount of speed the jet loses while it advances in a magnetic field which increases by ΔB. This measure is, however, difficult to estimate experimentally as the spacecraft intersects the jet at a given position and time, with no access to jet's history, i.e., its speed at the origin, close to the bow-shock. In this study we use a proxy to estimate the jet deceleration taking place between the moment when the jet enters the magnetosheath (assumed to be at the bow-shock) and the moment when the jet is detected by Cluster. We estimate the difference between the jet velocity observed by Cluster in the magnetosheath and its initial perpendicular speed. The latter is not known of course, we approximate it by the value of the solar wind velocity at the bow shock (from OMNI database) at the time when the jet is detected.

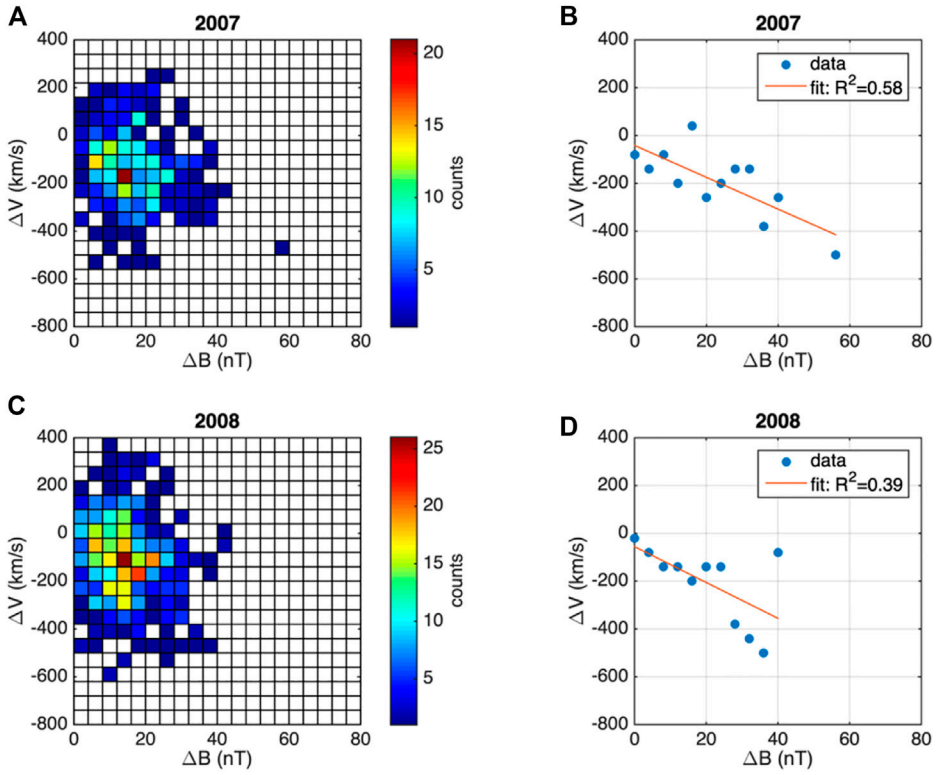


FIGURE 13
Left column: 2D histograms in the space defined by the jet perpendicular velocity breaking, quantified by the difference between the perpendicular jet speed at maximum dynamic pressure and the solar wind perpendicular speed, and total magnetic variation, ΔB, estimated as the difference between the jet magnetic field and the solar wind magnetic field, for 2007 **(A)** and 2008 **(C)**. Right column: Linear fit of ΔV function of ΔB, where ΔV is computed from the average over all bins in ΔV corresponding to each ΔB value in the 2D histogram at left for 2007 **(B)** and 2008 **(D)**. The results are obtained for a subset of jets in the subsolar magnetosheath $|Y_{GSE}| < 5R_E$.

Since only the perpendicular component is affected by the adiabatic braking (Lemaire, 1985; Voitcu and Echim, 2017), a measure of the braking/deceleration is given by $\Delta V = V_{jet} - V_{SW}$, where V_{jet}, V_{SW} denote the perpendicular component of the jet and solar wind plasma bulk velocity, respectively. Larger negative values of ΔV correspond to a stronger jet braking. Similarly, the increase in magnetic field is computed as the difference, $\Delta B = B_{jet} - B_{SW}$, between the magnetic field intensity measured in jet, at the moment of maximum dynamic pressure, and the magnetic field intensity measured in the solar wind at the moment of jet detection.

The two-dimensional histograms shown in Figure 13 illustrate the distribution of jets in the (ΔV, ΔB) plane, where ΔV is the variation of the perpendicular velocity estimated as the difference between the jet's perpendicular velocity when the jet dynamic pressure is maximum and the solar wind perpendicular speed at the moment of jet detection. The number of jets with large negative values of ΔV, corresponding to a strong deceleration/braking, increases with stronger magnetic field gradients, ΔB, for both years, 2007 and 2008. This correlation is consistent with the adiabatic breaking scenario. In other words, we observe that the jet perpendicular velocity decreases in an increasing magnetic field. This effect is consistent with the mechanism described in Eq. 4 leading to gradual transfer of jet's bulk speed into ion gyromotion and perpendicular heating.

One could question the contribution of the background magnetosheath on the dynamical properties of jets and braking revealed in Figures 11–13. There is certainly a breaking of the magnetosheath plasma, from the bow-shock to the magnetopause, which is a result of the general (magnetohydrodynamic) circulation flow of the shocked solar wind plasma surrounding the magnetospheric obstacle. The background magnetosheath plasma has a bulk velocity equal to zero at the magnetopause. However, this is not necessarily true for jets. A recent study based on MHD simulations and experimental data (Sibeck et al., 2022) suggests the anti-sunward component of the magnetosheath plasma bulk velocity decreases linearly from a quarter of the solar wind speed, at the bow-shock, to zero, at the magnetopause.

In Figure 14 we illustrate a 2D histogram of the background magnetosheath data in the (ΔV,ΔB) space, similar to Figure 13. The perpendicular velocity braking ΔV and maximum magnetic field variation, ΔB, are defined as for the jets. Generally, the distribution in (ΔV, ΔB) space look different for the background magnetosheath compared to the jets. Nevertheless, the linear regression shows better fitting parameters. Thus, one can argue that signatures of braking and correlation with magnetic gradient are retrieved also in the background magnetosheath. In a future study we will attempt to characterize in detail the magnetosheath braking and check against

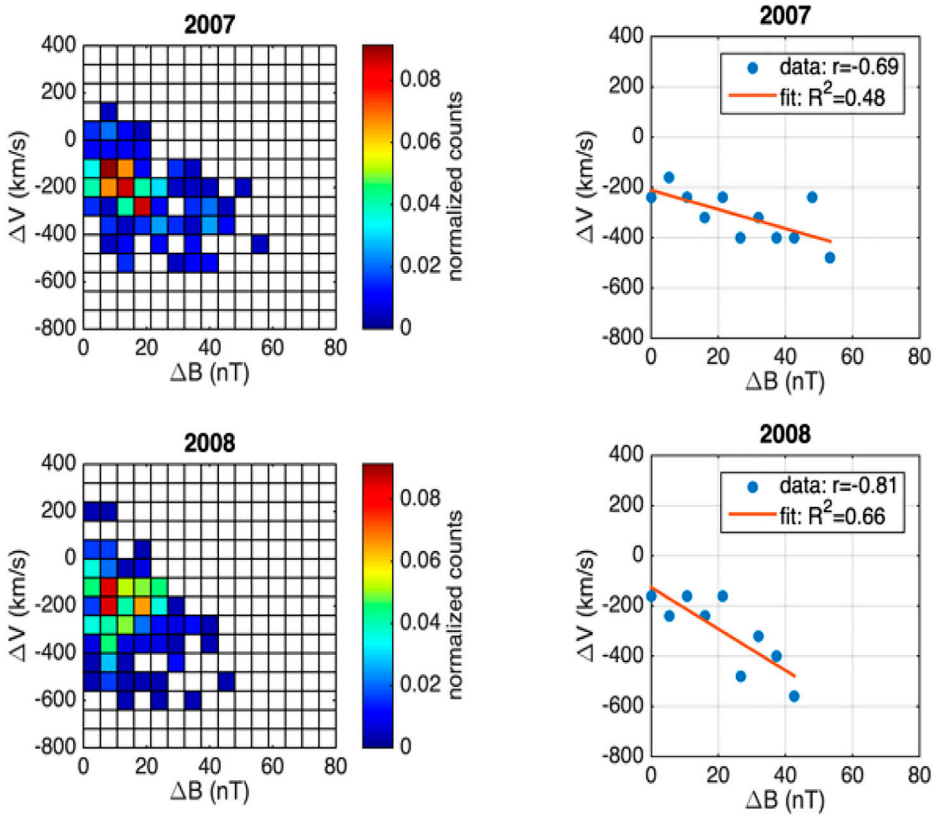


FIGURE 14
Same as Figure 13 but for the background magnetosheath data.

existing models, like, e.g., Sibeck et al. (2022), describing the decrease of background magnetosheath speed, from the bow-shock to the magnetopause.

Summary and perspectives

We discuss the dynamical properties of magnetosheath jets from the perspective of a phenomenology less considered in past studies devoted to this topic and supported by theory, numerical simulations and previous observational data, briefly reviewed in the introductory part of the manuscript. *In-situ* observations by fleet of spacecraft confirm that magnetosheath irregularities represent a quasi-permanent feature of the Earth's magnetosheath. Advocated in the past by theoretical models for the solar wind—magnetosphere interactions, magnetosheath irregularities/jets have an effective impact on the magnetopause and magnetosphere, like magnetic perturbations on ground, activation of auroral emissions, ionospheric effects.

Numerical simulations and models reveal that kinetic effects, like the electric self-polarization, have a role in the self-consistence of jets, implicitly assumed in many statistical studies based on *in-situ* observations. Collective plasma effects can also lead to a decoupling of jets from the background plasma and field, can foster the transport across the magnetopause, inside the magnetosphere. Such effects can also contribute to the deceleration of the jet bulk motion across an increasing magnetic field, like is the case in the vicinity of the magnetopause and/or inside the magnetosphere. The conservation of the adiabatic moment, the electric self-polarization of the jet, enable a decrease of the bulk motional energy perpendicular to the magnetic field, when the jet moves in and increasing magnetic field. The bulk motional energy is transferred into gyration energy. The jet forward motion stops where the magnetic field intensity is high enough such that the bulk motional energy is exhausted and fully transferred into gyration. *In-situ* jets observations by Cluster and THEMIS demonstrate the transport across the magnetopause, and the propagation well inside the magnetosphere (Dmitriev and Suvorova, 2015; Lyatsky et al., 2016b). Some of these penetrating jets loose completely their momentum inside the magnetosphere, become stagnant structures forming magnetospheric irregularities/cloud populated by magnetosheath plasma.

In this study we analyze observations of jets in the magnetosheath by Cluster 3 in 2007 and 2008, respectively. A number of 960 magnetosheath jets were identified by a searching procedure which detects the excess of the local dynamic pressure with respect to the background magnetosheath state. The latter is estimated from a running window average. The analysis provides insight on jet dynamical properties, relevant for the phenomenology discussed in first part of the manuscript. We find that:

1. The jet density, ion parallel and perpendicular temperature show a dawn-dusk asymmetry with a tendency of jet plasma being denser and colder in the dawn flank. This trend is found for data recorded in 2007 and is less clear for 2008. Nevertheless, the data analyzed for 2008 indicate a limited range of higher jet densities, 45–55 cm^{-3}, which has slightly higher probabilities in the dusk flank. The perturbation produced by jets on the background state is also asymmetric. The density perturbations produced by jets are stronger in the dusk than in the dawn flank, this tendency being more evident for data recorded in 2007. The strongest perturbations of the background magnetosheath temperature are detected in the dawn flank, generally closer to the magnetopause. However, one cannot assign a clear asymmetric trend for temperature perturbations produced by jets. The perturbations of the dynamical pressure are stronger in the dusk flank.
2. The jet speed decreases the closer the jets are detected to the Earth, in the subsolar magnetosheath. We find a correlation between jets perpendicular velocity decrease and the gradient of the magnetic field intensity, with a tendency of the perpendicular speed braking to increase with increasing magnetic field gradient. Such a behavior may be hallmark of an adiabatic breaking process.
3. The jet perpendicular ion temperature shows a tendency to increase with decreasing distance to the Earth. Data also show a correlation between jet perpendicular temperature and the local value of the magnetic field intensity, a possible signature of an adiabatic kinetic process converting forward bulk motion into perpendicular heating.

The tendencies described in conclusion 2 and 3 above are partially retrieved in background magnetosheath data. Future analyses on the same dataset will help us further discriminate between features pertaining exclusively to the dynamics of jets themselves and the background magnetosheath.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

ME: Conceptualization, data analysis, manuscript writing, project management MV: Conceptualization, data analysis, manuscript writing, project management CM: Conceptualization, data analysis, manuscript writing SC-B: Data analysis, manuscript writing ET: Conceptualization, data analysis, manuscript writing, project management ED: Conceptualization, data analysis GV: Conceptualization, data analysis, manuscript writing, project management CN: data analysis, manuscript writing.

Funding

The VESS project PN-III-P1-1.2-PCCDI-2017 0371, 18PCCDI/2018, funded by the Romanian Ministry of Research and Innovation, CCCDI-UEFISCDI, the PRODEX projects PEA4000134960 (MISION) and PEA4000130552 (CON-CLUSTER-III) funded by the European Space Agency, the project BRAIN-BE 2.0 B2//223/P1/PLATINUM funded by the Belgian Office for Research (BELSPO) are acknowledged for supporting this research. Partial funding was provided by the Romanian Ministry of Research, Innovation and Digitalization under Romanian National Core Program LAPLAS VII – contract no. 30N/2023 and the Belgian Solar Terrestrial Center of Excellence. CM, GV, ET, and CN acknowledge support from the PRODEX project MISION PEA. ME acknowledges support from the PRODEX project CLUSTER PEA and the Belgian Solar Terrestrial Center of Excellence (STCE).

Acknowledgments

The AMDA database, the ESA Cluster mission, the UK Cluster Data Centre and the SILS database are deeply thanked for their efforts and progress in science and technology and extended data provided. We used the OMNI definitive 1-min IMF and plasma data time-shifted to the nose of the Earth's bow shock (https://omniweb.gsfc.nasa.gov/html/HROdocum.html), downloaded from https://cdaweb.gsfc.nasa.gov. The European FP7 project STORM (http://www.storm-fp7.eu) is acknowledged for providing the time intervals when Cluster 3 crossed the magnetosheath.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amata, E., Savin, S. P., Ambrosino, D., Bogdanova, Y. V., Marcucci, M. F., Romanov, S., et al. (2011). High kinetic energy density jets in the Earth's magnetosheath: A case study. *Planet. Space Sci.* 59, 482–494. ISSN 0032-0633. doi:10.1016/j.pss.2010.07.021

Archer, M. O., Hietala, H., Hartinger, M. D., Plaschke, F., and Angelopoulos, V. (2019). Direct observations of a surface eigenmode of the dayside magnetopause. *Nat. Commun.* 10, 615. doi:10.1038/s41467-018-08134-5

Archer, M. O., Horbury, T. S., and Eastwood, J. P. (2012). Magnetosheath pressure pulses: Generation downstream of the bow shock from solar wind discontinuities. *J. Geophys. Res.* 117, A05228. doi:10.1029/2011ja017468

Archer, M. O., and Horbury, T. S. (2013). Magnetosheath dynamic pressure enhancements: Occurrence and typical properties. *Ann. Geophys.* 31, 319–331. doi:10.5194/angeo-31-319-2013

Balogh, A., Dunlop, M. W., Cowley, S. W. H., Southwood, D. J., Thomlinson, J. G., Glassmeier, K. H., et al. (1997). The cluster magnetic field investigation. *Space Sci. Rev.* 79 (1/2), 65–91. doi:10.1023/a:1004970907748

Borovsky, J. E., Birn, J., Echim, M. M., Fujita, S., Lysak, R. L., Knudsen, D. J., et al. (2020). Quiescent discrete auroral arcs: A review of magnetospheric generator mechanisms. *Space Sci. Rev.* 216, 1. doi:10.1007/s11214-019-0619-5

Borovsky, J. E. (2021). "Is our understanding of solar-wind/magnetosphere coupling satisfactory ?," in *Frontiers in Astronomy and space sciences* 8, 634073. doi:10.3389/fspas.2021.634073

Bostick, W. H. (1956). Experimental study of ionized matter projected across a magnetic field. *Phys. Rev.* 104, 292–299. doi:10.1103/physrev.104.292

Burgess, D., and Scholer, M. (2013). Microphysics of quasi-parallel shocks in collisionless plasmas. *Space Sci. Rev.* 178 (2-4), 513–533. doi:10.1007/s11214-013-9969-6

Dimmock, A. P., Pulkkinen, T. I., Osmane, A., and Nykyri, K. (2016). The dawn–dusk asymmetry of ion density in the dayside magnetosheath and its annual variability measured by THEMIS. *Ann. Geophys.* 34, 511–528. doi:10.5194/angeo-34-511-2016

Dmitriev, A. V., Lalchand, B., and Ghosh, S. (2021). Mechanisms and evolution of geoeffective large-scale plasma jets in the magnetosheath. *Universe* 7, 152. doi:10.3390/universe7050152

Dmitriev, A. V., and Suvorova, A. V. (2015). Large-scale jets in the magnetosheath and plasma penetration across the magnetopause: THEMIS observations. *J. Geophys. Res. Space Phys* 120, 4423–4437. doi:10.1002/2014JA020953

Echim, M., Chang, T., Kovacs, P., Wawrzaszek, A., Yordanova, E., Narita, Y., et al. (2021). "Turbulence and complexity of magnetospheric plasmas," in *Magnetospheres in the solar system*. Editors R. Maggiolo, N. André, H. Hasegawa, D. T. Welling, Y. Zhang, and L. J. Paxton Hoboken, NJ: Wiley-American Geophysical Union doi:10.1002/9781119815624

Echim, M., Lemaire, J. F., and Roth, M. (2005). Self-consistent solution for a collisionless plasma slab in motion across a magnetic field. *Phys. Plasmas* 12, 072904. doi:10.1063/1.1943848

Echim, M., and Lemaire, J. (2000). Laboratory and numerical simulations of the impulsive penetration mechanism. *Space Sci. Rev.* 92, 565–601. doi:10.1023/a:1005264212972

Echim, M., and Lemaire, J. (2002). Positive density gradients at the magnetopause: Interpretation in the framework of the impulsive penetration mechanism. *J. Atmos. Solar-Terrestrial Phys.* 64 (18), 2019–2028. doi:10.1016/s1364-6826(02)00229-8

Echim, M., and Lemaire, J. (2005). Two-dimensional Vlasov solution for a collisionless plasma jet across transverse magnetic field lines with a sheared bulk velocity. *Phys. Rev. E* 72, 036405. doi:10.1103/physreve.72.036405

Echim, M. M., Maggiolo, R., Roth, M., and De Keyser, J. (2009). A magnetospheric generator driving ion and electron acceleration and electric currents in a discrete auroral arc observed by Cluster and DMSP. *Geophys. Res. Lett.* 36, L12111. doi:10.1029/2009GL038343

Echim, M. M., Roth, M., and De Keyser, J. (2008). Ionospheric feedback effects on the quasi-stationary coupling between LLBL and postnoon/evening discrete auroral arcs. *Ann. Geophys.* 26, 913–928. doi:10.5194/angeo-26-913-2008

Echim, M. M., Roth, M., and De Keyser, J. (2007). Sheared magnetospheric plasma flows and discrete auroral arcs: A quasi-static coupling model. *Ann. Geophys.* 25, 317–330. doi:10.5194/angeo-25-317-2007

Escoubet, C. P., Hwang, K-J., Toledo-Redondo, S., Turc, L., Haaland, S. E., Aunai, N., et al. (2020). Cluster and MMS simultaneous observations of magnetosheath high speed jets and their impact on the magnetopause. *Front. Astron. Space Sci.* 6, 78. doi:10.3389/fspas.2019.00078

Escoubet, C. P., Schmidt, R., and Goldstein, M. (1997). Cluster - science and mission Overview. *Space Sci. Rev.* 79, 11–32. doi:10.1023/A:1004923124586

Farris, M. H., and Russell, C. T. (1994). Determining the standoff distance of the bow shock: Mach number dependence and use of models. *J. Geophys. Res.* 99 (A9), 17681–17689. doi:10.1029/94JA01020

Génot, V., Budnik, E., Jacquey, C., Bouchemit, M., Renard, B., Dufourg, N., et al. (2021). Automated Multi-Dataset Analysis (AMDA): An on-line database and analysis tool Automated Multi-Dataset Analysis (AMDA): An on-line database and analysis tool for heliospheric and planetary plasma data for heliospheric and planetary plasma data. *Planet. Space Sci.* 201, 105214. doi:10.1016/j.pss.2021.105214

Goncharov, O., Gunell, H., Hamrin, M., and Chong, S. (2020). Evolution of high-speed jets and plasmoids downstream of the quasi-perpendicular bow shock. *J. Geophys. Res. Space Phys.* 125, e2019JA027667. doi:10.1029/2019JA027667

Gunell, H., Nilsson, H., Stenberg, G., Hamrin, M., Karlsson, T., Maggiolo, R., et al. (2012). Plasma penetration of the dayside magnetopause. *Phys. Plasmas* 19 (072), 072906. doi:10.1063/1.4739446

Gustafsson, G., Bostrom, R., Holback, B., Holmgren, G., Lundgren, A., Stasiewicz, K., et al. (1997). The electric field and wave experiment for the cluster mission. *Space Sci. Rev.* 79 (1/2), 137–156. doi:10.1023/a:1004975108657

Han, D.-S., Hietala, H., Chen, X.-C., Nishimura, Y., Lyons, L. R., Liu, J.-J., et al. (2017). Observational properties of dayside throat aurora and implications on the possible generation mechanisms. *J. Geophys. Res. Space Phys.* 122, 1853–1870. doi:10.1002/2016ja023394

Han, D.-S., Liu, J.-J., Chen, X.-C., Xu, T., Li, B., Hu, Z.-J., et al. (2018). Direct evidence for throat aurora being the ionospheric signature of magnetopause transient and reflecting localized magnetopause indentations. *J. Geophys. Res. Space Phys.* 123, 2658–2667. doi:10.1002/2017JA024945

Hasegawa, H., Fujimoto, M., Phan, T.-D., Reme, H., Balogh, A., Dunlop, M. W., et al. (2004). Transport of solar wind into Earth's, magnetosphere through rolled-up, Kelvin–Helmholtz vortices. *NATURE* 430, 755–758. doi:10.1038/nature02799

Hietala, H., Laitinen, T. V., Andréeová, K., Vainio, R., Vaivads, A., Palmroth, M., et al. (2009). Supermagnetosonic jets behind a collisionless quasiparallel shock. *Phys. Rev. Lett.* 103, 245001. doi:10.1103/physrevlett.103.245001

Hietala, H., Partamies, N., Laitinen, T. V., Clausen, L. B. N., Facskó, G., Vaivads, A., et al. (2012). Supermagnetosonic subsolar magnetosheath jets and their effects: From the solar wind to the ionospheric convection. *Ann. Geophys.* 30, 33–48. doi:10.5194/angeo-30-33-2012

Johnson, J. R., and Wing, S. (2015). The dependence of the strength and thickness of field-aligned currents on solar wind and ionospheric parameters. *J. Geophys. Res. Space Phys.* 120, 3987–4008. doi:10.1002/2014JA020312

Karlsson, T., Brenning, N., Nilsson, H., Trotignon, J.-G., Vallières, X., and Facsko, G. (2012). Localized density enhancements in the magnetosheath: Three-dimensional morphology and possible importance for impulsive penetration. *J. Geophys. Res.* 117, A03227. doi:10.1029/2011JA017059

Karlsson, T., Kullen, A., Liljeblad, E., Brenning, N., Nilsson, H., Gunell, H., et al. (2015). On the origin of magnetosheath plasmoids and their relation to magnetosheath jets. *J. Geophys. Res. Space Phys.* 120, 7390–7403. doi:10.1002/2015JA021487

Karlsson, T., Plaschke, F., Hietala, H., Archer, M., Blanco-Cano, X., Kajdič, P., et al.(2018). Investigating the anatomy of magnetosheath jets – MMS observations. *Ann. Geophys.* 36, 655–677. doi:10.5194/angeo-36-655-2018

Katsavriasi, C., Raptis, S., Daglis, L. A., Karlsson, T., Georgiou, M., and Balasis, G. (2021). On the generation of Pi2 pulsations due to plasma flow patterns around magnetosheath jets. *Geophys. Res. Lett.* 48, e2021GL093611. doi:10.1029/2021GL093611

Koller, F., Temmer, M., Preisser, L., Plaschke, F., Geyer, P., Jian, L. K., et al. (2022). Magnetosheath jet occurrence rate in relation to CMEs and SIRs. *J. Geophys. Res. Space Phys.* 127, e2021JA030124. doi:10.1029/2021JA030124

Lai, H. R., and Russell, C. T. (2018). Nanodust released in interplanetary collisions. *Planet. Space Sci.* 156, 2–6. doi:10.1016/j.pss.2017.10.003

Lemaire, J. (1977). Impulsive penetration of filamentary plasma elements into the magnetospheres of the Earth and Jupiter. *Planet. Space Sci.* 25, 887–890. doi:10.1016/0032-0633(77)90042-3

Lemaire, J. (1985). Plasmoid motion across a tangential discontinuity - with application to the magnetopause. *J. Plasma Phys.* 33, 425–436. doi:10.1017/s0022377800002592

Lemaire, J., and Roth, M. (1981). Differences between solar wind plasmoids and ideal magnetohydrodynamic filaments. *Planet. Space Sci.* 29, 843–849. doi:10.1016/0032-0633(81)90075-1

Lemaire, J., and Roth, M. (1978). Penetration of solar wind plasma elements into the magnetosphere. *J. Atmos. Terr. Phys.* 40, 331–335. doi:10.1016/0021-9169(78)90049-1

Liu, T. Z., Hietala, H., Angelopoulos, V., Omelchenko, Y., Vainio, R., and Plaschke, F. (2020). Statistical study of magnetosheath jet-driven bow waves. *J. Geophys. Res. Space Phys.* 125, e2019JA027710. doi:10.1029/2019JA027710

Lundin, R., and Aparicio, B. (1982). Observations of penetrated solar wind plasma elements in the plasma mantle. *Planet. Space Sci.* 30, 81–91. doi:10.1016/0032-0633(82)90075-7

Lundin, R., and Dubinin, E. (1984). Solar wind energy transfer regions inside the dayside magnetopause. I - evidence for magnetosheath plasma penetration. *Planet. Space Sci.* 32, 745–755. doi:10.1016/0032-0633(84)90098-9

Lundin, R. (1988). On the magnetospheric boundary layer and solar wind energy transfer into the magnetosphere. *Space Sci. Rev. (ISSN 0038-6308* 48 (3-4), 263–320. doi:10.1007/bf00226010

Lundin, R., Sauvaud, J.-A., Rème, H., Balogh, A., Dandouras, I., Bosqued, J. M., et al. (2003). Evidence for impulsive solar wind plasma penetration through the dayside magnetopause. *Ann. Geophys.* 21, 457–472. doi:10.5194/angeo-21-457-2003

Lyatsky, W., Pollock, C., Goldstein, M. L., Lyatskaya, S., and Avanov, L. (2016a). Penetration of magnetosheath plasma into dayside magnetosphere: 1. Density, velocity, and rotation. *J. Geophys. Res. Space Phys.* 121, 7699–7712. doi:10.1002/2015JA022119

Lyatsky, W., Pollock, C., Goldstein, M. L., Lyatskaya, S., and Avanov, L. (2016b). Penetration of magnetosheath plasma into dayside magnetosphere: 2. Magnetic field in

plasma filaments. *J. Geophys. Res. Space Phys.* 121, 7713–7727. doi:10.1002/2015JA022120

Munteanu, C., Hamada, A., and Mursula, K. (2019). High-speed solar wind streams in 2007-2008: Turning on the Russell-McPherron effect. *J. Geophys. Res. Space Phys.* 124, 8913–8927. doi:10.1029/2019JA026846

Negrea, C., Munteanu, C., and Echim, M. M. (2021). Global ionospheric response to a periodic sequence of HSS/CIR events during the 2007–2008 solar minimum. *J. Geophys. Res. Space Phys.* 126, e2020JA029071. doi:10.1029/2020JA029071

Nemecek, Z., Šafránková, J. M., Prech, L., Sibeck, D. G., Kokubun, S., Mukai, T., et al. (1998). Transient flux enhancements in the magnetosheath. *Geophys. Res. Lett.* 25, 1273–1276. doi:10.1029/98GL50873

Nykyri, K. (2013). Impact of MHD shock physics on magnetosheath asymmetry and Kelvin-Helmholtz instability. *J. Geophys. Res. Space Phys.* 118 (8), 5068–5081. doi:10.1002/jgra.50499

Nykyri, K., and Otto, A. (2001). Plasma transport at the magnetospheric boundary due to reconnection in Kelvin-Helmholtz vortices. *Geophys. Res. Lett.* 28, 3565–3568. doi:10.1029/2001GL013239

Parkhomov, V. A., Eselevich, V. G., Eselevich, M. V., Dmitriev, A. V., Suvorova, A. V., Khomutov, S. Yu., et al. (2021). Magnetospheric response to the interaction with the sporadic solar wind diamagnetic structure. *Solar-Terr. Phys.* 7 (3), 11–28. doi:10.12737/stp-73202102

Parkhomov, V. A., Eselevich, V. G., Eselevich, M. V., Tsegmed, B., Khomutov, S. Yu., Raita, T., et al. (2022). Correspondence of a global isolated substorm to the McPherron statistical model. *Solar-Terrestrial Phys.* 8 (1), 37–46. doi:10.12737/stp-82202206

Plaschke, F., Hietala, H., and Angelopoulos, V. (2013). Anti-sunward high-speed jets in the subsolar magnetosheath. *Ann. Geophys.* 31 (10), 1877–1889. doi:10.5194/angeo-31-1877-2013

Plaschke, F., Hietala, H., Angelopoulos, V., and Nakamura, R. (2016). Geoeffective jets impacting the magnetopause are very common. *J. Geophys. Res. Space Phys.* 121, 3240–3253. doi:10.1002/2016JA022534

Plaschke, F., Hietala, H., Archer, M., Blanco-Cano, X., Kajdiõc, P., Karlsson, T., et al. (2018). Jets downstream of collisionless shocks. *Space Sci. Rev.* 214, 81. doi:10.1007/s11214-018-0516-3

Plaschke, F., Jernej, M., Hietala, H., and Vuorinen, L. (2020). On the alignment of velocity and magnetic fields within magnetosheath jets. *Ann. Geophys.* 38, 287–296. doi:10.5194/angeo-38-287-2020

Plaschke, F., Karlsson, T., Hietala, H., Archer, M., Voros, Z., Nakamura, R., et al. (2017). Magnetosheath high-speed jets: Internal structure and interaction with ambient plasma. *J. Geophys. Res. Space Phys.* 122 (10), 175. 157–10. doi:10.1002/2017JA024471

Preisser, L. (2020). Magnetosheath jets and plasmoids: Characteristics and formation mechanisms from hybrid simulations. *Astrophysical J. Lett.* 900, abad2b. doi:10.3847/2041-8213/abad2b

Raptis, S., Aminalragia-Giamini, S., Karlsson, T., and Lindberg, M. (2020). Classification of magnetosheath jets using neural networks and high resolution OMNI(HRO) data. *Front. Astronomy Space Sci.* 7, 24. doi:10.3389/fspas.2020.00024

Raptis, S., Karlsson, T., Plaschke, F., Kullen, A., and Lindqvist, P.-A. L. (2019). Classifying magnetosheath jets using MMS - statistical properties. *ESSOAr*. doi:10.1002/essoar.10501493.2

Reme, H. (1997). "The cluster ion spectrometry (CIS) experiment," in *The cluster and phoenix missions*. Editors C. P. Escoubet, C. T. Russell, and R. Schmidt (Dordrecht: Springer). doi:10.1007/978-94-011-5666-012

Roth, M. (1995). "Impulsive transport of solar wind into the magnetosphere in Physics of the magnetopause," in *Geophysical monograph*. Editors P. Song, B. U. Ö. Sonnerup, and M. F. Thomsen (Washington, D. C: AGU), 90, 343–348.

Savin, S., Amata, E., Zelenyi, L., Budaev, V., Consolini, G., Treumann, R., et al. (2008). High energy jets in the Earth's magnetosheath: Implications for plasma dynamics and anomalous transport. *JETP Lett.* 87, 593–599. doi:10.1134/S0021364008110015

Schmidt, G. (1960). Plasma motion across magnetic fields. *Phys. Fluids* 3, 961–965. doi:10.1063/1.1706163

Shue, J.-H., Chao, J.-K., Song, P., McFadden, J. P., Suvorova, A., Angelopoulos, V., et al. (2009). Anomalous magnetosheath flows and distorted subsolar magnetopause for radial interplanetary magnetic fields. *Geophys. Res. Lett.* 36, L18112. doi:10.1029/2009GL039842

Shue, J.-H., Chao, J. K., Fu, H. C., Russell, C. T., Song, P., Khurana, K. K., et al. (1997). A new functional form to study the solar wind control of the magnetopause size and shape. *J. Geophys. Res.* 102 (A5), 9497–9511. doi:10.1029/97JA00196

Sibeck, D. G., Silveira, M. V. D., and Collier, M. R. (2022). Tracking the Subsolar Bow Shock and Magnetopause. *Journal of Geophysical Research: Space Physics* 127 (9), e30704. doi:10.1029/2022JA030704

Teodorescu, E., Echim, M., and Voitcu, G. (2021). A perspective on the scaling of magnetosheath turbulence and effects of bow shock properties. *Astrophysical J.* 910, 13. doi:10.3847/1538-4357/abe12d

Vaisberg, O. L., Smirnov, V. N., Avanov, L. A., Waite, J. H., Jr., Burch, J. L., Russell, C. T., et al. (1998). Observation of isolated structures of the low latitude boundary layer with the INTERBALL/tail probe. *Geophys. Res. Lett.* 25 (23), 4305–4308. doi:10.1029/1998GL900167

Voitcu, G., and Echim, M. (2018). Crescent-shaped electron velocity distribution functions formed at the edges of plasma jets interacting with a tangential discontinuity. *Ann. Geophys.* 36, 1521–1535. doi:10.5194/angeo-36-1521-2018

Voitcu, G., and Echim, M. (2017). Tangential deflection and formation of counterstreaming flows at the impact of a plasma jet on a tangential discontinuity. *Geophys. Res. Lett.* 44, 5920–5927. doi:10.1002/2017GL073763

Voitcu, G., and Echim, M. (2016). Transport and entry of plasma clouds/jets across transverse magnetic discontinuities: Three-dimensional electromagnetic particle-in-cell simulations. *J. Geophys. Res. Space Phys.* 121, 4343–4361. doi:10.1002/2015JA021973

Vörös, Z., Yordanova, E., Echim, M., Consolini, G., and Narita, Y., TURBULENCE-GENERATED proton-scale structures in the terrestrial magnetosheath, *Astrophysical J. Lett.*, 819 L15, doi:10.3847/2041-8205/819/1/L152019

Vuorinen, L., Hietala, H., and Plaschke, F. (2019). Jets in the magnetosheath: IMF control of where they occur. *Ann. Geophys.* 37, 689–697. doi:10.5194/angeo-37-689-2019

Wang, B., Nishimura, Y., Hietala, H., Lyons, L., Angelopoulos, V., Plaschke, F., et al. (2018). Impacts of magnetosheath high-speed jets on the magnetosphere and ionosphere measured by optical imaging and satellite observations. *J. Geophys. Res.-Space* 123, 4879–4894. doi:10.1029/2017JA024954

Wing, S., Johnson, J. R., Chaston, C. C., Echim, M., Escoubet, C. P., Lavraud, B., et al. (2014). Review of solar wind entry into and transport within the plasma sheet. *Space Sci. Rev.* 184, 33–86. doi:10.1007/s11214-014-0108-9

Wing, S., Johnson, J. R., Newell, P. T., and Meng, C.-I. (2005). Dawn-dusk asymmetries, ion spectra, and sources in the northward interplanetary magnetic field plasma sheet. *J. Geophys. Res.* 110, A08205. doi:10.1029/2005JA011086

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Mini Review
PUBLISHED 06 March 2023
DOI 10.3389/fspas.2023.1121230

Check for updates

OPEN ACCESS

EDITED BY
Olga V. Khabarova,
Tel Aviv University, Israel

REVIEWED BY
Laxman Adhikari,
University of Alabama in Huntsville, United States
Olga Gutynska,
Charles University, Czechia

*CORRESPONDENCE
Liudmila Rakhmanova,
rakhlud@gmail.com

SPECIALTY SECTION
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 11 December 2022
ACCEPTED 10 February 2023
PUBLISHED 06 March 2023

CITATION
Rakhmanova L, Riazantseva M, Zastenker G and Yermolaev Y (2023), Role of the variable solar wind in the dynamics of small-scale magnetosheath structures. *Front. Astron. Space Sci.* 10:1121230. doi: 10.3389/fspas.2023.1121230

COPYRIGHT
© 2023 Rakhmanova, Riazantseva, Zastenker and Yermolaev. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Role of the variable solar wind in the dynamics of small-scale magnetosheath structures

Liudmila Rakhmanova*, Maria Riazantseva, Georgy Zastenker and Yuri Yermolaev

Space Research Institute, Russian Academy of Sciences, Moscow, Russia

Varying solar wind plays a crucial role in the processes inside the magnetosphere. Statistical studies generally reveal the geoeffectiveness of solar wind streams of different origins and types, characterized by various parameters such as dynamic pressure and magnetic field orientation. However, the predictions of the space weather are still not completely reliable. Small-scale structures (observed as high-amplitude variations with frequencies above 0.01 Hz) involved in the turbulent flow of the solar wind and in the magnetosheath may contribute to the models' inaccuracies. Spacecraft measurements obtained during last 2 decades provide a great amount of new information about small-scale plasma processes in near-Earth space. However, the influence of solar wind on the dynamics of the small-scale structures in the magnetosheath has been rarely addressed. The present review summarizes experimental studies on this influence including features of turbulence around ion scales. The study aims to give a general picture of the problem and underline the gaps in current understanding of the role of the dynamics of the small-scale structures and turbulence in the solar–terrestrial relations.

KEYWORDS

solar wind, bow shock, magnetosheath, Sun–Earth coupling, turbulence

1 Introduction

Interaction of the supersonic and super-Alfvenic solar wind (SW) flow with the Earth's magnetosphere results in the formation of outstanding bow shock (BS) and a region behind it filled with turbulent shocked plasma—magnetosheath (MSH). The MSH plasma and magnetic field are in contact with the Earth's magnetosphere. Therefore, the analysis of the MSH processes is of great importance for the exploration of solar–terrestrial relations and space weather forecasts.

Predictions of the magnetosphere response on the changing SW parameters are usually based on spacecraft measurements in the Lagrange L1 point (Yermolaev et al., 2010; Pallocchia et al., 2006; Boynton et al., 2012; Podladchikova and Petrukovich, 2012). These predictions do not account for alteration in plasma and magnetic field parameters in the course of passage from L1 to the magnetosphere, including substantial distortion of the flow in the foreshock region (Sibeck et al., 2008) and inside the MSH (Rakhmanova et al., 2015; 2016) and changes in the magnetic field orientation in front of the magnetopause (Šafránková et al., 2009; Pulinets et al., 2014). This approach provides high reliability only for intensive magnetosphere disturbances such as geomagnetic storms (Petrukovich et al., 2001). In addition to the statistical approach, models of the solar–terrestrial relations are based typically on MHD descriptions of plasma flow around the magnetosphere (Spreiter and

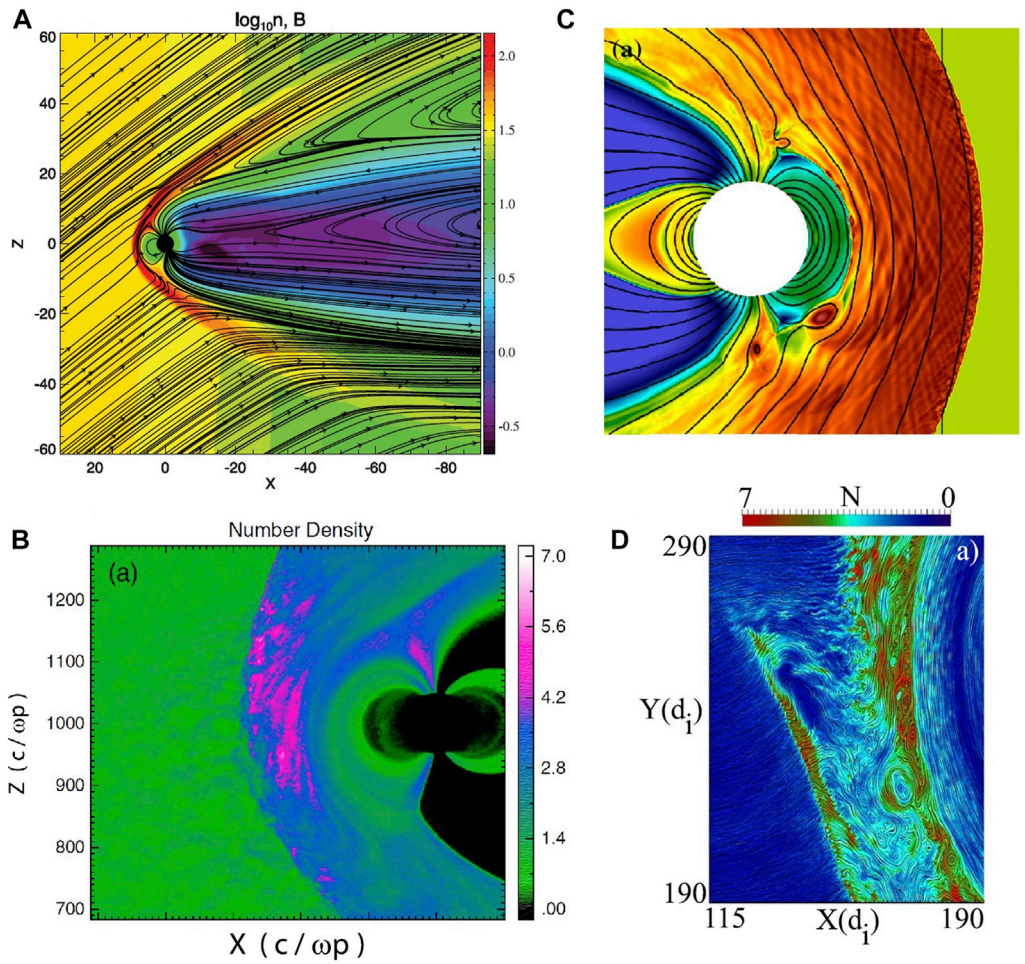


FIGURE 1
Magnetosheath density (shown by color) reproduced by **(A)** MHD model (adopted from Toth et al., 2005; figure 13), **(B)** 2.5-D electromagnetic hybrid simulations (adopted from Omidi et al., 2014; figure 10a), **(C)** hybrid Vlasov simulation (adopted from Palmroth et al., 2018; figure 7a (distributed under the terms of the Creative Commons Attribution 4.0 http://creativecommons.org/licenses/by/4.0/), and **(D)** 2D global hybrid simulation (the bow shock structure is shown, adopted from Karimabadi et al., 2014; figure 13a).

Stahara, 1980; Kartalev et al., 1996; Toth et al., 2005). Though the MHD description is of great importance for global understanding of the MSH–magnetosphere system, it cannot reproduce high-amplitude small-scale variations (Zastenker et al., 2002; 2008; Hayosh et al., 2006). These variations are characterized by frequencies above 10^{-3}–10^{-2} Hz and amplitudes above 5–10% and occur throughout the MSH (Němeček et al., 2000; Shevyrev et al., 2003; Lucek et al., 2005; Shevyrev and Zastenker, 2005; Gutynska et al., 2008; Gutynska et al., 2009; Gutynska et al., 2015). Moreover, typically, the description of the SW and MSH plasma and magnetic field is based on validity of the frozen-in hypothesis, which may be violated for low-speed plasma in the MSH (see Antonova et al., 2012 and references therein). Suggested mechanisms of plasma penetration through the magnetopause include finite gyroradius scattering, diffusion, wave-induced diffusion, impulsive penetration, and magnetic reconnection (Sibeck et al., 1999; Kirpichev et al., 2017; Fuselier, 2021). However, the presence of small-scale variations in the MSH favors additional ways for plasma transport across the magnetopause (Echim and Lemaire, 2002; Dmitriev and Suvorova, 2015); therefore, they can affect the magnetosphere processes. Observed discrepancies between models of the solar–terrestrial relations and reality may be a result of disregarding processes taking place at the scales around and below ion gyroradius (small-scale processes) in the MSH. Recent progress in technology made it possible to use hybrid models (Karimabadi et al., 2014; Omidi et al., 2014; Palmroth et al., 2018; Lu et al., 2020) which can successfully reproduce the basic properties of highly fluctuating MSH environments. Figure 1 presents the comparison of the MHD-modeled MSH (panel A, adopted from Toth et al., 2005) with results of three hybrid (electron fluid, kinetic ions) models (panels B–D). Small-scale structures are clearly captured by latter models, demonstrating the importance of their consideration. However, the stochastic nature of turbulent small-scale fluctuations makes them highly irreproducible both for distinct cases and monitoring of the MSH parameters. Therefore, in the exploration of mean characteristics of small-scale structures and turbulence, differences in its features for various background conditions may give information on when the existing forecasts would work well and when they should be corrected to account for processes in the MSH.

For a long time, MSH fluctuations have been considered in a framework of waves (Anderson et al., 1994; Hubert, 1994; Lacombe and Belmont, 1995; Schwartz et al., 1996). Great contributions have been made to the understanding of the MSH fluctuations (and are still in progress) with the help of Cluster mission measurements (Lucek et al., 2005). Recent measurements by MMS helped to better understand the small-scale fluctuations and reconnection in the turbulent MSH plasma (Stawarz et al., 2016; 2022; Vörös et al., 2019; Yordanova et al., 2020). A number of studies treated the MSH fluctuations in a framework of turbulence (see Alexandrova, 2008; Sahraoui et al., 2020; Rakhmanova et al., 2021 and references therein). Thus, there is a general understanding of sources, spectral features, and spatial scales of the fluctuations. However, the effect of the varying SW conditions on the properties of the MSH fluctuations was addressed quite rarely.

The present mini-review aims to give an overview of recent studies of the SW influence on the small-scale structures behind the BS. Here, the small-scale MSH structures/variations refer to variations of plasma and/or magnetic field parameters, which are registered by a spacecraft during seconds to minutes. The review does not discuss reconnection, current sheets, waves, and instabilities, however, as this topic is far too large and these have been extensively described before (Retinò et al., 2007; Yordanova et al., 2016; 2020; Khabarova et al., 2021 and references therein; Lacombe and Belmont, 1995; Schwartz et al., 1996). Only those studies have been mentioned here, which include observational analysis of the SW influence on the small-scale MSH structures. The authors believe that the review will help find gaps in the exploration of the space weather forecasts and motivate revision of available large datasets of high-quality MSH measurements from this point of view.

2 Basic MSH properties

MSH formation is determined by the interaction of the SW and the magnetosphere. The basic parameter which influences the power of variations behind the BS is the angle between the interplanetary magnetic field and local BS normal—θ_{BN} (Greenstadt, 1972; Shevyrev and Zastenker, 2005). For quasi-perpendicular BS ($\theta_{BN} > 45^{\circ}$), the low-energy SW ions (namely, protons and alfa particles, which are dominant in the SW) pass through the BS front and are advected away from it. Generally, the amplitude of the downstream small-scale fluctuations (with frequencies 0.01–1 Hz) is around 7–10% of the mean parameter value in this case (Shevyrev and Zastenker, 2005). For quasi-parallel BS ($\theta_{BN} < 45^{\circ}$), the SW ions are reflected from the front, go upstream (forming the foreshock region), and are simultaneously advected with the plasma flow to the downstream region. This complex path leads to powerful fluctuations behind the BS. The amplitude of the fluctuations behind quasi-parallel BS is typically twice as high as that behind the quasi-perpendicular BS and reaches 30% for almost-parallel BS ($\theta_{BN} < 10^{\circ}$) (Shevyrev and Zastenker 2005). Another well-known difference between the two configurations of the MSH lies in wave activity. At the quasi-perpendicular BS, significant temperature anisotropy $T_{\perp} >> T_{\parallel}$ arises with respect to the mean magnetic field vector. Temperature anisotropy leads to excitation of wave processes and instabilities, with their type depending on local plasma beta parameter (Lacombe and Belmont, 1995; Schwartz et al., 1996). Substantial differences in properties of the two configurations of the MSH imply that they should be considered separately to reveal the SW influence on the processes there. The foreshock region, populated with ions reflected from the BS, presents a complex, highly fluctuating, and anisotropic environment in front of the quasi-parallel BS (see Blanco-Cano et al., 2006; Sibeck et al., 2008; Gutynska et al., 2015 and references in these papers). Effects of the foreshock processes on the downstream MSH variations combine both foreshock intrinsic processes and pure SW effects. For this reason, foreshock processes are not focused on in the present review.

Another property of the MSH parameters is their dawn–dusk asymmetry which partly represents differences of the flow parameters behind the different BS types. A number of studies were focused on the problem of asymmetry in parameters' values and MSH thickness (Němeček et al., 2000; Paularena et al., 2001; Němeček et al., 2003; Longmore et al., 2005; Dimmock and Nykyri, 2013). Generally, the dusk flank of the MSH is characterized by higher magnetic field magnitude, plasma speed, lower density, and ion temperature than the dawn flank (Dimmock et al., 2017). Most of the studies suggested SW speed and direction of the IMF to be the main parameters which affect the asymmetry.

3 Properties of the MSH fluctuations for changing SW conditions

A number of studies have tried to reveal if there was a connection between the MSH fluctuations properties and the SW parameters. Němeček et al. (2002) analyzed statistics of Interball-1 crossing of the MSH at 30-min intervals, which corresponded to fluctuations in the range ~[0.001–0.01] Hz. The authors showed the fluctuation level weakly depends on the IMF B_z component and not on the SW Alfven Mach number M_A and plasma β. However, a higher level of fluctuations was observed for decreasing the IMF cone angle, which could refer to the influence of the BS topology.

Gutynska et al. (2009) presented extensive statistics of magnetic field fluctuations in the frequency range [0.001–0.125] Hz measured by Cluster and analyzed the correlation properties of these fluctuations as a function of the SW parameters. Correlation length of the fluctuations (being 1 R_E in average) was shown to increase significantly for increasing SW velocity, stayed unchanged for different SW densities, and was slightly higher for southward IMF compared to the northward IMF. Gutynska et al. (2015) analyzed properties of density transients (density increase by a factor of 1.5 and more during 10–60 s) in the dayside MSH and concluded that they tended to occur during high SW speeds and low densities. The authors also showed preferable occurrence of the structures during low IMF cone and θ_{BN} angles and concluded that density transients were often associated with foreshock processes.

Dimmock et al. (2014) studied the influence of the SW velocity and IMF direction on the distribution of the MSH magnetic field fluctuations in the frequency range [0.1, 2] Hz and their dawn–dusk asymmetry based on 6 years of THEMIS measurements. The authors concluded that fluctuations with higher amplitudes were observed for periods of fast SW (with velocities above 400 km/s). Additionally, amplified fluctuations in the vicinity of the magnetopause and

absence of dawn–dusk asymmetry were shown for southward IMF. The authors used a magnetosheath interplanetary medium reference frame (Verigin et al., 2006) which allowed to account for aberration and for dynamics of the BS and magnetopause positions with variations of the SW parameters.

Rakhmanova et al. (2015), Rakhmanova et al. (2016), and Rakhmanova et al. (2018a) presented a statistical study of the two-point correlation function in the SW and MSH with the help of THEMIS data for density and magnetic field magnitude variations in the range [0.001–0.01] Hz. The authors suggested that the SW variations are more likely to stay unchanged during the BS crossing for higher SW densities and IMF magnitude. Additionally, SW variations with higher amplitudes were shown to be less modified in the MSH. No significant dependencies were found on the SW plasma β, IMF direction, or SW speed.

4 MSH jets and their dependence on the SW parameters

Presence of jets, usually described as transient enhancements of the dynamic pressure or density, is an important phenomenon in the magnetosheath as they can lead to local deformations in the magnetopause and plasma penetration (Lemaire, 1977; Lemaire and Roth, 1978; Dmitriev and Suvorova, 2015). There are several causes suggested for their formation including the BS ripples (Hietala et al., 2009; Hietala et al., 2012) and interaction of the SW discontinuities with the BS (Němeček et al., 1998; Savin et al., 2004; Savin et al. 2011; Savin et al. 2012; Dmitriev et al., 2021). A full overview of the phenomena can be found in Plaschke et al. (2018) and Dmitriev et al. (2021). Here, we highlight the results referring to the dependence of the jets on the SW parameters.

Archer and Horbury (2013) analyzed distribution of the high-amplitude transient (during the time intervals [10, 180] sec) enhancements of the dynamical pressure inside the MSH and their dependence on the SW conditions. The authors distinguished two basic regimes characterized by decrease or increase of density. The first group was typically observed during high-IMF cone angles, while the second group occurred predominantly for low-IMF cone angles. For the second regime (the most common), the enhancements were more frequent for increasing solar wind speed, while no dependencies were found on the IMF clock angle, SW plasma parameter β, or Mach number.

Plaschke et al. (2013) considered SW conditions favorable for jet formation in the subsolar MSH and concluded that the only significant factor was the IMF cone angle, which implied that the quasi-parallel BS was more favorable for jet formation. The authors used statistics of THEMIS measurements during 2008–2011, i.e., during solar minimum. Thus, the study included limited statistics of high values of the SW parameters such as ion density, speed, and IMF magnitude. Further analysis by LaMoury et al. (2021) considered enlarged statistics which covered the years 2008–2018 and showed additional SW parameters favorable for jet formation: SW speed around 300 km/s and 600 km/s, low IMF magnitude and dynamic pressure, high plasma β, and Alfven Mach number. The authors considered separately the SW conditions favorable for jets to reach the magnetopause if they were formed near the BS. These conditions included low-IMF cone angle and plasma β, high values of SW speed, IMF magnitude, and dynamic pressure. Thus, some of the SW properties (namely, the IMF magnitude) may contribute not only to the processes of jets' formation but also to their dissipation in the MSH, and the relative portion of these effects are not clear so far.

The multiple SW factors, which may affect the magnetosphere and MSH, are not physically independent. Typically, their values are grouped according to the type of the SW, e.g., an origin of the SW stream at the Sun (Yermolaev et al., 2009; Borovsky et al., 2019). Generally, main types of the SW include slow or fast wind, interplanetary coronal mass ejections (ICMEs), and co-rotating interaction regions CIRs (or stream interaction regions, SIRs). In front of the ICMEs or CIRs, compressed regions may be formed (ICME Sheath or CIR Sheath); some of the catalogs of large-scale SW phenomena list them as a distinct SW type (Yermolaev et al., 2009). Different large-scale flows have varying geoeffectiveness (Borovsky and Denton, 2006; Yermolaev et al., 2012). Recently, the combination of the SW properties started to be considered as driving small-scale MSH processes as well. Koller et al. (2022) compared a list of the ICMEs and SIRs with the lists of the MSH jets and found clear differences in the jet's occurrence rate for distinct large-scale SW types. For SIRs and high-speed streams, the number of jets was increased by up to 50% compared to the mean occurrence rate. On the other hand, the occurrence rate was decreased by 50% for ICMEs and associated Sheath regions. The authors noted that the study included only those jets which were aligned with the Sun–Earth line and are, therefore, geoeffective as they reached the magnetopause. This finding uncovers new information on geoeffectiveness of the SW events, as ICMEs and their Sheath regions are typically supposed to cause the most intense magnetospheric response.

5 Role of the large-scale SW streams in dynamics of the small-scale MSH structures

As mentioned in the previous section, large-scale SW structures have different effects on the small-scale MSH variations. However, typically their properties are considered as constant during the flow from the Lagrange L1 point to the magnetopause. At small scales, the plasma and magnetic field variations embedded in the large-scale structures could be modified across the BS and inside the MSH.

Turc et al. (2014) presented a study of simultaneous registration of the magnetic clouds (MCs and ICMEs with strong and regular magnetic field) in the MSH based on four representative cases. They have shown that behind the quasi-parallel BS, the smooth rotation of the magnetic field orientation in the MC may be dominated by the MSH variations, which results in difficulties in the magnetosphere response forecasts based on the SW data. Furthermore, Turc et al. (2017) presented a statistical study on simultaneous variations of the magnetic field in the MC upstream and downstream from the BS for 82 cases. The B_z component (which is believed to be one of the most geoeffective factors due to the processes of reconnection at the magnetopause) was shown to change its sign in the MSH, sometimes for periods as long as 30 min and more. These cases were observed for the IMF which lay in the equatorial plane. This result corresponds well with earlier results by Šafránková et al. (2009)

who showed statistical differences between the sign of the B_z component in the SW and MSH. Turc et al. (2017) demonstrated that if the predominant IMF component was B_y, both positive and negative B_z could have equal geoeffectiveness.

Rakhmanova et al. (2018a) tried to connect the modification of the small-scale plasma structures in the MSH with the type of the upstream SW flow. Their results showed that during slow undisturbed SW, the small-scale SW structures stayed unchanged in the MSH just for 30% of cases. On the other hand, this percentage was higher by the factor of 2 during periods of the CIRs in the SW, that is, small-scale variations inside CIRs tend to survive across the BS. However, for other SW phenomena such as ICME ejecta, the small-scale variations tended to stay unmodified only for half of the cases.

Recently, Ala-Lahti et al. (2021) considered passage of the ICME Sheath through the BS. The authors compared magnetic field variations inside the Sheath, registered simultaneously in three points in the MSH, in the SW close to the BS, and at the Lagrange L1 point. The IMF B_z component was shown to be enhanced for a set of substructures, which made them more geoeffective. For some parts of the Sheath, the upstream small-scale structures survived during the transmission through the BS. On the other hand, for other periods, there was no correlation between upstream and downstream variations. The authors showed that variations with frequencies less than 0.01 Hz had increasing probability of surviving during passage through the BS.

Thus, the small-scale MSH structures may evolve in a different way for different large-scale SW types. Statistical relations between the SW parameters and magnetosphere responses often consider the SW type as a parameter. Therefore, an additional study of differences in its effect on the MSH small-scale structures would help improve predictability of the models.

6 SW effects on the features of the MSH turbulence and its evolution

Turbulence of the SW has been proposed to be one of the drivers of the magnetosphere dynamics (see D'Amicis et al., 2020 and references therein). Changes in the turbulence properties in the MSH for various SW conditions may be, thus, important for the magnetosphere response on the varying SW flow. Typically, around ion scales, the fluctuation spectra of the magnetic field and ion flux/density are characterized by two power laws with a break both in the SW and MSH; at frequencies lower than the break (MHD scales), the power law exponent (slope) is close to −5/3 (so-called Kolmogorov spectrum), while at higher frequencies (kinetic scales) the slope is around −2.8 (Alexandrova et al., 2013; Sahraoui et al., 2020; Rakhmanova et al., 2021). Statistical studies demonstrated modification of the shape of turbulent cascade around ion scales at the BS and its restoration at the flanks and near the magnetopause (Gutynska et al., 2009; Huang et al., 2017a; Rakhmanova et al., 2018b; Rakhmanova et al., 2018c). These statistical studies addressed the mean parameters and clearly demonstrated non-Kolmogorov scaling at MHD scales at the dayside MSH regions in the vicinity of the BS. However, several case studies (Rakhmanova et al., 2020a; Rakhmanova et al., 2020b) showed differences in the evolution of spectra between two points inside the MSH and different scenarios of the spectrum modification at the BS. The authors have suggested that turbulent cascade may evolve differently for specific SW conditions. A number of studies considered the influence of the SW parameters on the MSH turbulent spectra. Gutynska et al. (2009) showed changes of the mean spectral slopes at the MHD scales from −1.2 to −0.9 with increasing SW velocity. A statistical study (Rakhmanova et al., 2020c) showed a tendency to observe typical Kolmogorov scaling at the MHD scales for high-amplitude variations of the SW density and not to observe Kolmogorov scaling for high SW density values. Also, the authors showed that the spectral slope at the kinetic scales tended to deviate substantially from typical values and from theoretical predictions for northward IMF, while for southward IMF, the slope values were around −2.8, usually registered in the SW and MSH. The authors did not find any significant dependence on the SW velocity. On the other hand, Li et al. (2020) used MMS statistics and suggested no influence of the SW velocity and IMF B_z value on the evolution of the turbulence inside the MSH.

The preliminary statistical study of Rakhmanova et al. (2020b) demonstrated that changes in turbulent cascade at the BS may be influenced by the large-scale SW type. Recently, Rakhmanova et al. (2022) presented a set of multipoint case studies, which compared properties of the turbulent spectra in the SW, in the dayside MSH and at the MSH flank, taking into account the plasma propagation delays. The study included 12 cases from three groups of the SW stream types: slow undisturbed SW, ICMEs, and compressed SW (Sheath and CIRs). The results demonstrated that for undisturbed SW, the quasi-perpendicular BS affected the turbulent cascade in a confined region close to the BS, and the shape of the SW spectra restored at the flank MSH. On the other hand, interaction of the disturbed SW streams with the BS always resulted in violation of the Kolmogorov scaling at the MHD scales and in significant steepening of spectra at the kinetic scales in the whole dayside MSH. Interestingly, for the ICMEs, the spectra at the MSH flank tended to have properties similar to those registered in the SW, but for compressed SW flows, the properties of the cascade at the kinetic scales were not restored. This result is demonstrated in Figure 2, where slopes of the spectra in the dayside MSH and in the flank MSH at frequencies higher (P_2) than the ion spectral break are shown *versus* the same slopes in the SW for three groups of the large-scale SW types. The panel on the left demonstrates linear dependence of the turbulence properties downstream from the BS in the dayside MSH on those in the upstream region for the three groups of the SW types. At the flank, the right panel exhibits substantial differences in development of kinetic-scale turbulence for compressed SW flows.

7 Summary

To date, few works were focused directly on the solar wind parameters, which influence the small-scale magnetosheath structures. The present mini-review underlines that

- Varying solar wind conditions contribute to the dynamics of the small-scale structures inside the magnetosheath; however, the contribution of this factor is much less examined than the influence of the foreshock, reconnection, and wave activity.
- IMF parameters (magnitude, B_z component, and cone angle) and ion bulk speed are most frequently reported solar wind parameters, which affect the dynamics of small-scale structures behind the bow shock.

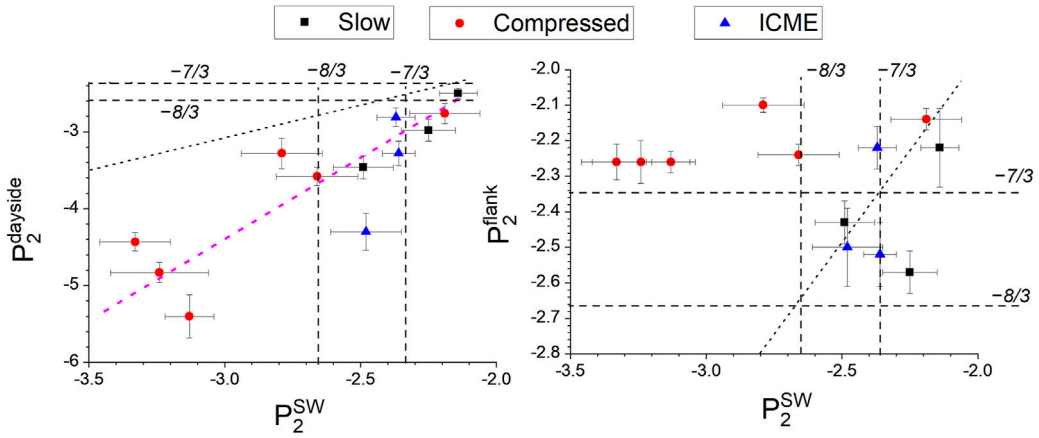


FIGURE 2
Spectral slopes at the frequencies higher than the ion spectral break in the dayside magnetosheath (left) and at the flank (right) *versus* same slopes upstream from the bow shock for three types of the solar wind; dotted lines show equal values of the downstream and upstream slopes, and magenta dashed line presents the linear fit (adopted from Rakhmanova et al., 2022; figure 6c, d, distributed under the terms of the Creative Commons Attribution 4.0 http://creativecommons.org/licenses/by/4.0/).

- Recently, enlarged statistics of high-quality measurements in the magnetosheath helped to find additional factors which contributed to the dynamics of the small-scale structures in the magnetosheath. These factors were plasma density, ion parameter β, and Alfven Mach number.
- Ratio between the contribution of the aforementioned factors and their dependence on each other are still a matter of investigation; complex factors such as type of the solar wind flow may be more perspective for further exploration.

Solar wind parameters contribute to the solar–terrestrial relations not only determining the large-scale dynamics of the magnetosphere system, but through the changes in dynamics of small-scale MSH structures as well. A set of small-scale structures, found with the help of kinetic simulations and reported in the observations, was still not considered in the context of their relation to the upstream solar wind conditions. These structures include magnetic islands (Huang et al., 2017b), magnetic holes (Huang et al., 2016; Yao et al., 2021), plasmoids (Goncharov et al., 2020), and Alfven vortices (Alexandrova, 2008). A combination of different mission measurements during all parts of the solar cycle is required to get a full picture of properties of the small-scale MSH structures, their sources and contribution to the solar–terrestrial relations. To date, a set of missions with high-quality plasma and magnetic field measurements, complemented with realistic simulations, allows for such a complex exploration (Dimmock et al., 2020). The present review shows what has been achieved and what could be achieved in this scope for understanding the role of small-scale magnetosheath structures.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication. In particular, the paper was conceived and written by LR with the support and critical feedback of MR, GZ, and YY.

Funding

The study is supported by the Russian Ministry of Education by the theme “Plasma”.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ala-Lahti, M., Dimmock, A. P., Pulkkinen, T. I., Good, S. W., Yordanova, E., Turc, L., et al. (2021). Transmission of an ICME sheath into the Earth’s magnetosheath and the occurrence of traveling foreshocks. *J. Geophys. Res. Space Phys.* 126, e2021JA029896. doi:10.1029/2021JA029896

Alexandrova, O., Chen, C. H. K., Sorriso-Valvo, L., Horbury, T. S., and Bale, S. D. (2013). Solar wind turbulence and the role of ion instabilities. *Space Sci. Rev.* 178, 101–139. doi:10.1007/s11214-013-0004-8

Alexandrova, O. (2008). Solar wind vs magnetosheath turbulence and Alfvén vortices. *Geophys* 15, 95–108. doi:10.5194/npg-15-95-2008

Anderson, B. J., Fuselier, S. A., Gary, S. P., and Denton, R. E. (1994). Magnetic spectral signatures in the Earth's magnetosheath and plasma depletion layer. *J. Geophys. Res.* 99, 5877–5891. doi:10.1029/93ja02827

Antonova, E. E., Pulinets, M. S., Riazantseva, M. O., Znatkova, S. S., Kirpichev, I. P., and Stepanova, M. V. (2012). in *Turbulence in the magnetosheath and the problem of plasma penetration inside the magnetosphere Exploring the solar wind*. Editor M. Lazar (Berlin, Germany: Springer), 417–438.

Archer, M. O., and Horbury, T. S. (2013). Magnetosheath dynamic pressure enhancements: Occurrence and typical properties. *Ann. Geophys.* 31 (2), 319–331. doi:10.5194/angeo-31-319-2013

Blanco-Cano, X., Omidi, N., and Russell, C. T. (2006). Macrostructure of collisionless bow shocks: 2. ULF waves in the foreshock and magnetosheath. *J. Geophys. Res.* 111, A10205. doi:10.1029/2005JA011421

Borovsky, J. E., and Denton, M. H. (2006). Differences between CME-driven storms and CIR-driven storms. *J. Geophys. Res.* 111, A07S08. doi:10.1029/2005ja011447

Borovsky, J. E., Denton, M. H., and Smith, C. W. (2019). Some properties of the solar wind turbulence at 1 AU statistically examined in the different types of solar wind plasma. *J. Geophys. Res. Space Phys.* 124, 2019JA026580–2424. doi:10.1029/2019JA026580

Boynton, R. J., Balikhin, M. A., Billings, S. A., Sharma, A. S., and Amariutei, O. A. (2012). Data derived NARMAX Dst model. *Ann. Geophys.* 29, 965–971. doi:10.5194/angeo-29-965-2011

D'Amicis, R., Telloni, D., and Bruno, R. (2020). The effect of solar-wind turbulence on magnetospheric activity. *Front. Phys.* 8, 604857. doi:10.3389/fphy.2020.604857

Dimmock, A. P., Hietala, H., and Zou, Y. (2020). Compiling magnetosheath statistical data sets under specific solar wind conditions: Lessons learnt from the dayside kinetic southward IMF GEM challenge. *Earth Space Sci.* 7, e2020EA001095. doi:10.1029/2020EA001095

Dimmock, A. P., Nykyri, K., Osmane, A., Karimabadi, H., and Pulkkinen, T. I. (2017). "Dawn-dusk asymmetries of the Earth's dayside magnetosheath in the magnetosheath interplanetary medium reference frame," in *Dawn-dusk asymmetries in planetary plasma environments* (Washington, D.C., United States: American Geophysical Union), 49–72. doi:10.1002/9781119216346.ch5

Dimmock, A. P., Nykyri, K., and Pulkkinen, T. I. (2014). A statistical study of magnetic field fluctuations in the dayside magnetosheath and their dependence on upstream solar wind conditions. *J. Geophys. Res. Space Phys.* 119 (8), 6231–6248. doi:10.1002/2014JA020009

Dimmock, A. P., and Nykyri, K. (2013). The statistical mapping of magnetosheath plasma properties based on THEMIS measurements in the magnetosheath interplanetary medium reference frame. *J. Geophys. Res. Space Phys.* 118, 4963–4976. doi:10.1002/jgra.50465

Dmitriev, A. V., Lalchand, B., and Ghosh, S. (2021). Mechanisms and evolution of geoeffective large-scale plasma jets in the magnetosheath. *Universe* 7, 152. doi:10.3390/universe7050152

Dmitriev, A. V., and Suvorova, A. V. (2015). Large-scale jets in the magnetosheath and plasma penetration across the magnetopause: THEMIS observations. *J. Geophys. Res. Sp. Phys.* 120, 4423–4437. doi:10.1002/2014JA020953

Echim, M. M., and Lemaire, J. (2002). Positive density gradients at the magnetopause: Interpretation in the framework of the impulsive penetration mechanism. *J. Atmos. Sol. Terr. Phys.* 64, 2019–2028. doi:10.1016/s1364-6826(02)00229-8

Fuselier, S. A. (2021) *Dayside magnetopause processes space Physics and aeronomy collection volume 2: Magnetospheres in the solar system, geophysical monograph 259*. Editors Romain Maggiolo, Nicolas André, Hiroshi Hasegawa, and Daniel T. Welling First Edition (Netherland: John Wiley & Sons). doi:10.1002/9781119815624.ch10

Goncharov, O., Gunell, H., Hamrin, M., and Chong, S. (2020). Evolution of high-speed jets and plasmoids downstream of the quasi-perpendicular bow shock. *J. Geophys. Res. Space Phys.* 125, e2019JA027667. doi:10.1029/2019JA027667

Greenstadt, E. W. (1972). Binary index for assessing local bow shock obliquity. *J. Geophys. Res.* 77, 927 5467–5479. doi:10.1029/JA077i028p05467

Gutynska, O., Šafránková, J., and Němeček, Z. (2008). Correlation length of magnetosheath fluctuations: Cluster statistics. *Ann. Geophys.* 26, 2503–2513. doi:10.5194/angeo-26-2503-2008

Gutynska, O., Šafránková, J., and Němeček, Z. (2009). Correlation properties of magnetosheath magnetic field fluctuations. *J. Geophys. Res. Sp. Phys.* 114, A8. doi:10.1029/2009JA014173

Gutynska, O., Sibeck, D. G., and Omidi, N. (2015). Magnetosheath plasma structures and their relation to foreshock processes. *J. Geophys. Res. Space Phys.* 120, 7687–7697. doi:10.1002/2014JA020880

Hayosh, M., Šafránková, J., and Němeček, Z. (2006). MHD-modelling of the magnetosheath ion plasma flow and magnetic field and their comparison with experiments. *Adv. Space Res.* 37, 507–514. doi:10.1016/j.asr.2005.07.059

Hietala, H., Andréeová, K., Vainio, R., Vaivads, A., Palmroth, M., Rème, H., et al. (2009). Supermagnetosonic jets behind a collisionless quasiparallel shock. *Phys. Rev. Lett.* 103 (24), 245001–245023. doi:10.1103/PhysRevLett.103.245001

Hietala, H., Laitinen, T. V., Clausen, L. B., Facskò, G., Vaivads, A., Lucek, E. A., et al. (2012). Supermagnetosonic subsolar magnetosheath jets and their effects: From the solar wind to the ionospheric convection. *Ann. Geophys.* 30 (1), 33–48. doi:10.5194/angeo-30-33-2012

Huang, S. Y., Hadid, L. Z., Sahraoui, F., Yuan, Z. G., and Deng, X. H. (2017a). On the existence of the Kolmogorov inertial range in the terrestrial magnetosheath turbulence. *Astrophys. J. Lett.* 836, L10. doi:10.3847/2041-8213/836/1/L10

Huang, S. Y., Sahraoui, F., Retino, A., Le Contel, O., Yuan, Z. G., Chasapis, A., et al. (2016). MMS observations of ion-scale magnetic island in the magnetosheath turbulent plasma. *Geophys. Res. Lett.* 43, 7850–7858. doi:10.1002/2016GL070033

Huang, S. Y., Sahraoui, F., Yuan, Z. G., He, J. S., Zhao, J. S., et al. (2017b). A statistical study of kinetic-size magnetic holes in turbulent magnetosheath: MMS observations. *J. Geophys. Res. Space Phys.* 122, 8577–8588. doi:10.1002/2017JA024415

Hubert, D. (1994). Nature and origin of wave modes in the dayside Earth magnetosheath. *Adv. Sp. 960 Res.* 14, 55–64. doi:10.1016/0273-1177(94)90048-5

Karimabadi, H., Roytershteyn, V., Vu, H. X., Omelchenko, Y. A., Scudder, J., Daughton, W., et al. (2014). The link between shocks, turbulence, and magnetic reconnection in collisionless plasmas. *Phys. Plasmas* 21, 062308. doi:10.1063/1.4882875

Kartalev, M. D., Nikolova, V. I., Kamenetsky, V. F., and Mastikov, I. P. (1996). On the self-consistent determination of dayside magnetopause shape and position. *Planet. Space Sci.* 44, 1195–1208. doi:10.1016/S0032-0633(96)00040-2

Khabarova, O., Malandraki, O., Malova, H., Kislov, R., Greco, A., Bruno, R., et al. (2021). Current sheets, plasmoids and flux ropes in the heliosphere. *Space Sci. Rev.* 217, 38. doi:10.1007/s11214-021-00814-x

Kirpichev, I. P., Antonova, E. E., and Stepanova, M. (2017). Ion leakage at dayside magnetopause in case of high and low magnetic shears. *J. Geophys. Res. Space Phys.* 122, 8078–8095. doi:10.1002/2016JA023735

Koller, F., Temmer, M., Preisser, L., Plaschke, F., Geyer, P., Jian, L. K., et al. (2022). Magnetosheath jet occurrence rate in relation to CMEs and SIRs. *J. Geophys. Res. Space Phys.* 127, e2021JA030124. doi:10.1029/2021JA030124

Lacombe, C., and Belmont, G. (1995). Waves in the Earth's magnetosheath: Observations and interpretations. *Adv. Sp. Res.* 15, 329–340. doi:10.1016/0273-1177(94)00113-F

LaMoury, A. T., Hietala, H., Plaschke, F., Vuorinen, L., and Eastwood, J. P. (2021). Solar wind control of magnetosheath jet formation and propagation to the magnetopause. *J. Geophys. Res. Space Phys.* 126 (9), e2021JA029592. doi:10.1029/2021ja029592

Lemaire, J. (1977). Impulsive penetration of filamentary plasma elements into the magnetospheres of the Earth and Jupiter. *Planet. Space Sci.* 25, 887–890. doi:10.1016/0032-0633(77)90042-3

Lemaire, J., and Roth, M. (1978). Penetration of solar wind plasma elements into the magnetosphere. *J. Atmos. Terr. Phys.* 40, 331–335. doi:10.1016/0021-9169(78)90049-1

Li, H., Jiang, W., Wang, C., Verscharen, D., Zeng, C., Russell, C. T., et al. (2020). Evolution of the Earth's magnetosheath turbulence: A statistical study based on MMS observations. *Astrophys. J.* 898, L43. doi:10.3847/2041-8213/aba531

Longmore, M., Schwartz, S. J., Geach, J., Cooling, B. M. A., Dandouras, I., Lucek, E. A., et al. (2005). Dawn-dusk asymmetries and sub-Alfvénic flow in the high and low latitude magnetosheath. *Ann. Geophys.* 23, 3351–3364. doi:10.5194/angeo-23-3351-2005

Lu, Q., Wang, H., Wang, X., Lu, S., Wang, R., Gao, X., et al. (2020). Turbulence-Driven magnetic reconnection in the magnetosheath downstream of a quasi-parallel shock: A three-dimensional global hybrid simulation. *Geophys. Res. Lett.* 47, e2019GL085661. doi:10.1029/2019gl085661

Lucek, E. A., Constantinescu, D., Goldstein, M. L., Pickett, J., Pinçon, J. L., Sahraoui, F., et al. (2005). The magnetosheath. *Space Sci. Rev.* 118, 95–152. doi:10.1007/s11214-005-3825-2

Němeček, Z., Hayosh, M., Šafránková, J., Zastenker, G. N., and Richardson, J. D. (2003). The dawn-dusk asymmetry of the magnetosheath: INTERBALL-1 observations. *Adv. Space Res.* 31, 1333–1340. doi:10.1016/S0273-1177(03)00007-3

Němeček, Z., Šafránková, J., Přech, L., Sibeck, D. G., Kokubun, S., and Mukai, T. (1998). Transient flux enhancements in the magnetosheath. *Geophys. Res. Lett.* 25, 1273–1276. doi:10.1029/98GL50873

Němeček, Z., Šafránková, J., Zastenker, G. N., Pišoft, P., and Jelínek, K. (2002). Low-frequency variations of the ion flux in the magnetosheath. *Planet. Space Sci.* 50, 567–575. doi:10.1016/s0032-0633(02)00036-3

Němeček, Z., Šafránková, J., Zastenker, G. N., Pišoft, P., Paularena, K. I., and Richardson, J. D. (2000). Observations of the radial magnetosheath profile and a comparison with gasdynamic model predictions. *Geophys. Res. Lett.* 27, 2801–2804. doi:10.1029/2000GL000063

Omidi, N., Sibeck, D., Gutynska, O., and Trattner, K. J. (2014). Magnetosheath filamentary structures formed by ion acceleration at the quasi-parallel bow shock. *J. Geophys. Res. Space Phys.* 119, 2593–2604. doi:10.1002/2013JA019587

Pallocchia, G., Amata, E., Consolini, G., Marcucci, M. F., and Bertello, I. (2006). Geomagnetic <i>D_{st}</i&

gt; index forecast based on IMF data only. *Ann. Geophys.* 24 (3), 989–999. doi:10.5194/angeo-24-989-2006

Palmroth, M., Ganse, U., Pfau-Kempf, Y., Battarbee, M., Turc, L., Brito, T., et al. (2018). Vlasov methods in space physics and astrophysics. *Living Rev. comput. Astrophys.* 4, 1. doi:10.1007/s41115-018-0003-2

Paularena, K. I., Richardson, J. D., Kolpak, M. A., Jackson, C. R., and Siscoe, G. L. (2001). A dawn-dusk density asymmetry in Earth's magnetosheath. *J. Geophys. Res.* 106, 25377–25394. doi:10.1029/2000JA000177

Petrukovich, A. A., Lazarus, A., Lepping, R. P., and Klimov, S. I. (2001). Comparison of the solar wind energy input to the magnetosphere measured by Wind and Interball-1. *J. Atmos. Sol. Terr. Phys.* 63, 1643–1647. doi:10.1016/s1364-6826(01)00039-6

Plaschke, F., Hietala, H., and Angelopoulos, V. (2013). Anti-sunward high-speed jets in the subsolar magnetosheath. *Ann. Geophys.* 31 (10), 1877–1889. doi:10.5194/angeo-31-1877-2013

Plaschke, F., Hietala, H., Blanco-Cano, X., Kajdič, P., Karlsson, T., Sibeck, D., et al. (2018). Jets downstream of collisionless shocks. *Space Sci. Rev.* 214 (5), 81. doi:10.1007/s11214-018-0516-3

Podladchikova, T. V., and Petrukovich, A. A. (2012). Extended geomagnetic storm forecast ahead of available solar wind measurements. *Space weather.* 10, S07001. doi:10.1029/2012SW000786

Pulinets, M. S., Antonova, E. E., Riazantseva, M. O., Znatkova, S. S., and Kirpichev, I. P. (2014). Comparison of the magnetic field before the subsolar magnetopause with the magnetic field in the solar wind before the bow shock. *Adv. Space Res.* 54, 604–616. doi:10.1016/j.asr.2014.04.023

Rakhmanova, L., Riazantseva, M. O., Zastenker, G. N., Yermolaev, Y. I., Lodkina, I. G., and Chesalin, L. S. (2020a). Turbulent cascade in the magnetosheath affected by the solar wind's plasma turbulence. *Cosm. Res.* 57, 443–450. doi:10.1134/S0010952519060066

Rakhmanova, L., Riazantseva, M., and Zastenker, G. (2016). Correlation level between solar wind and magnetosheath plasma and magnetic field parameters. *Adv. Space Res.* 58, 157–165. doi:10.1016/j.asr.2015.09.036

Rakhmanova, L., Riazantseva, M., and Zastenker, G. (2021). Plasma and magnetic field turbulence in the Earth's magnetosheath at ion scales. *Front. Astron. Space Sci.* 7, 616635. doi:10.3389/fspas.2020.616635

Rakhmanova, L., Riazantseva, M., Zastenker, G., and Šafránková, J. (2015). Modification of small- and middle-scale solar wind structures by the bow shock and magnetosheath: Correlation analysis. *Planet. Space Sci.* 115, 12–18. doi:10.1016/j.pss.2015.03.003

Rakhmanova, L., Riazantseva, M., Zastenker, G., and Yermolaev, Y. (2022). Large-scale solar wind phenomena affecting the turbulent cascade evolution behind the quasi-perpendicular bow shock. *Universe* 8 (12), 611. doi:10.3390/universe8120611

Rakhmanova, L., Riazantseva, M., Zastenker, G., Yermolaev, Y., and Lodkina, I. (2020c). Dependence of the properties of a turbulent cascade behind the bow shock on the dynamics of the solar wind parameters. *Cosm. Res.* 58, 478–486. doi:10.1134/S0010952520060088

Rakhmanova, L., Riazantseva, M., Zastenker, G., Yermolaev, Y., and Lodkina, I. (2020b). Dynamics of plasma turbulence at Earth's bow shock and through the magnetosheath. *Astrophys. J.* 901, 30. doi:10.3847/1538-4357/abae00

Rakhmanova, L. S., Riazantseva, M. O., Zastenker, G. N., and Verigin, M. I. (2018c). Effect of the magnetopause and bow shock on characteristics of plasma turbulence in the Earth's magnetosheath. *Geomagn. Aeron.* 58, 718–727. doi:10.1134/S0016793218060129

Rakhmanova, L. S., Riazantseva, M. O., Zastenker, G. N., Verigin, M. I., Yermolaev, Yu. I., and Lodkina, I. G. (2018a). The influence of parameters of the interplanetary medium and magnetosheath boundaries on the correlation coefficient between the ion flux measured in the solar wind and the magnetosheath. *Geomagn. Aeron.* 58, 449–456. doi:10.1134/S0016793218040138

Rakhmanova, L. S., Riazantseva, M., Zastenker, G., and Verigin, M. (2018b). Kinetic-scale ion flux fluctuations behind the quasi-parallel and quasi-perpendicular bow shock. *J. Geophys. Res. Sp. Phys.* 123, 5300–5314. doi:10.1029/2018JA025179

Retinò, A., Sundkvist, D., Vaivads, A., Mozer, F., André, M., and Owen, C. J. (2007). *In situ* evidence of magnetic reconnection in turbulent plasma. *Nat. Phys.* 3, 235–238. doi:10.1038/nphys574

Šafránková, J., Hayosh, M., Gutynska, O., Němeček, Z., and Přech, L. (2009). Reliability of prediction of the magnetosheathB_{Z}component from interplanetary magnetic field observations. *J. Geophys. Res. Space Phys.* 114. doi:10.1029/2009JA014552

Sahraoui, F., Hadid, L., and Huang, S. (2020). Magnetohydrodynamic and kinetic scale turbulence in the near-earth space plasmas: A (short) biased review. *Rev. Mod. Plasma Phys.* 4, 4–1132. doi:10.1007/s41614-020-0040-2

Savin, S., Amata, E., Zelenyi, L., Lutsenko, V., Safrankova, J., Nemecek, Z., et al. (2012). Super fast plasma streams as drivers of transient and anomalous magnetospheric dynamics. *Ann. Geophys.* 30, 1–7. doi:10.5194/angeo-30-1-2012

Savin, S., Budaev, V., Zelenyi, L., Amata, E., Sibeck, D., Lutsenko, V., et al. (2011). Anomalous interaction of a plasma flow with the boundary layers of a geomagnetic trap. *JETP Lett.* 93, 754–762. doi:10.1134/S0021364011120137

Savin, S. P., Zelenyi, L. M., Amata, E., Buechner, J., Blecki, J., Klimov, S. I., et al. (2004). Dynamic interaction of plasma flow with the hot boundary layer of a geomagnetic trap. *Jetp Lett.* 79, 368–371. doi:10.1134/1.1772433

Schwartz, S. J., Burgess, D., and Moses, J. J. (1996). Low-frequency waves in the Earth's magnetosheath: Present status. *Ann. Geophys.* 14, 1134–1150. doi:10.1007/s00585-996-1134-z

Shevyrev, N. N., Zastenker, G. N., Nozdrachev, M. N., Němeček, Z., Šafránková, J., and Richardson, J. D. (2003). High and low frequency large amplitude variations of plasma and magnetic field in the magnetosheath: Radial profile and some features. *Adv. Space Res.* 31, 1389–1394. doi:10.1016/S0273-1177(03)00008-5

Shevyrev, N. N., and Zastenker, G. N. (2005). Some features of the plasma flow in the magnetosheath behind quasi-parallel and quasi-perpendicular bow shocks. *Planet. Space Sci.* 53, 95–102. doi:10.1016/j.pss.2004.09.033

Sibeck, D. G., Omidi, N., Dandouras, I., and Lucek, E. (2008). On the edge of the foreshock: Model-data comparisons. *Ann. Geophys.* 26, 1539–1544. doi:10.5194/angeo-26-1539-2008

Sibeck, D. G., Paschmann, G., Treumann, R. A., Fuselier, S. A., Lennartsson, W., Lockwood, M., et al. (1999). Plasma transfer processes at the magnetopause. *Space Sci. Rev.* 88, 207–283. doi:10.1023/A:1005255801425

Spreiter, J. R., and Stahara, S. S. (1980). A new predictive model for determining solar wind-terrestrial planet interactions. *J. Geophys. Res.* 85, 6769–6777. doi:10.1029/JA085iA12p06769

Stawarz, J. E., Eastwood, J. P., Phan, T. D., Gingell, I. L., Pyakurel, P. S., Shay, M. A., et al. (2022). Turbulence-driven magnetic reconnection and the magnetic correlation length: Observations from Magnetospheric Multiscale in Earth's magnetosheath. *Phys. Plasmas* 29, 012302. doi:10.1063/5.0071106

Stawarz, J. E., Eriksson, S., Wilder, F. D., Ergun, R. E., Schwartz, S. J., Pouquet, A., et al. (2016). 1177 observations of turbulence in a kelvin-helmholtz event on 8 september 2015 by the magnetospheric 1178 multiscale mission. *J. Geophys. Res. Sp. Phys.* 121 (11), 021–111. doi:10.1002/2016JA023458

Tóth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., De Zeeuw, D. L., et al. (2005). Space weather modeling framework: A new tool for the space science community. *J. Geophys. Res. Space Phys.* 110, A12226. doi:10.1029/2005JA011126

Turc, L., Fontaine, D., Escoubet, C. P., Kilpua, E. K. J., and Dimmock, A. P. (2017). Statistical study of the alteration of the magnetic structure of magnetic clouds in the Earth's magnetosheath. *J. Geophys. Res. Space Phys.* 122 (3), 2956–2972. doi:10.1002/2016JA023654

Turc, L., Fontaine, D., Savoini, P., and Kilpua, E. K. J. (2014). Magnetic clouds' structure in the magnetosheath as observed by Cluster and Geotail: Four case studies. *Ann. Geophys.* 32 (10), 1247–1261. doi:10.5194/angeo-32-1247-2014

Verigin, M. I., Tátrallyay, M., Erdös, G., and Kotova, G. A. (2006). Magnetosheath interplanetary medium reference frame: Application for a statistical study of mirror type waves in the terrestrial plasma environment. *Adv. Space Res.* 37, 515–521. doi:10.1016/j.asr.2005.03.042

Vörös, Z., Yordanova, E., Graham, D. B., Khotyaintsev, Y. V., and Narita, Y. (2019). MMS observations of whistler and lower hybrid drift waves associated with magnetic reconnection in the turbulent magnetosheath. *J. Geophys. Res. Space Phys.* 124, 8551–8563. doi:10.1029/2019ja027028

Yao, S. T., Yue, Z. S., Shi, Q. Q., Degeling, A. W., Fu, H. S., Tian, A. M., et al. (2021). Statistical properties of kinetic-scale magnetic holes in terrestrial space. *Earth Planet. Phys.* 5 (1), 63–72. doi:10.26464/epp2021011

Yermolaev, Y. I., Nikolaeva, N. S., Lodkina, I. G., and Yermolaev, M. Y. (2009). Catalog of large-scale solar wind phenomena during 1976-2000. *Cosm. Res.* 47, 81–94. doi:10.1134/S0010952509020014

Yermolaev, Y. I., Nikolaeva, N. S., Lodkina, I. G., and Yermolaev, M. Y. (2012). Geoeffectiveness and efficiency of CIR, Sheath and ICME in generation of magnetic storms. *J. Geophys. Res.* 117. doi:10.1029/2011JA017139

Yermolaev, Yu. I., Nikolaeva, N. S., Lodkina, I. G., and Yermolaev, M. Yu. (2010). Specific interplanetary conditions for CIR-Sheath-and ICME-induced geomagnetic storms obtained by double superposed epoch analysis. *Ann. Geophys.* 28, 2177–2186. doi:10.5194/angeo-28-2177-2010

Yordanova, E., Vörös, Z., Raptis, S., and Karlsson, T. (2020). Current sheet statistics in the 1209 magnetosheath. *Front. Astron. Sp. Sci.* 7, 2. doi:10.3389/fspas.2020.00002

Yordanova, E., Vörös, Z., Varsani, A., Graham, D. B., Norgren, C., Khotyaintsev, Y. V., et al. (2016). Electron scale structures and magnetic reconnection signatures in the turbulent magnetosheath. *Geophys. Res. Lett.* 43, 5969–5978. doi:10.1002/2016GL069191

Zastenker, G. N., Kartalev, M. D., Dobreva, P. S., Shevyrev, N. N., and Koval, A. (2008). A comparison of measured parameters of magnetosheath plasma with predictions of a new magnetosheath–magnetosphere model. *Cosmic Res.* 46, 469–483. doi:10.1134/S0010952508060014

Zastenker, G. N., Nozdrachev, M. N., Němeček, Z., Šafránková, J., Paularena, K. I., Richardson, J. D., et al. (2002). Multispacecraft measurements of plasma and magnetic field variations in the magnetosheath: Comparison with spreiter models and motion of the structures. *Planet. Space Sci.* 50, 601–612. doi:10.1016/S0032-0633(02)00039-9

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 16 March 2023
DOI 10.3389/fspas.2023.1105797

Check for updates

OPEN ACCESS

EDITED BY
Teimuraz Zaqarashvili,
University of Graz, Austria

REVIEWED BY
Jiansen He,
Peking University, China
Vladimir Florinski,
University of Alabama in Huntsville,
United States

*CORRESPONDENCE
Man Zhang,
mzhang@spaceweather.ac.cn
Huichao Li,
lihuichao@hit.edu.cn

SPECIALTY SECTION
This article was submitted to Stellar and Solar Physics, a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 23 November 2022
ACCEPTED 21 February 2023
PUBLISHED 16 March 2023

CITATION
Zhang M, Feng X, Li H, Xiong M, Shen F, Yang L, Zhao X, Zhou Y and Liu X (2023), Numerical modeling of solar wind and coronal mass ejection in the inner heliosphere: A review.
Front. Astron. Space Sci. 10:1105797.
doi: 10.3389/fspas.2023.1105797

COPYRIGHT
© 2023 Zhang, Feng, Li, Xiong, Shen, Yang, Zhao, Zhou and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Numerical modeling of solar wind and coronal mass ejection in the inner heliosphere: A review

Man Zhang[1]*, Xueshang Feng[1,2,3], Huichao Li[3]*, Ming Xiong[1], Fang Shen[1,2,3], Liping Yang[1], Xinhua Zhao[1], Yufen Zhou[1] and Xiaojing Liu[1]

[1]SIGMA Weather Group, State Key Laboratory for Space Weather, National Space Science Center, Chinese Academy of Sciences, Beijing, China, [2]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, China, [3]HIT Institute of Space Science and Applied Technology, Shenzhen, China

The predictions of plasma parameters in the interplanetary medium are the core of space weather forecasts, and the magnetohydrodynamics (MHD) numerical simulation is an important tool in the prediction of plasma parameters. Operational space weather forecasts are commonly produced by a heliosphere model whose inner boundary is set at 18 R_s or beyond. Such predictions typically use empirical/physics-based inner boundary conditions to solve the MHD equations for numerical simulation. In recent years, significant progress has been made in the numerical modeling of the inner heliosphere. In this paper, the numerical modeling of solar wind and coronal mass ejection in the inner heliosphere is reviewed. In particular, different inner boundary conditions used in the simulation are investigated since the MHD solutions are predetermined by the treatment of the inner boundary conditions to a large extent. Discussion is made on further development of the heliosphere model.

KEYWORDS

numerical simulation, solar wind, coronal mass ejection, heliosphere, magnetohydrodynamics

1 Introduction

Over the past few decades, the effects of space weather on technology have become an important field of research. The most serious space weather impact in near-Earth space is usually related to interplanetary coronal mass ejections (ICMEs), which are affected by the surrounding solar wind when propagated to Earth orbit (Wu et al., 2006; Kilpua et al., 2019; Li et al., 2020). For this reason, solar wind condition is a necessary prerequisite for the propagation of CMEs, which governs the propagation of ICMEs in the interplanetary space. The prediction of solar wind and ICME plasma properties plays a crucial role in the space weather forecast. However, the *in situ* observations of plasma parameters are only applicable to several points where spacecraft are located, for example, the Advanced Composition Explorer (ACE) (Stone et al., 1998), Wind (Bochsler et al., 1995), Parker Solar Probe (PSP) (Fox et al., 2016), Solar Orbiter (SO) (Müller et al., 2020), and Solar Terrestrial Relations Observatory (STEREO) spacecraft (Howard et al., 2008; Kaiser et al., 2008). Therefore, we have to rely on numerical technology to predict the plasma parameters.

Currently, a wide variety of techniques have been developed to generate plasma parameters in the interplanetary space. In particular, magnetohydrodynamics (MHD)

simulations are important tools in this endeavor (Pizzo, 1982; Usmanov, 1993; Odstrcil, 2003; Detman et al., 2006; Nakamizo et al., 2009; Hayashi, 2012; Riley et al., 2012; Feng et al., 2014; Zhang and Feng, 2015; Shiota and Kataoka, 2016; Zhang and Xueshang, 2016; Pomoell and Poedts, 2018; Shen et al., 2018; Scolini et al., 2019; Feng, 2020; Scolini et al., 2020; Singh et al., 2020). The MHD simulations can be classified into two types: the full physics-based models and the hybrid empirical/physics-based models. The full physics-based models solve the MHD simulation from the solar surface to 1 au or beyond, and on the other hand, the inner boundary of the hybrid empirical/physics-based models is usually beyond the corona region, with a set of empirical formulas to specify the solar wind distribution at the lower boundary. Since the plasma β in the lower corona is low, the time step determined by the Courant–Friedrichs–Levy condition is small in the corona region. In this sense, the full physics-based models are computationally expensive and the hybrid empirical/physics-based models have their merits from the perspective of forecasting application because of the inexpensive computational cost and relative simplicity of implementation. Also, Owens et al. (2008) showed that the hybrid empirical/physics-based MHD model can reproduce large-scale solar wind structures comparable with the full physics-based models. The purpose of this study is to review the development status of numerical simulation for solar wind and coronal mass ejection in the inner heliosphere, and in particular, we investigate the different inner boundary conditions used in the models.

2 Numerical simulation of solar wind in the inner heliosphere

The lower boundary of the heliospheric models is usually located around 0.1 au, where the solar wind becomes supersonic and super-Alfvénic. Thus, the MHD solutions are predetermined by the treatment of the inner boundary conditions to a large extent. Different kinds of lower-boundary conditions existed in the inner heliosphere models.

In general, the radial component of the magnetic field B_r at the lower boundary of the heliospheric models is usually obtained by the potential field source surface (PFSS) model (Altschuler and Newkirk, 1969; Schatten et al., 1969; Mackay and Yeates, 1999). In the PFSS model, a source surface (around 2.5 R_s) is defined, and the magnetic field between the solar surface and the source surface is assumed to be current-free and becomes open and purely radial above the source surface. With the magnetograph observations of the photosphere as the input, the PFSS model can rapidly extrapolate the large-scale coronal magnetic structures with reasonable accuracy. However, magnetic fields derived from the PFSS model have no thin heliospheric current sheet (HCS) or Parker spiral in the interplanetary space. To solve this problem, some authors combined the PFSS model with the Schatten current sheet (SCS) model (Schatten, 1971). The SCS model takes the absolute value of the (radial) field on the source surface obtained from the PFSS solution as a new lower-boundary condition, the new potential field is solved between the source surface and infinity, and then, the proper sign is restored. Another method for reconstructing the coronal magnetic field is the current sheet source surface (CSSS) model (Zhao and Hoeksema, 1995) that explicitly takes into account additional sheet currents. The CSSS model includes a cusp surface height R_{CS} and a source surface height R_{SS}, where the magnetic field becomes open above R_{CS} and becomes radial above R_{SS}. The Ulysses observations show that the strength of the interplanetary magnetic field B_r has no dependence on latitude. However, the large-scale structures of the coronal magnetic field derived from the PFSS model have a systematic gradient of B_r in the latitudinal direction. The CSSS model can better reproduce the latitude-invariant nature of B_r.

Based on the coronal magnetic topology parameters obtained from the previous models, the Wang–Sheeley (WS) or Wang–Sheeley–Arge (WSA)-type models (Arge and Pizzo, 2000; Arge et al., 2004; Mcgregor et al., 2011) are the most widely used empirical models that predict the solar wind velocity at the lower boundary. The specific form of V_r (km/s) in the WS model can be written as

$$V_r(f_s) = V_{\min} + \frac{V_{\max}}{f_s^{\alpha}}, \tag{1}$$

where f_s is the magnetic expansion factor which reads $f_s = (\frac{1}{R})^2 \frac{B_{R_s}}{B_R}$, where B_{R_s} and B_R are the magnetic field strength at the solar surface and at the source surface, respectively. The distribution of f_s at the lower boundary of the heliospheric models is the same as its distribution at the source surface. $V_{\min}$ sets the minimum possible solar wind speed, while $V_{\max}$ sets the maximum possible solar wind speed. α determines the effect of f_s on V_r. By including an additional variable θ_b, the specific form of V_r (km/s) in the WSA model can be written as

$$V_r(f_s,\theta_b) = V_{\min} + \frac{V_{\max}}{(1+f_s)^{a_1}}\left[1 - 0.8\exp\left(-\left(\frac{\theta_b}{a_2}\right)^{a_3}\right)\right]^{a_4}, \tag{2}$$

where θ_b denotes the minimum angular separation between an open-field foot point and its nearest coronal hole boundary. The angle a_2 and exponent a_3 determine the angular extent and influence of the open-flux boundary layer, respectively. Any of the parameters in Eqs **1**, **2** can be modified.

The other components of the magnetic field and velocity, such as the meridional magnetic field B_θ and the meridional flow velocity V_θ, are always set to 0. The azimuthal flow velocity V_ϕ is always set to zero in the inertial frame and set as $\Omega_s R_{\rm gb} \sin\theta$ in the rotating frame, with Ω_s denoting the solar angular speed and $R_{\rm gb}$ standing for the lower boundary of the heliospheric models. The azimuthal magnetic field B_ϕ is determined by $B_\phi = (B_r/V_r)\,(V_\phi - \Omega_s R_{gb} \sin\theta)$. Other solar wind parameters, including the density and temperature, are always prescribed by an assumption of the constant momentum flux and total pressure (sum of thermal and magnetic pressures). In the following section, the lower-boundary conditions in the existing inner heliospheric models are introduced in detail.

2.1 Boundary conditions based on the PFSS model

Using the PFSS + WS model, Odstrcil (2003) modeled the ambient solar wind from 0.1 au to 1 au using the WS-ENLIL mode. The solar wind velocity at 0.1 au was calculated according to Equation **1** with $V_{\min}$ = 285 km/s, $V_{\max}$ = 575 km/s, and α = 1/1.7. The mass density ρ was determined by an assumption of the

constant momentum flux, and the temperature T was chosen to assure the total pressure is uniform on the source surface. Hayashi (2003) reconstructed three-dimensional solar wind structure from 50 R_s to 1 au. The radial flow velocity V_r at the inner boundary was determined by solar wind data obtained from interplanetary scintillation (IPS) observations, and the number density and temperature were determined by the empirical relation obtained from the Helios data through least-squares fitting, $N(V_r) = 62.98 + 866.4(\frac{V_r - 154.9}{100})^{-3.402}$ and $T(V_r) = -0.455 + \frac{0.1943 V_r}{100}$. They found that the solar wind research and space weather prediction algorithms can be enhanced by the combination of MHD simulation and IPS tomography. Based on that work, Hayashi (2012) modeled time-dependent solar wind from 50 R_s up to 1,250 R_s. The solar wind was simulated with the time-dependent boundary condition. They applied temporal linear interpolation from one Carrington rotation to the next Carrington rotation. Shiota et al. (2014) developed an inner heliospheric MHD solar wind model for practical use in real-time space weather forecasting. The numerical domain was set to $25R_s \le r \le 425R_s$. The radial flow velocity was determined using Equation **1** with $V_{min} = 267.5$ km/s, $V_{max} = 410$ km/s, and $\alpha = 0.4$. The number density and temperature were obtained using empirical relationships that have been used by Hayashi (2003) and Hayashi (2012), respectively. Since Hayashi (2003) and Hayashi (2012) showed the relation at 50 R_s, their model specified the boundary condition at 25 R_s. For modification, the number density was multiplied by 4 assuming constant velocity propagation from 25 R_s to 50 R_s, and the temperature was multiplied by $4^{\gamma-1}$ with the assumption of polytropic expansion.

Based on the PFSS + WSA model, Merkin et al. (2011); Pahud et al. (2012); and Merkin et al. (2016) simulated the solar wind from 0.1 au to 1 au or beyond using the Lyon–Fedder–Mobarry (LFM) MHD simulation code. This new version of the code is referred to as LFM-helio. The values of V_r at the inner boundary were all calculated according to Equation **2** but with different parameters. In Merkin et al. (2011), $V_{min} = 200$ km/s, $V_{max} = 750$ km/s, $a_1 = 2/9$, $a_2 = 3.8°$, $a_3 = 3.6$, and $a_4 = 3$. The number density and temperature were determined by using an empirical fit to Helios data: $N[\mathrm{cm}^{-3}] = 112.64 + 9.49 \cdot 10^7/(V_r[km/s])^2$, $T = N_0 T_0/N$, where $T_0 = 8 \cdot 10^5$ K and $N_0 = 300\mathrm{cm}^{-3}$. **Figure 1** shows the radial velocity and magnetic field at the inner boundary of the simulation (0.1 au) (Merkin et al., 2011). The magnetic field at 0.1 au was scaled down by a factor 2/3 × 100.

The inner boundary conditions used by Pahud et al. (2012) were similar to those used by Merkin et al. (2011), except for the following parameters: $V_{min} = 240$ km/s, $V_{max} = 675$ km/s, $a_1 = 2/9$, $a_2 = 2.8°$, $a_3 = 1.25$, and $a_4 = 3$. By comparing the MHD results with MESSENGER and ACE spacecraft observations, Pahud et al. (2012) found that the uncertainty in the specification of the boundary conditions, rather than a poor performance of the solar wind model, led to the discrepancies between *in situ* measurements and simulations. Merkin et al. (2016) used the LFM-helio MHD model to explore the heliospheric consequences of time-dependent changes at the Sun. The Air Force Data Assimilate Photospheric flux Transport (ADAPT) model was used to obtain daily updated photospheric magnetograms. These time-dependent magnetograms were then used to obtain the solar wind parameters at 21.5 R_s with the WSA model. In particular, the changes of the longitudinal and latitudinal components of the magnetic field at the inner boundary were induced by the tangential electric fields, which were derived by a Poisson equation. The V_r at the inner boundary was calculated according to Equation **2** with $V_{min} = 240$ km/s, $V_{max} = 675$ km/s, $a_1 = 2/9$, $a_2 = 1.9°$, $a_3 = 2$, and $a_4 = 3$. The number density was determined in the same way as in Merkin et al. (2011). The temperature was defined from the approximate full pressure balance in the angular directions.

$$NkT + \frac{B_r^2}{8\pi} = N_{\mathrm{slow}} k T_{\mathrm{slow}}, \tag{3}$$

where k is the Boltzmann constant, $N_{\mathrm{slow}} = \max(N)$, and $T_{\mathrm{slow}} = 5 \cdot 10^5$ K is the nominal slow wind plasma temperature. In some cases, one can specify the fast wind temperature parameter T_{fast} instead to avoid a negative temperature:

$$NkT + \frac{B_r^2}{8\pi} = N^* k T_{\mathrm{fast}} + \frac{B_{max}^2}{8\pi}, \tag{4}$$

where $T_{\mathrm{fast}} = 10^6$ K, $B_{\max}$ is the maximum magnetic field located somewhere in the fast solar wind, and N^* represents the plasma density at that location.

Shen et al. (2018) and Wang et al. (2020) used a new boundary treatment to model the solar wind in interplanetary space with MHD equations. The magnetic field was derived by the PFSS model combining with the OMNI database.

$$B_r = \mathrm{sign}\left(B^{\mathrm{PFSS}}\right) \times \frac{1}{\sqrt{2}} \mathrm{mean}\left(B^{1\ \mathrm{au}}\right) \left(\frac{1\mathrm{au}}{R_{gb}}\right)^2, \tag{5}$$

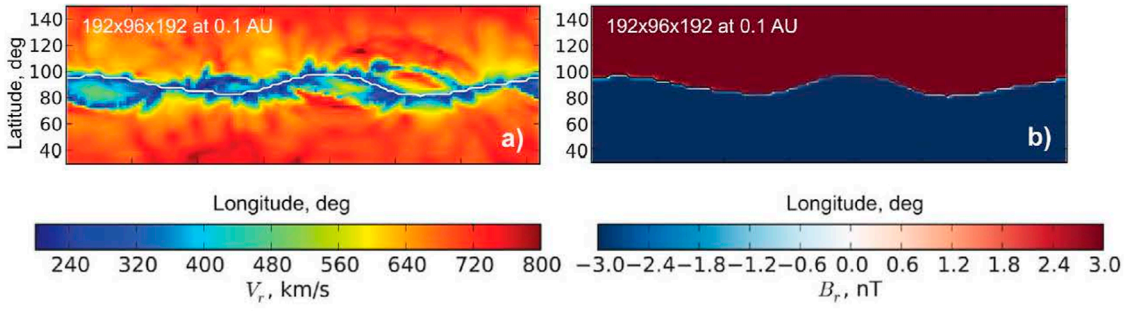


FIGURE 1
Radial velocity **(A)** and magnetic field **(B)** at the inner boundary of the simulation (0.1 au) (Merkin et al., 2011).

where R_{gb} = 0.1 au. The average value of the observed magnetic field, mean(B^{1au}), was determined by the past three Carrington rotations (CRs) at 1 au from the OMNI database. The solar wind velocity at 0.1 au was obtained from Equation 2. V_{max} = 675 km/s, a_1 = 0.22 rad, a_3 = 1.0, and a_4 = 1.0 were constants over time, while V_{min} and a_2 can be adjusted with different solar cycles. The number density was calculated by assuming that the solar wind energy flux is constant over the solar cycle.

$$N = N_{1au}\left(\frac{1\,\text{au}}{R_{gb}}\right)^2 V_{1au}\left(\frac{1}{2}V_{1au}^2 + \frac{GM_s}{Rs}\right)\left[V_r\left(\frac{1}{2}V_r^2 + \frac{GM_s}{Rs}\right)\right]^{-1}, \tag{6}$$

where V_{1au} = 750 km/s is velocity at 1 au and N_{1au} is deduced from the average solar wind energy flux based on the OMNI observations. Various studies showed that there are linear or quadratic relationships between T and V at 1 au (Lopez and Freeman, 1986; Richardson and Cane, 1995; Elliott et al., 2005; Verbanac et al., 2011; Chat et al., 2012). The relationship of T_p and V_r at the lower boundary could be obtained as follows:

$$T_p = \frac{1}{2}V_r^2\left(\frac{1\,\text{au}}{R_{gb}}\right)^{2\gamma-1}. \tag{7}$$

Then, the pressure was obtained by $p = 2NKT$. **Figure 2** presents the maps of solar wind parameters for CR2053 at the lower boundary of 0.1 au (Shen et al., 2018).

Gonzi et al. (2021) investigated the impact of inner heliospheric boundary conditions on solar wind predictions at Earth. The radial component B_r was computed using two methods, the PFSS model or the magnetofrictional (MF) model. The solar wind velocity was obtained from three different methods. The first was the WSA model in Equation 2 using V_{min} = 240 km/s, V_{max} = 675 km/s, a_1 = 2/9, a_2 = 2.8°, a_3 = 1.25, and a_4 = 3. The second was a modified DCHB model $V_r(\theta_b) = V_{min} + (V_{max} - V_{min})(\theta_b \cdot \omega)^{\delta}$, and the parameters were chosen by comparing model-predicted solar wind speed histograms at 1 au with observations V_{min} = 200 km/s, V_{max} = 700 km/s, ω = 7 radian, and δ = 1.25. The third was determined by the interplanetary scintillation solar wind data. The number density was determined from the interplanetary scintillation solar wind data or the assumptions of the constant momentum flux NV^2, $N = N_s\left(\frac{V_s}{V_r}\right)^2$. The temperature was set as $T = T_s\frac{N_s}{N}$. Specifically, V_s = 300 km/s, N_s = 500cm^{-3}, and T_s = 500 kK in the slow stream. The results showed that the selection of the coronal magnetic field model had a significant impact on the model results, and IPS data had a high success rate in reproducing high-speed solar wind.

2.2 Boundary conditions based on the PFSS + SCS model

Detman et al. (2006, 2011) introduced a Hybrid Heliospheric Modeling System (HHMS) for background solar wind to aid in the operational forecasting of geomagnetic activity, where the lower boundary was also set at 0.1 au. The radial magnetic field was provided by the PFSS + SCS model. The principle in determining the solar wind speed at 0.1 au was similar to the WS model given by

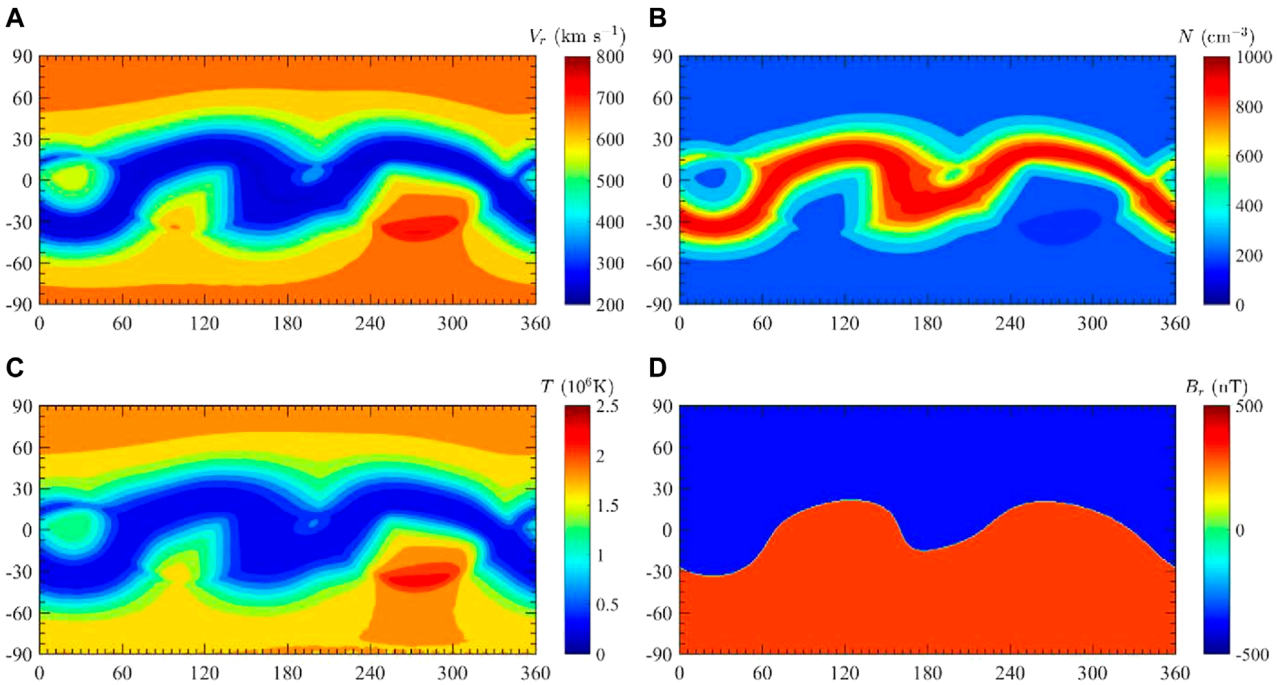


FIGURE 2
Maps of solar wind parameters at the lower boundary of 0.1 au for CR 2053. **(A–D)** The panels show the radial velocity Vr (km/s), number density N (cm^{-3}), temperature T (K), and radial magnetic field Br (nT) (Shen et al., 2018).

the equation

$$V_r(f_s) = V_{\min} + (V_{\max} - V_{\min}) / \left(1 + \frac{f_s}{f_0}\right)^{V_{xp}}, \quad (8)$$

where $V_{\max}$ and $V_{\min}$ were set at 700 and 200 km/s, respectively. The parameter $V_{xp} = 1.5$ determined the sharpness of the fast–slow transition in the relationship. The parameter f_0 scaled the parameter f_s, which was adjustable to match the observed solar wind speeds. The mass density $\rho = \frac{F_{\text{mass}}}{V_r}$ was determined by the assumption of an constant mass flux, F_{mass} was obtained by taking observed values of ρV_r at Earth and applying an r^2 scaling to 0.1 au. The temperature $T = \frac{p_{\text{total}} - \frac{\mathbf{B}^2}{8\pi}}{2\rho R_{\text{gas}}}$ was determined by an assumption of the uniform total pressure p_{total} on the source surface, and p_{total} was a free parameter that required tuning. **Figure 3** shows the block diagram of the HHMS in Detman et al. (2006).

Narechania et al. (2021) described a 3D MHD-based heliospheric model based on a semi-empirical data-driven approach. The inner boundary of the heliospheric model was set at 25 R_s, the magnetic field was derived by the PFSS + SCS model as in Detman et al. (2006), and the velocity was provided by Equation **2** with $V_{\min} = 250$ km/s, $V_{\max} = 850$ km/s, $a_1 = 0.2$, $a_2 = 2.6°$, $a_3 = 1.25$, and $a_4 = 2.5$. The remaining parameters were determined as in Shiota et al. (2014).

Pomoell and Poedts (2018) presented a new space weather forecasting-targeted inner EUropean Heliospheric FORecasting Information Asset (EUHFORIA) model. The plasma parameters at 0.1 au were obtained by using the empirical solar wind model. The radial component B_r was computed as

$$B_r = sgn(B_{\text{corona}}) B_{\max} (V_r / V_{\max}), \quad (9)$$

where $\text{sgn}(B_{\text{corona}})$ is the sign of the magnetic field as given by the PFSS + SCS model. $B_{\max} = 300$ nT. The boundary value for the solar wind velocity was calculated according to Equation **2** with $V_{\min} = 240$ km/s, $V_{\max} = 675$ km/s, $a_1 = 2/9$, $a_2 = 0.02$ rad, $a_3 = 1.25$, and $a_4 = 3$. A constant value of 50 km/s is subtracted from the solar wind velocity to avoid systematically overestimating the wind speed. The number density is given as follows:

$$N = N_{\text{fsw}} (V_{\max} / V_r)^2, \quad (10)$$

where $N_{\text{fsw}} = 300\text{cm}^{-3}$ is the number density of the fast solar wind. The temperature is given as $T = T_{\text{fsw}} (\rho_{\text{fsw}} / \rho)^2$ with $T_{\text{fsw}} = 0.8$ MK. **Figure 4** shows the sequence of steps that provides the boundary conditions for the heliospheric MHD model (Pomoell and Poedts, 2018).

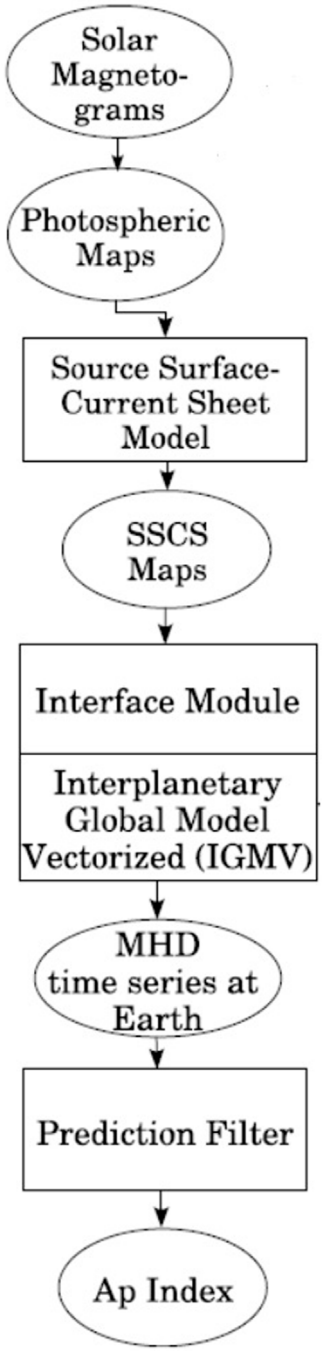


FIGURE 3
Block diagram of the HHMS (Detman et al., 2006).

2.3 Boundary conditions based on the CSSS model

The inner heliospheric solar wind was simulated from 21.5 R_s to 1 au with the MHD code CRONOS in Wiengarten et al. (2013) and Wiengarten et al. (2014). In Wiengarten et al. (2013), the CSSS model was used to obtain B_r at 21.5 R_s. The magnetic flux at the solar surface was obtained by the solar surface flux transport (SFT) model, which was then extrapolated to 10 solar radii by the CSSS model with $R_{CS} = 1.55R_s$ and $R_{SS} = 10R_s$. Then, the radial magnetic field at the source surface was scaled to the lower boundary at 21.5 R_s by using the scale factor r^{-2}. In Wiengarten et al. (2014), the magnetic field at 0.1 au was derived by the PFSS + SCS model as in Detman et al. (2006). A Finite Difference Iterative Poisson Solver (FDIPS) was used to obtain the PFSS solution to avoid the numerical artifacts generated by the usual spherical harmonics expansion of the coronal potential field. At 0.1 au, the radial magnetic field was given a value of 300 nT, while keeping the

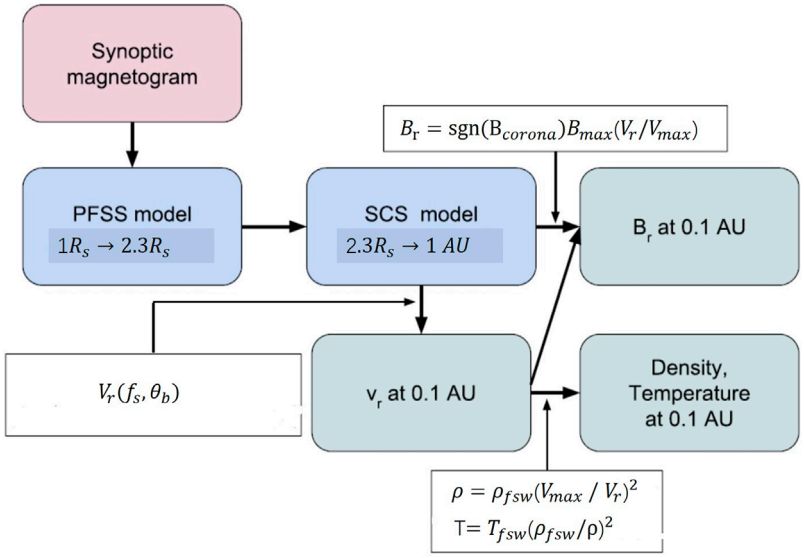


FIGURE 4
Sequence of steps that provide the boundary conditions for the heliospheric MHD model (Pomoell and Poedts, 2018).

orientation derived by the PFSS + SCS model. The radial velocity at 0.1 au was obtained with different forms for Wiengarten et al. (2013) and Wiengarten et al. (2014). Wiengarten et al. (2013) used $V_r = V_{min} + \frac{V_{del}}{f_s^{V_{xp}}} + \alpha_1 \cdot (\frac{\theta_b}{f_s^{V_{xp}/2}} - \alpha_2)$, where $V_{\min} = 154$ km/s, $V_{del} = 300$ km/s, $V_{xp} = 3$, $\alpha_1 = 7.4$ km/s, and $\alpha_2 = 3.5$. The velocity V_r in Wiengarten et al. (2014) was calculated according to Equation 2 with $V_{\min} = 200$ km/s, $V_{\max} = 675$ km/s, $a_1 = 2/9$, $a_2 \in (2.8°, 3.2°)$, $a_3 = 2$, and $a_4 = 3$. The remaining MHD quantities at the inner boundary were all similar to that described in Detman et al. (2006).

Li et al. (2020) simulated the interplanetary B_z using a data-driven heliospheric solar wind model. Following Merkin et al. (2016), the time-dependent boundary conditions were also used in their study. The radial component B_r at the inner boundary was derived from the CSSS model with $R_{CS} = 2.5R_s$ and $R_{SS} = 15R_s$. The value of B_r given by the CSSS model was scaled by a factor $b_{up} = \Phi_{OB}/\Phi_{CSSS}$ to provide enough magnetic flux at the inner boundary, where Φ_{CSSS} was the magnetic flux from the original CSSS solution and $\Phi_{OB} = 4\pi r_{1au}^2 \mathrm{mean}(B^{1\,au})/\sqrt{2}$ was the observed average magnetic flux at 1 au derived from OMNI data. The WSA model (Equation 2) was used to determine the solar wind velocity with $V_{\min} = 240$ km/s, $V_{\max} = 560$ km/s, $a_1 = 2/9$, $a_2 = 2°$, $a_3 = 1$, and $a_4 = 3$. As in Detman et al. (2006), the density was obtained by an assumption of a constant mass flux. The distribution of temperature at the heliobase was similar to that used in Shen et al. (2018) and Wang et al. (2020).

2.4 Boundary conditions based on *in situ* measurements

Biondo et al. (2021) described a new approach to determine the plasma parameters at the inner boundary of 0.1 au. This scheme applied a back reconstruction technique to remap the *in situ* measurements acquired at 1 au into the inner heliosphere. The observation data at 1 au were thought as a ring of L cells.

$$\left[\phi_i(rad), N_i(cm^{-3}), V_{r,i}(km/s), B_{r,i}(nT), B_{\phi,i}(nT)\right]. \tag{11}$$

The Parker Spiral can be used to obtain their respective cells of origin $(0.1au; \phi_i')$.

$$\phi_i' = \phi_i + \frac{\Omega_s}{v_{r,i}}(1au - 0.1au). \tag{12}$$

The radial speed and numerical densities at 0.1 au can be obtained as follows:

$$V_{r,i}' = V_{r,i},$$

$$V_{\theta,i}' = 0,$$

$$V_{\phi,i}' = -\Omega_s, r$$

$$N_i' = N_i\left(\frac{1au}{0.1au}\right)^2.$$

The treatment of the magnetic field was similar.

$$B_{r,i}' = B_{r,i}\left(\frac{1au}{0.1au}\right)^2,$$

$$B_{\theta,i}' = 0,$$

$$B_{\phi,i}' = B_{\phi,i}\left(\frac{1au}{0.1au}\right) + B_{r,i}\frac{\Omega_s}{V_{r,i}}\left[(1au - R_{ss}) - (0.1au - R_{ss})\left(\frac{0.1au}{1au}\right)^2\right].$$

This method could generate an interplanetary spiral reconstruction well with observations, and additionally, many small-scale features were also generated. **Figure 5** presents the steps involved in RIMAP for the reconstruction of heliospheric plasma conditions (Biondo et al., 2021).

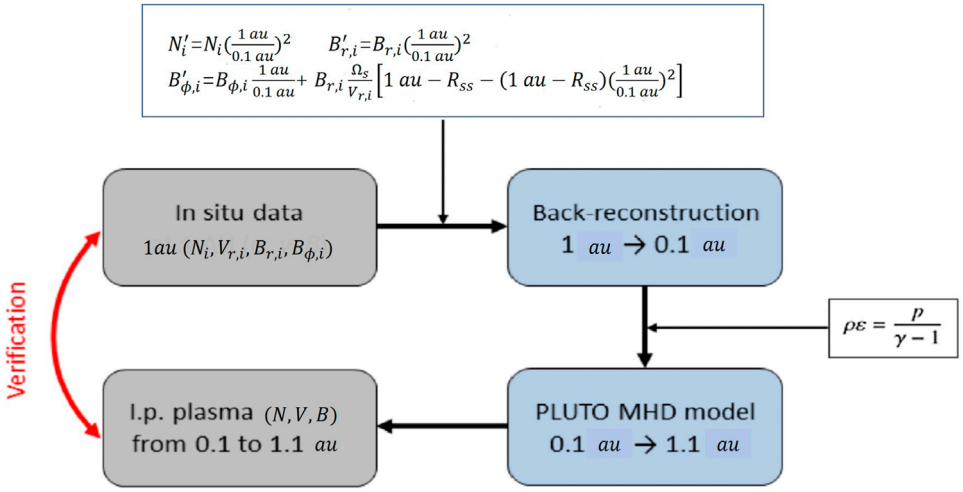


FIGURE 5
Schematic of the steps involved in RIMAP for the reconstruction of heliospheric plasma conditions starting from *in situ* measurements (Biondo et al., 2021).

2.5 Numerical simulation of solar wind with the Heliospheric Upwind eXtrapolation model

To provide a computationally efficient solution, the Heliospheric Upwind eXtrapolation (HUX) model (Riley and Lionello, 2011) was widely used to extrapolate velocity from near the Sun to 1 au without providing physical insight. The HUX model was essentially a 1D extrapolation with velocity in a radial direction.

$$V_{r+1,\phi} = V_{r,\phi} + \frac{\Delta r \Omega_s}{V_{r,\phi}}\left(\frac{V_{r+1,\phi} - V_{r,\phi}}{\Delta\phi}\right), \tag{13}$$

with Δr and $\Delta\phi$ representing the grid spacing in radial and longitude directions, respectively. A comparative study of extrapolation models and empirical relations in forecasting solar wind was made in Kumar et al. (2020). The PFSS + SCS model was used to extrapolate magnetic fields. The velocity was obtained using the WS model or the WSA model. Then, the HUX model and a physics-based model PLUTO (Mignone et al., 2007) were used to extrapolate these velocities into the inner heliosphere zone, and the different magnetic field extrapolation models combined with velocity formulations were compared to predict the solar wind properties at L1. **Figure 6** shows the sequence of steps used for each of these models.

They showed that the WSA model captures the contrast between the slow and fast solar wind better than the WS model. Additionally, the PFSS + SCS magnetic field extrapolation model combined with the WSA model had the best performance in all the cases.

Mayank et al. (2022) used a physics-based inner heliospheric model to forecast the ambient solar wind from 0.1 au to 2.1 au. The three-dimensional MHD equations were solved using the PLUTO code. The properties in the inner heliospheric model were derived from the PFSS + SCS model in combination with the WSA model (Equation **2**). Additionally, the HUX algorithm was used to find the optimal set of values of independent parameters in the WSA model. The following parameters were chosen: $V_{\min}$ = 250 km/s, $V_{\max}$ = 650 km/s, a_1 = 2/9, a_2 = median of θ_{bE}, a_3 = 1.25, and a_4 = 3. θ_{bE} was calculated on the field lines that reached the location of Earth. Then, the WSA speed profile was rotated in the longitudinal direction by angle κ to account for solar rotation.

$$\kappa = 5 + \left(\frac{2\pi}{\Omega_s}\right)\left(\frac{21.5R_s}{(V_r)_{\min}}\right). \tag{14}$$

Ω_s ranged from 27.21 to 27.34 days, depending on the location of Earth in its orbit. The radial component B_r and plasma number density were similar to those used by Pomoell and Poedts (2018). The plasma thermal pressure was chosen to be constant on the boundary and equal to p = 6.6 nPa. **Figure 7** presents the process flow diagram of the proposed solar wind model showing the range of numerical models involved in the inner heliospheric model.

The time-dependent form of the HUX model, which was referred to the HUXt model, was used to give solar wind velocity at Earth in Owens et al. (2020) and Bunting and Morgan (2022). Owens et al. (2020) set the inner boundary at 30 R_s, and the computed V_r at 30 R_s from the Magnetohydrodynamics Algorithm outside a Sphere (MAS) coronal model serves as the input to HUXt. They suggested that HUXt can act as a surrogate for full three-dimensional MHD models when very large ensembles are required. In Bunting and Morgan (2022), the inner boundary condition of the HUXt model was based on the coronal plasma density gained from coronagraph observations. The solar wind velocity was obtained using a simple linear relationship as follows:

$$V = V_{\max} - \left[\left(\frac{N - N_{\min}}{N_{\max} - N_{\min}}\right)(V_{\max} - V_{\min})\right], \tag{15}$$

where $N_{\max}$ and $N_{\min}$ are the maximum and minimum densities in the equatorial plane, respectively. The ambient solar wind velocity at Earth was obtained by using the HUXt model. Compared to an HUXt model with a traditional MAS as the input, the results showed that the tomography/HUXt model can predict solar wind velocity much better.

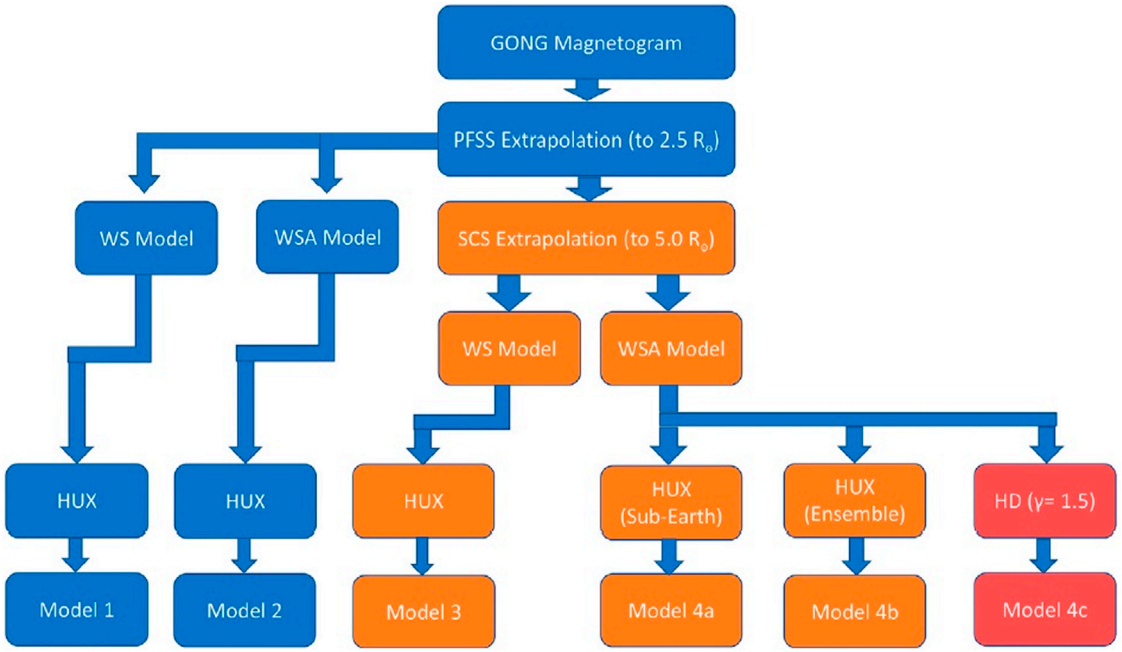


FIGURE 6
Flowchart of the models utilized in the work (Kumar et al., 2020).

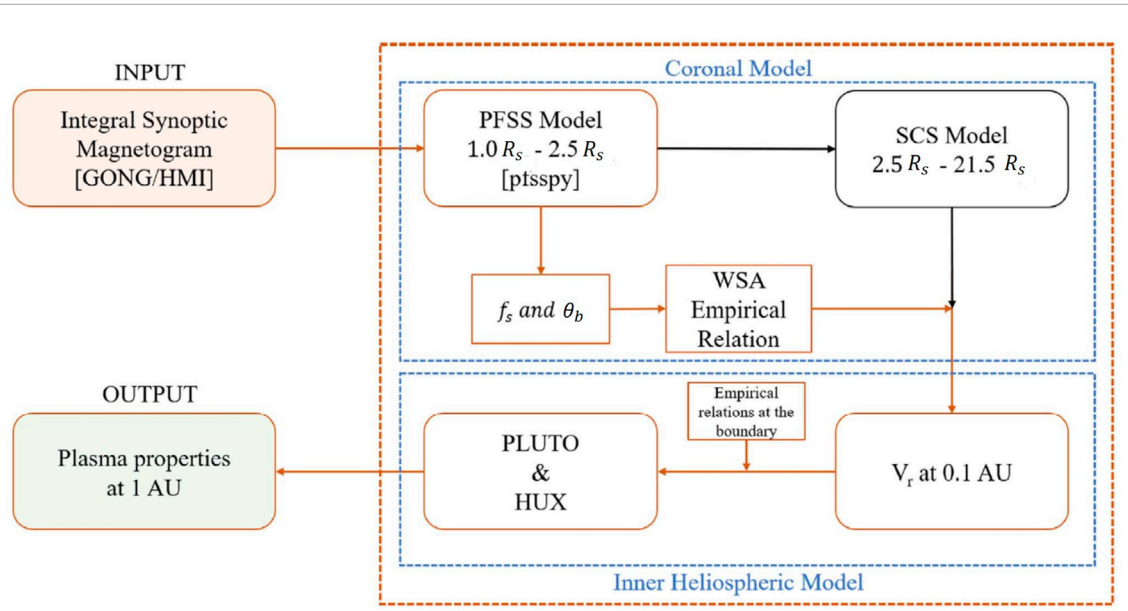


FIGURE 7
Process flow diagram of the proposed solar wind model showing the range of numerical models involved in the inner heliospheric model (Mayank et al., 2022).

Overall, various inner boundary conditions have been developed in recent years, which provide important input conditions for space weather prediction. Table 1 summarizes the various methods to set the solar wind parameters at a lower boundary. The MHD model and the HUX model have been widely used for numerical simulation of solar wind. The HUX model can act as a surrogate for full three-dimensional MHD models when very large ensembles are required. However, it can only be used to obtain the velocity in the interplanetary space. Thus, the MHD model is necessary when we want to get other parameters, such as magnetic field, temperature, and density. For MHD simulation, these variables are all needed to be determined at the lower boundary

initially.The magnetic field and velocity distribution are important because temperature and density are always determined from them. To calculate the magnetic field distribution, PFSS + SCS may be a better choice because the magnetic fields derived from the PFSS model have no thin HCS or Parker spiral in the interplanetary space. To better reflect the latitude-invariant nature of B_r, the CSSS model can be used. For calculating velocity distribution, the WSA model is better than the WS model in capturing the contrast between the slow and fast solar wind. The IPS observation has also been used to obtain velocity distribution, and the results show that IPS data has high success rate in detecting high-speed solar wind and the space weather prediction can be enhanced by the combination of MHD simulation and the IPS data.

It can be seen that the distribution of the inner boundary is fixed in many studies, the numerical processing on this fixed boundary is simple, and the steady-state solar wind can well represent the background conditions of the short-term transient phenomena using this fixed boundary. Since some studies used time-dependent boundary (Hayashi, 2012; Merkin et al., 2016; Li et al., 2020), it is expected that time-dependent boundary conditions can simulate the inner heliosphere more accurately and in real time. Fixed boundary conditions can accommodate observational data made in only one CR period. However, time-dependent boundary conditions can determine the MHD state of the solar wind at distant regions from the Sun, which are determined in a longer time period than one CR. Also, the heliospheric consequences of time-dependent conditions can be explored, that is, what kind of plasma and magnetic field structures are created due to the changing boundaries of coronal holes. Merkin et al. (2016) showed that time-dependent boundary conditions can reproduce the gross-scale structure of the heliosphere with higher fidelity, and provide important insights to interpret data on smaller spatial and faster time scales (e.g., 1 day). Thus, the time-dependent boundary condition is a promising direction of research both for space weather applications and fundamental physics of the heliosphere. For the time-dependent boundary, how to maintain the divergence-free condition is one difficulty. Since the time-dependent radial magnetic field $\frac{\partial B_r}{\partial t}$ is always obtained by observations, we need to derive an electric field tangential to the inner boundary spherical surface that would evolve the radial magnetic field. To do that, the Helmholtz theorem is applied on a sphere and the tangential electric field is expressed as the sum of the curl and the gradient of scalar potentials.

$$\mathbf{E}_t = \nabla_t \times \psi \mathbf{r} + \nabla_t \Phi. \tag{16}$$

Here, ψ is determined by solving the Poisson equation following Faraday's law.

$$\nabla_t^2 \psi = \frac{\partial B_r}{\partial t}. \tag{17}$$

Because there is no available information to determine Φ, Φ is always set to 0. Then, the transverse components of the magnetic field in the corotational frame can be calculated as follows:

$$B_\phi = B_r \frac{V_\phi}{V_r} + \frac{E_\theta}{V_r}, \tag{18}$$

$$B_\theta = -\frac{E_\phi}{V_r}. \tag{19}$$

3 Numerical simulation of the coronal mass ejection in the inner heliosphere

CMEs are giant clouds of magnetized plasma erupting from the Sun. When CMEs propagate through the corona and the interplanetary space, they interact with the solar wind plasma or other ICMEs and undergo changes. At larger heliocentric distances, CMEs are known to exhibit rotations and deflections, which can play a crucial role in their geomagnetic impact (Maharana et al., 2022). MHD simulations are often used for CME modeling purposes. There are two categories of MHD simulation models, one is anchored to the low corona and the other is fully embedded into the middle or upper corona. The first category is built through a large number of observations and enables the incorporation of more detailed physics, but the total computation time is much larger. On the other hand, the second category is suitable for space-weather forecasts because of the inexpensive computational cost and relative simplicity of implementation. Also, with the launch PSP and SO, the ICME will be constrained by more observations. In the past decades, many different models have been developed to simulate CME propagation in the heliosphere, such as ENLIL (Odstrcil, 2003), EUHFORIA (Pomoell and Poedts, 2018), SUSANOO-CME (Shiota and Kataoka, 2016), and MS-FLUKSS (Singh et al., 2020). In these models, CMEs are inserted at 0.1 au and evolved self-consistently by solving the MHD equations and a suitable CME initialization model is necessary. In this section, we focus on the initial CME parameters that need to be determined and extrapolated to the lower boundary.

The initial CME parameters rely heavily on CME reconstruction methods. The cone model (Zhao et al., 2002; Xie et al., 2004; Michaek, 2006) and the graduated cylindrical shell (GCS) model (Thernisien et al., 2006, 2009; Thernisien, 2011) are widely used to estimate the CME kinematic and geometric parameters in the past years. The cone model represents CME as a hydrodynamic pulse with a constant angular width, propagation direction, and speed. The angular width and source position of the CME are obtained by matching the modeled halos with the observed halos. The radial velocity and acceleration of the CME are determined by a series of times and radial distances. **Figure 8** shows the cone model topology and relationship between the heliocentric coordinate system and the cone coordinate system (Michaek, 2006). The heliocentric coordinate system is defined such that x_h points to Earth, z_h points north, and y_h, z_h defines the sky plane. The cone coordinate system has the origin at the apex of the cone, z_c is the cone axis, and x_c, y_c defines the plane parallel to the base of the cone. The geometry of the cone can be determined by heliographic longitude ϕ, latitude θ, and angular width ω.For the GCS model, the CME is described as a flux-rope-like structure that expands in a self-similar fashion. The three-dimensional shape of the CME consists of two legs and a curved front. **Figure 9** shows a face-on schematic of the GCS model (Thernisien et al., 2009) and schematic of the detailed geometric parameter (Thernisien, 2011). The dashed–dotted line shows the axis of the model, and the solid line the represents of the shell where the density is placed. The cross section of the model is a group of circular annuli with a gradually varying radii $r_1 = \kappa r$, where r is the distance from a point on the shell to the center of the Sun and κ is a constant depending on the studied event. The angle between the axes of two conical legs is 2ω, and the height of the cone is h. The geometry of the shell can be completely defined with these three parameters. In the

TABLE 1 Various methods used to set the solar wind parameters at the lower boundary.

Solar wind parameters at the lower boundary of the inner heliosphere models						
CME model					HUX model	
Number density	Temperature	Magnetic field	Solar wind velocity		Solar wind velocity	
Constant momentum flux	Total pressure balance	PFSS	WS	Odstrcil 2003	PFSS + SCS combined with the WS/WSA model	Kumar et al., 2020
		PFSS + SCS	WS	Deman et al., 2006, Deman et al., 2011	MAS coronal model	Owens et al., 2020
			WSA	Wiengarten et al., 2014	Coronagraph observation	Bunting and Morgan (2022)
		CSSS	WSA	Wiengarten et al., 2013		
	Empirical relation from velocity	CSSS	WSA	Li et al., 2020		
	Thermal pressure balance	PFSS	WSA/IPS/DCHB	Gonzi et al., 2021		
		PFSS + SCS	WSA	Pomoell and Poedts, 2018; Mayank et al., 2022		
Empirical relation from Helios data	Empirical relation from Helios data	PFSS	IPS data	Hayashi 2003, Hayashi 2012		
			WS	Shiota et al., 2014		
			WSA	Merkin et al., 2011; Pahud et al., 2012		
		PFSS + SCS	WSA	Narechania et al., 2021		
	Total pressure balance	PFSS	WSA	Merkin et al., 2016		
Constant energy flux	Empirical relation from velocity	PFSS	WSA	Shen et al., 2018; Wang et al., 2020		
In situ measurements	*In situ* measurements	*In situ* measurements	*In situ* measurements	Biondo et al., 2021		

10.3389/fspas.2023.1105797

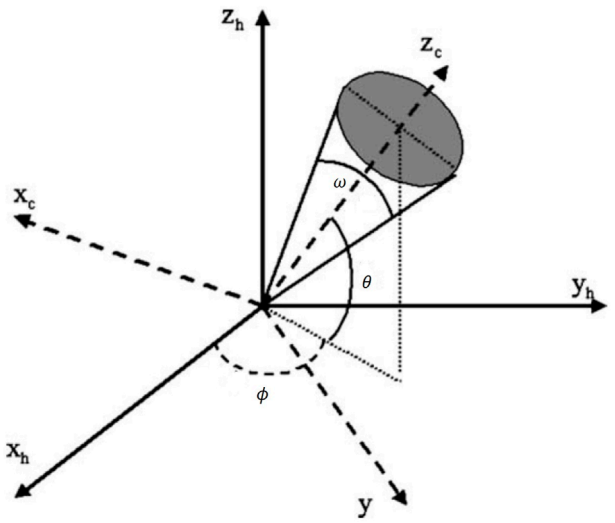


FIGURE 8
Cone model topology and relationship between the heliocentric coordinate system and the cone coordinate system (Michaek, 2006).

following section, lower boundary conditions in the existing inner heliospheric models are introduced in detail.

3.1 Numerical simulation of coronal mass ejection with hydrodynamic plasma cloud

Based on the cone model, the ENLIL model was used to simulate transient heliosphere disturbances (Odstrcil, 2003; Odstrcil et al., 2004; Odstrcil, 2005). A CME was simulated by launching a time-dependent hydrodynamic plasma cloud through the inner boundary of the heliospheric. The cone model was used to describe the location, speed, and angular width of the CME. The CME disturbance had uniform velocity, density, and temperature. The CME's density (temperature) was taken to be four times larger than (equal to) the mean values in the fast stream. With the simple geometry, the ENLIL model had been successfully used to study the global evolution of CMEs in the heliosphere and predicted the CME arrival times at Earth. **Figure 10** presents the distribution of solar wind parameters with the introduction of the input pulse on 12 May 1997 at 1800 UT (Odstrcil, 2005).

With the development of the EUHFORIA heliospheric model (Pomoell and Poedts, 2018), a new CME propagation model simulation was performed. The cone CME model was used to determine the CME angular width, propagation direction, and speed. Then, the CME was introduced to the inner radial boundary at 0.1 au as a time-dependent boundary condition. Specifically, the velocity, density, and temperature for solar wind values were replaced by constant values at $R_{gb} = 0.1$ au.

$$(\theta - \theta_{\mathrm{CME}})^2 + (\phi - \phi_{\mathrm{CME}})^2 < \alpha(t)^2,$$

where the width as a function of time is

$$\alpha(t) = \frac{\omega_{\mathrm{CME}}}{2} \sin\left[\frac{\pi}{2}(t - t_{\mathrm{onset}})/t_{\mathrm{half}}\right],$$

where $t_{\mathrm{half}} = R_{gb} \tan(\omega_{\mathrm{CME}}/2)/v_{\mathrm{CME}}$, θ_{CME}, and ϕ_{CME} were the co-latitude and longitude of the propagation direction of the CME center, ω_{CME} was the angular width of the CME, and v_{CME} was the velocity of the CME. These parameters were provided by the cone model. For simplicity, the density and temperature of the modeled CMEs were taken to be density $\rho_{\mathrm{CME}} = 10^{-18} kgm^{-3}$ and

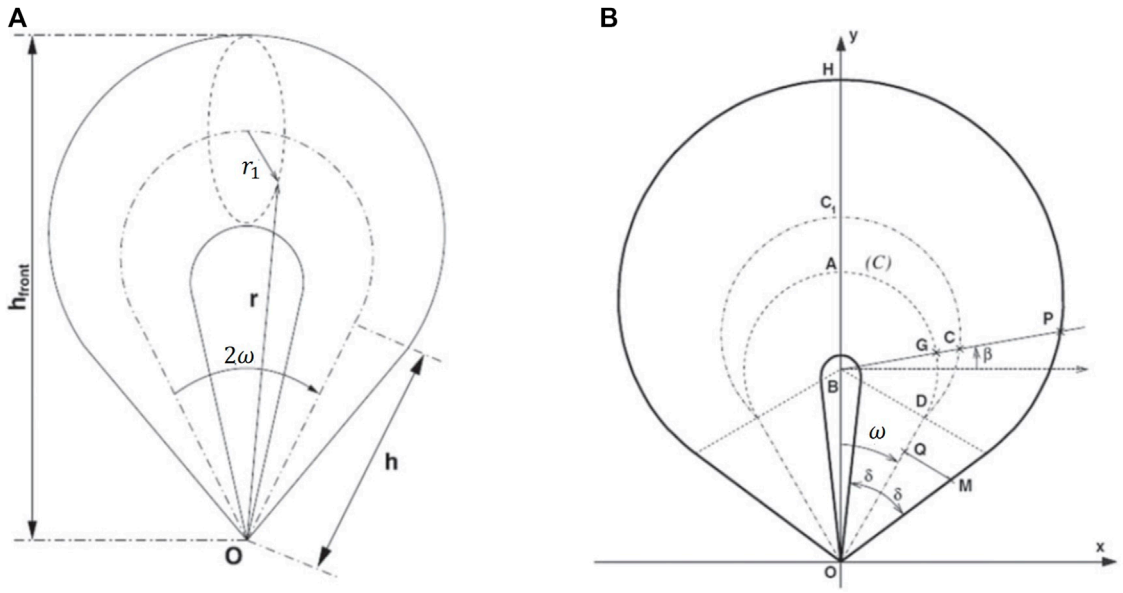


FIGURE 9
(A) Face-on schematic of the graduated cylindrical shell model (Thernisien et al., 2009), **(B)** Schematic of the detailed geometric parameter (Thernisien, 2011).

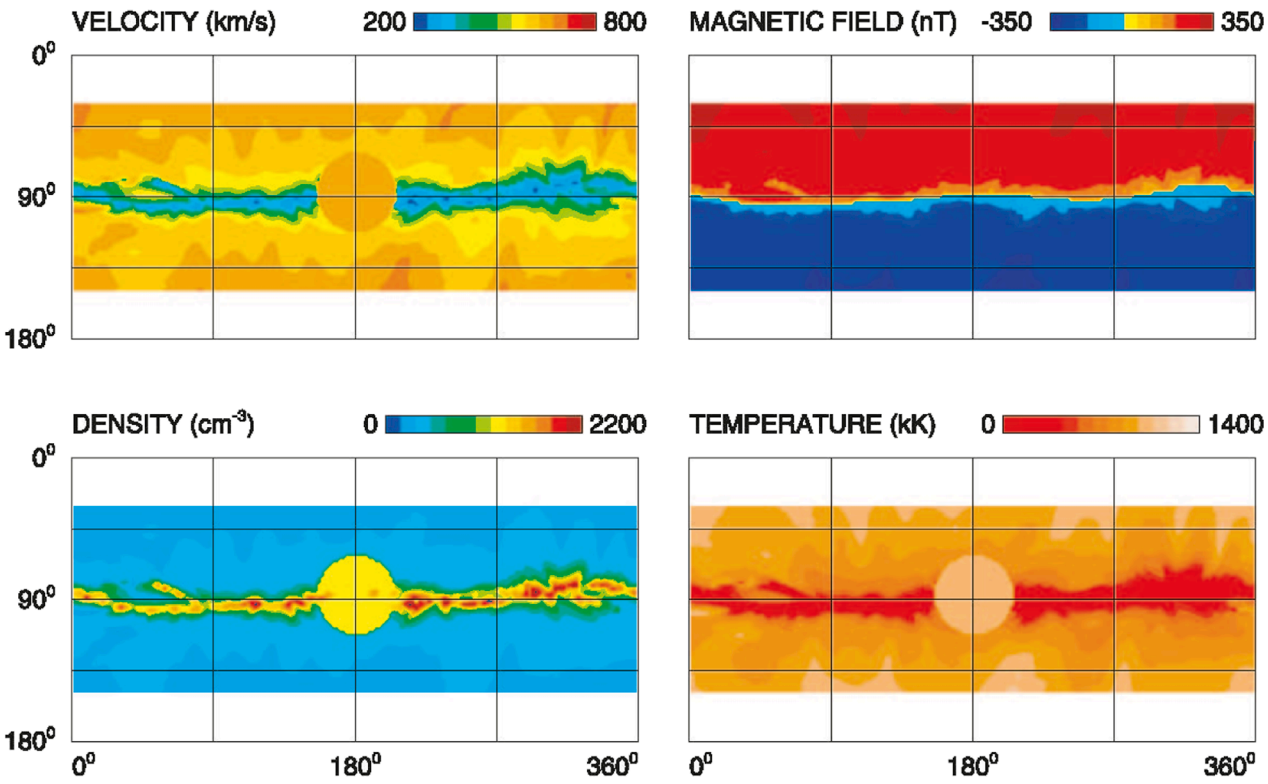


FIGURE 10
Distribution of solar wind parameters with the introduction of the input pulse on 12 May 1997 at 1800 UT (Odstrcil, 2005).

temperature $T_{CME} = 0.8\ MK$, respectively. Scolini et al. (2018) used the EUHFORIA heliospheric model to test the effect of different CME shapes on simulation outputs, such as CME radius, the CME cross-section, and velocity initialization at the inner boundary. They showed that these geometrical parameters particularly affect predictions at locations hit by the CME, and the definition of the CME radius was the parameter having the greatest impact on simulation outputs. Verbeke et al. (2022) presented a new inner heliospheric model, ICARUS, for the simulation of the propagation and evolution of superposed CMEs. The cone CME model described in Scolini et al. (2018) was employed. For both scientific and forecasting purposes, the authors showed that the properties of radial grid stretching and solution adaptive mesh refinement were useful to save computational time.

3.2 Numerical simulation of coronal mass ejection with the flux-rope model

The geoeffectiveness of a CME is largely due to its magnetic field. If the CME contains a positive Bz component, it will be favorable for magnetic reconnection with Earth's magnetosphere and results in a strong geomagnetic storms. The model described so far has not yet predicted the magnetic field structure of CMEs arriving at Earth. In order to overcome the limitations of the cone model, recent work has focused on modeling the CME using more realistic flux-rope models, such as spheromaks or toroidal-like structures.

The SUSANOO-CME model, developed by Shiota and Kataoka (2016), was used to simulate the interplanetary propagation of multiple CMEs with an internal magnetic flux-rope structure. The CME passed through the inner boundary at 30 R_s as the time evolution of the MHD variables. They adopted a spheromak-type magnetic flux rope, and radial compression was adopted for the flux rope to make the CME pass quickly. First, the CME model was defined as a simple linear force-free spheromak model near the Sun.

$$\tilde{B}_{\tilde{r}}\left(\tilde{r},\tilde{\theta},\tilde{\phi}\right) = \left(2B_0/\alpha\tilde{r}\right) j_1\left(\alpha\tilde{r}\right)\cos\tilde{\theta},$$

$$\tilde{B}_{\tilde{\theta}}\left(\tilde{r},\tilde{\theta},\tilde{\phi}\right) = -B_0\left[\frac{j_1\left(\alpha\tilde{r}\right)}{\alpha\tilde{r}} + j_1'\left(\alpha\tilde{r}\right)\right]\sin\tilde{\theta},$$

$$\tilde{B}_{\tilde{\phi}}\left(\tilde{r},\tilde{\theta},\tilde{\phi}\right) = H\cdot B_0 j_1\left(\alpha\tilde{r}\right)\sin\tilde{\theta},$$

where j_1 and j_0 are the first-order and zero-order spherical Bessel function, respectively. The CME was defined in local spherical coordinates $(\tilde{r},\tilde{\theta},\tilde{\phi})$, the center of the CME was chosen to be the origin. Then, the CME model was shifted outward in the radial direction from the Sun to the inner boundary. The spherical cut R'_{cut} of the shifted magnetic configuration moved across the computational boundary as

$$R'_{cut}(t) = R'_{cut}(t_0) - v_{CME}(t - t_0),$$

where $R'_{cut}(t_0) = 30R_s$ defines the nose of the deformed spheromak. $t_0 = t_{onset} + \frac{(30R_s - r_{mc})}{v_{CME}}$ is the start time when the CME appeared at the

TABLE 2 Initial parameters of the CME (Liu et al., 2019).

$\rho_{max}(kgm^{-3})$	$v_{max}(km/s)$	$T_{max}(K)$	$H_m(Mx^2)$	$h(R_s)$	$\omega(°)$	$\delta(°)$
1.15×10^{-18}	1,200	6×10^6	1×10^{42}	25	20	20

inner boundary. Finally, at the inner boundary, the magnetic field is given as

$$B'\left(30,\tilde{\theta},\tilde{\phi}\right)=B'\left(R'_{\rm cut}(t),\tilde{\theta},\tilde{\phi}\right)=\tilde{B}\left(R'_{\rm cut}(t)-a,\tilde{\theta},\tilde{\phi}\right),$$

where $a=30R_s-r_0-r_{\rm mc}$ is the radial shift. The CME radius was $r_{\rm mc}=1R_s\cdot\sin\omega/2$, and $r_0=1R_s+r_{\rm mc}$ was the distance between the central position of the spheromak and the heliospheric center. The model had 10 free parameters. $t_{\rm onset}$ was the onset time and v_{CME} was the propagation speed, which can be found from the LASCO CME catalog; λ_s and ϕ_s were heliographic latitude and longitude, respectively, which came from the flare list in National Geophysical Data Center (NGDC, https://www.ngdc.noaa.gov/stp/spaceweather.html); τ was the tilt angle of the spheromak (±90° with Hale–Nicholson law), χ was the inclination angle of the spheromak, H was the chirality of helicity in the spheromak (1, set −1 if opposite to the Bothmer–Schwenn rule), $\Phi_{\rm mag}$ was the magnetic flux within CME, and $w=60°$ was the angular width of CME. The density inside the CME was in proportion to the magnetic field strength. The CME pressure was assumed to be a constant, which was defined as $\rho(v_{\rm CME}-v_{\rm SW})^2/2$. Because of the constant pressure and high density, the temperature was the lowest at the center of the flux rope.

Liu et al. (2019) used a three-dimensional flux-rope CME initialization model based on the GCS model to simulate the propagation and deflection of a CME from 21.5 R_s to 244 R_s. The Lundquist force-free CME model in a cylindrical geometry (r,ϕ,z) was used.

$$B_r=0,$$

$$B_\phi=HB_0j_1(\alpha\tilde{r}),$$

$$B_z=B_0j_1(\alpha\tilde{r}).$$

The axial component of the magnetic field was assumed to be zero at the edge of the flux rope; thus, they could get

$$\alpha R=2.405,$$

where R stands for the radius of the cylindrical shell, which was obtained from the GCS model. The maximum of magnetic field B_0 is defined as

$$B_0=\sqrt{\frac{2.405H_m}{4\pi LRj}},$$

with $j=\int_0^R j_1^2(\alpha\tilde{r})dr$, H_m is the magnetic helicity of the Lundquist flux rope and $L=2\omega(H_{\rm front}-R_{\rm front})$ is the flux-rope length, and $H_{\rm front}$ and $R_{\rm front}$ are the maximum height and maximum radius of the front, respectively, that were obtained from the GCS model. The initial density, velocity, and temperature at point P in the CME are given as follows:

$$\rho_{\rm CME}=\rho_{\rm max}\left(1-\frac{d^2}{(2r)^2}\right),$$

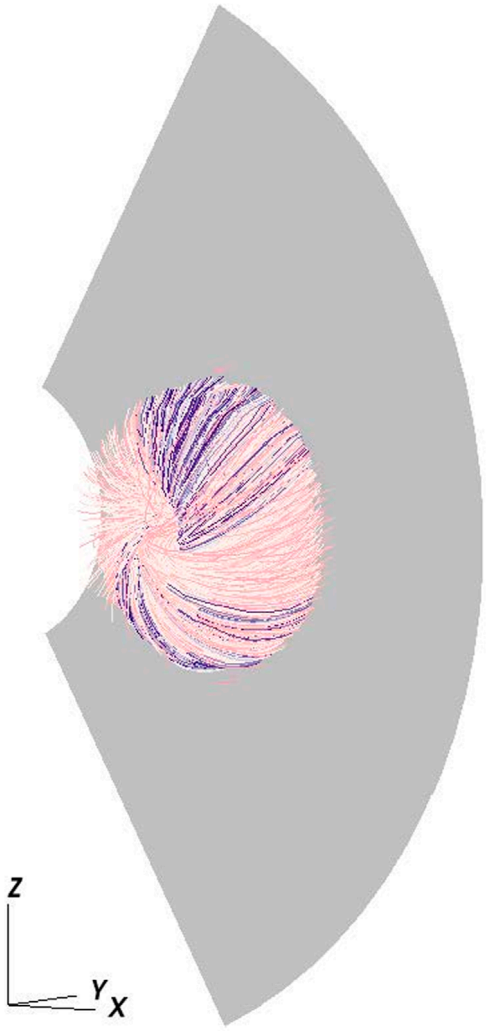


FIGURE 11
Magnetic field lines depicting the structure of the LFFS CME model with a 90.0° tilt (Verbeke et al., 2019).

$$v_{\rm CME}=v_{\rm max}\left(1-\frac{d^2}{(2r)^2}\right),$$

$$T_{\rm CME}=T_{\rm max}\left(1-\frac{d^2}{(2r)^2}\right).$$

Within the GCS model, r represents the radius of the cross section and d stands for distance from point P to the center point at the front part, while defined as the distance from point P to the inner side of the cone in the cross-section plane. **Table 2** shows the initial CME parameters (Liu et al., 2019).

Shen et al. (2021) made a numerical research on how the initial CME parameters can affect simulation results. They found that the initial density and geometric size of the CME both had an effect on its arrival time at Earth. The initial magnetic field had a large effect on the CME's geomagnetic effect. All of these results confirmed that the initial geometric and physical parameters had an important role on space-weather research and forecasts.

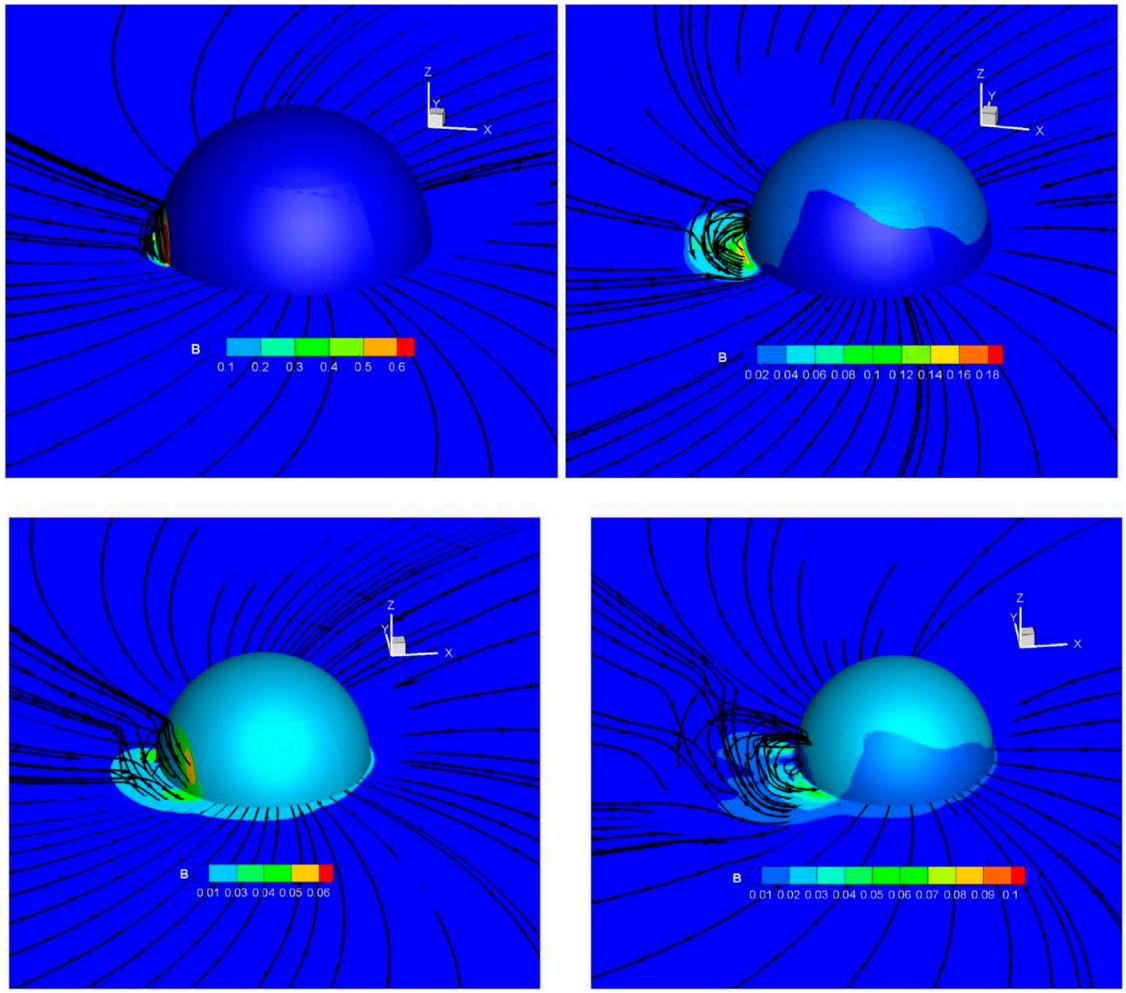


FIGURE 12
Three-dimensional view of the magnetic field for CASE-1 and CASE-2 at different times (Zhang et al., 2021).

The EUHFORIA CME model in Pomoell and Poedts (2018) and Scolini et al. (2018) did not contain an intrinsic magnetic field and was injected at the inner radial boundary with the enhancement of velocity, density, and pressure. Verbeke et al. (2019) introduced a linear force-free spheromak (LFFS) solution to the EUHFORIA CME model. The CME was considered to be a sphere with a radius $r_{mc} = 0.1au \cdot \sin\omega/2$. The CME latitude and longitude $(\theta_{\mathrm{CME}}, \phi_{\mathrm{CME}})$, angular width ω, and speed v_{CME} were the same as in Pomoell and Poedts (2018). Located at $(0.1au - r_{mc}, \theta_{\mathrm{CME}}, \phi_{\mathrm{CME}})$, the CME moved through the 0.1 au boundary with speed v_{CME}. A point at $(0.1au, \theta, \phi)$ satisfied the following formula was inside the CME

$$(x_{\mathrm{CME}} - x_{\mathrm{bound}})^2 + (y_{\mathrm{CME}} - y_{\mathrm{bound}})^2 + (z_{\mathrm{CME}} - z_{\mathrm{bound}})^2 \le r_{mc}, \quad (20)$$

where $(x_{\mathrm{CME}}; y_{\mathrm{CME}}; z_{\mathrm{CME}})$ and $(x_{\mathrm{bound}}; y_{\mathrm{bound}}; z_{\mathrm{bound}})$ were the coordinates of the CME center and the considered point on the boundary in Cartesian HEEQ coordinates, respectively. The magnetic field was defined in a local spherical coordinate system $(\tilde{r}, \tilde{\theta}, \tilde{\phi})$ as in Shiota and Kataoka (2016), with the center of the spheromak as the origin. As in Pomoell and Poedts (2018), density $\rho_{\mathrm{CME}} = 10^{-18} kgm^{-3}$ and temperature $T_{\mathrm{CME}} = 0.8\ MK$. The initial speed required for the spheromak CME model should reflect only the radial speed.

$$v_{\mathrm{rad}} = v_{\mathrm{cme}} \cdot \left[1 - \frac{1}{1 + \frac{1}{2}(1 + \cot w)}\right]. \quad (21)$$

The remaining magnetic input parameters required in the spheromak CME model were the flux-rope handedness, tilt, and the toroidal magnetic flux at 0.1 au. The total toroidal flux Φ was used to define the magnetic field strength B_0.

$$\Phi = \iint \tilde{B}_{\tilde{\phi}} \tilde{r} d\tilde{r} d\tilde{\theta} = \frac{2B_0}{\alpha^2} \cdot \left[-\sin(\alpha r_0) + \int_0^{\alpha r_0} \frac{\sin t}{t} dt\right]. \quad (22)$$

In that paper, a +1 handedness with a 90.0° tilt angle were selected to obtain a negative Bz magnetic field when the CME was arriving at Earth. Lastly, the total toroidal flux was set to be 10^{14} Wb to produce the best results they performed so far. **Figure 11** shows the magnetic field lines depicting the structure of the LFFS CME model with a 90.0° tilt (Verbeke et al., 2019).

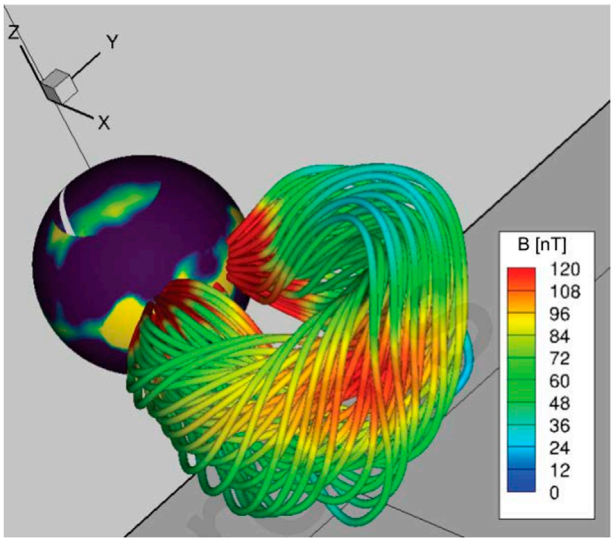


FIGURE 13
FRi3D flux rope emerging out of the EUHFORIA's inner heliospheric boundary at 0.1 AU (Maharana et al., 2022).

Zhang et al. (2021) performed a numerical study of two injection methods for CME in the inner heliosphere, where CASE-1 CME was introduced into the inner boundary with a radial compression (Shiota and Kataoka, 2016) and CASE-2 CME was introduced into the inner boundary without a radial compression (Scolini et al., 2019; Scolini et al., 2020; Verbeke et al., 2019). The CME angular width was obtained by using the sky-plane-projected speed range (Gopalswamy et al., 2010). The CME mass density and temperature were set to be homogeneous as in Pomoell and Poedts (2018); Verbeke et al. (2019). The CME position, velocity, and magnetic input parameters were estimated from image modeling and geometric triangulation analysis following Liu et al. (2010). They found that CASE-2 CME overestimated the radial extension at 1 au and the modeled magnetic fields at 1 au were lower compared to CASE-1 CME. **Figure 12** shows the three-dimensional view of the magnetic field for CASE-1 and CASE-2 at different times. The top row shows the three-dimensional view of the magnetic field at 0.4 h (left), 2.0 h (right) after the addition of the CME into the ambient solar wind for CASE-1, while the bottom row shows the magnetic field at 5 h (left), 15 h (right) after the addition of the CME into the ambient solar wind for CASE-2. Solid black lines display the magnetic field lines and the color code represents the magnetic field B in units of 10^5 T. The blue sphere represents the 20 R_s (Zhang et al., 2021).

Cone CMEs in Pomoell and Poedts (2018) were initialized using a set of seven input parameters during the CME insertion at the inner boundary of the heliospheric model. For example, the CME insertion time, speed v_{CME}, source $(\theta_{\mathrm{CME}}, \phi_{\mathrm{CME}})$, and half angular width $\omega/2$ at 0.1 au were usually derived from cone model. In addition, the CME mass density and temperature were set to be homogeneous with density $\rho_{\mathrm{CME}} = 10^{-18} kgm^{-3}$ and temperature $T_{\mathrm{CME}} = 0.8\ MK$. When simulated CMEs use the spheromak model as in Verbeke et al. (2019), three additional input parameters were needed: the helicity sign (chirality), the tilt, and the toroidal magnetic flux at 0.1 au. Since the spheromak CME in Verbeke et al. (2019) was initialized from observations partially, Scolini et al. (2019, 2020) aimed to constrain all the CME input parameters from remote-sensing observations. The CME geometric and kinematic parameters were derived using the GCS model: the CME direction, the height of the CME apex h_{front}, the tilt angle around the axis of symmetry γ (with respect to the solar equator), the half angle between the legs ω, and the half angle of the cone δ related to the "aspect ratio" κ by the relation $\kappa = \sin\delta$. The 3D speed at the CME apex was derived from the GCS model with a sequence of images.

$$v_{\mathrm{CME}} = \frac{dh_{\mathrm{front}}}{dt}.$$

Then, the radial and expansion contribution were obtained as follows:

$$v_{\mathrm{rad}} = \frac{1}{1+\kappa}\frac{dh_{\mathrm{front}}}{dt},$$

$$v_{\mathrm{exp}} = \frac{\kappa}{1+\kappa}\frac{dh_{\mathrm{front}}}{dt}.$$

The flux-rope handedness was determined from pre-eruption EUV sigmoids observations. The flux-rope tilt angle/orientation was inferred from the orientation of the source region polarity inversion line (PIL) and/or from the orientation of the post-eruption arcades (PEAs). The flux-rope toroidal magnetic flux was based on the reconnected flux, which was computed by the FRED method described by Gopalswamy et al. (2017). Using the PEA area, the FRED method computed the total (unsigned) magnetic flux from line-of-sight magnetic field data. Then, the reconnected flux ϕ_{RC} could be obtained as follows:

$$\phi_{\mathrm{RC}} = \frac{1}{2}\int_{\mathrm{PEA}} |B_{\mathrm{los}}| dA = \frac{1}{2}|B_{\mathrm{los}}|_{\mathrm{tot}} A_{\mathrm{PEA}}.$$

The axial field strength B_0 was estimated with the following formula:

$$B_0 = \frac{\alpha^3}{2\pi}\frac{\phi_p(r_*)\, r_*}{(\sin(\alpha r_*) - \alpha r_* \cos(\alpha r_*))},$$

where r_* is the distance from the center of the spheromak, on the plane $\theta = \frac{\pi}{2}$, where the magnetic field becomes completely axial. $\phi_p(r_*)$ was the poloidal magnetic flux that is transferred from the reconnected flux.

Maharana et al. (2022) introduced a fully analytical "Flux Rope in 3D" (FRi3D) CME model to EUHFORIA (Pomoell and Poedts, 2018) to improve the modeling of CME flank encounters. The geometrical parameters for the FRi3D flux rope contained the half-width φ_{hw} in the azimuthal direction, the angular half-height φ_{hh} in the polar direction, the toroidal height R_t defined the heliocentric distance to the apex of the CME axis, a coefficient n defined the deformation at the CME front, pancaking φ_p meant the deformation due to radial expansion, and skew φ_s defined the deformation due to solar rotation. The geometrical parameters were obtained by using the forward modeling tool of FRi3D to remote multi-viewpoint observations. The Lundquist model in cylindrical geometry was used to get the magnetic field configuration of the CME. The magnetic field parameters, such as the tilt, the magnetic flux, the twist, the chirality, and the polarity, were obtained

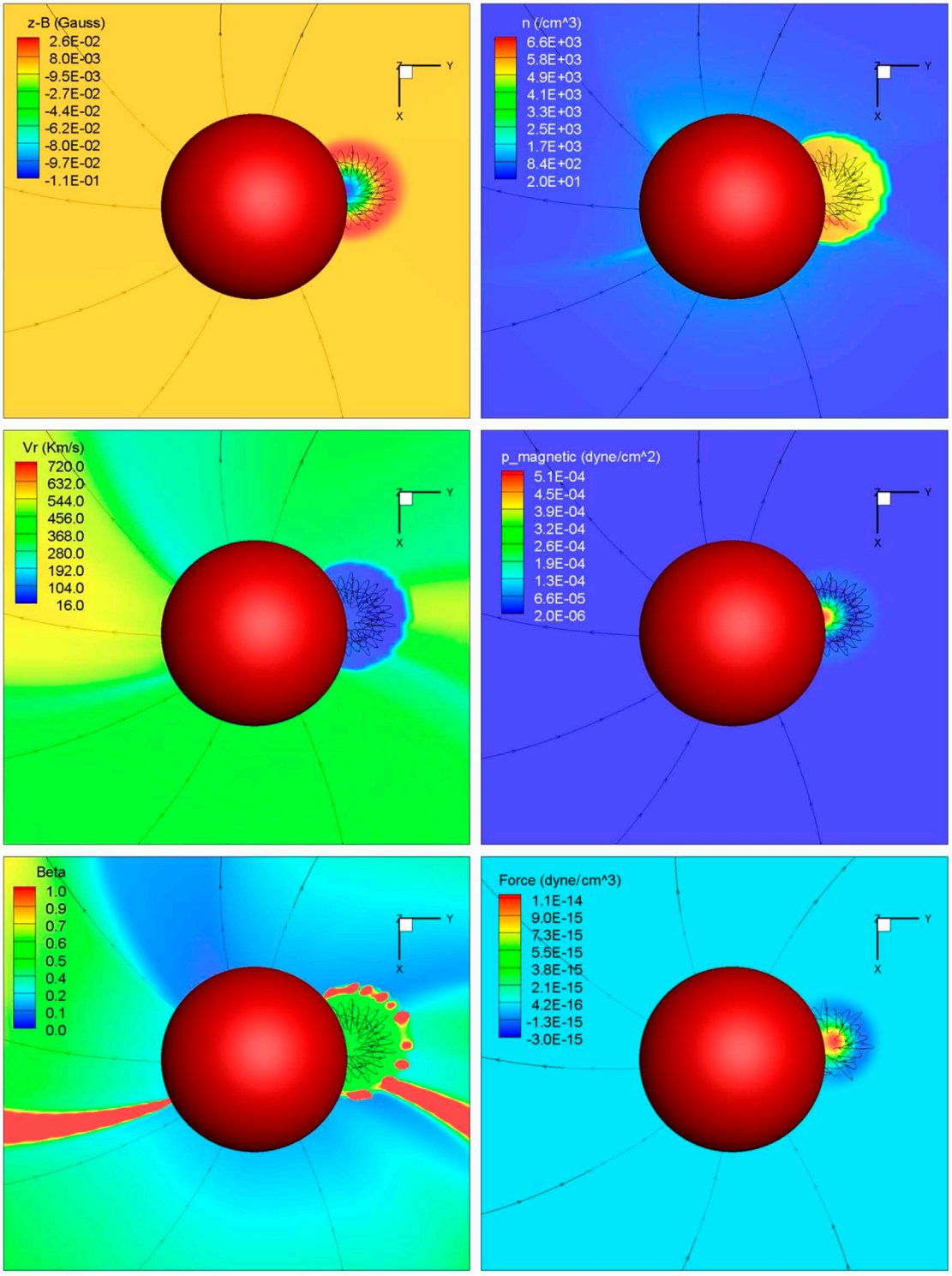


FIGURE 14
Magnetic field line configuration of the modified spheromak inserted at the inner boundary R = 0.1 AU, shown with the red sphere (Singh et al., 2020).

by fitting the FRi3D CME model to WIND spacecraft at 1 au (Isavnin, 2016). The FRi3D CME model was also filled with uniform density plasma with density $\rho_{CME} = 10^{-17} kgm^{-3}$ and temperature $T_{CME} = 0.8\ MK$. **Figure 13** shows a FRi3D flux rope emerging out of EUHFORIA's inner heliospheric boundary at 0.1 au. The color of the field lines is based on the magnetic field strength. The field lines are twisted and deformed as per a particular CME geometry and have maximum strength close to the axis which reduces outward (typical Lundquist model characteristics) (Maharana et al., 2022).

TABLE 3 Various methods to set the CME parameters at the lower boundary.

CME parameters at the lower boundary of the inner heliosphere models							
—	Density	Temperature	Velocity	Location	Angular width	Magnetic fields	—
Hydrodynamic plasma cloud	Four times larger than the mean values in the fast stream	Equal to the mean values in the fast stream	Cone model	Cone model	Cone model	Null	Odstrcil 2003; Odstrcil et al., 2004; Odstrcil 2005
	Constant	Constant	Cone model	Cone model	Cone model	Null	Pomoell and Poedts 2018; Verbeke et al., 2022
Flux rope	Constant	Derived from constant pressure and density	LASCO	NGDC	60°	Some empirical constant	Shiota and Kataoka 2016
	Constant	Constant	GCS model	GCS model	GCS model	GCS model	Liu et al., 2019; Shen et al., 2021
	Constant	Constant	Cone model	Cone model	Cone model	Cone empirical constant	Verbeke et al., 2019
	Constant	Constant	Image modeling	Image modeling	Derived from sky-plane-projected	Image modeling	Zhang et al., 2021
	Constant	Constant	GCS model	GCS model	GCS model	Remote-sensing observations	Scolini et al., 2019, Scolini et al., 2020
	Constant	Constant	Forward modeling	Forward modeling	Forward modeling	By fitting the FRi3D CME model to WIND	Maharana et al., 2022

Singh et al. (2020) applied a modified spheromak model for simulations of the CME in the inner heliosphere. The simulations were carried out using a MultiScale Fluid-Kinetic Simulation Suite (MS-FLUKSS) code. They introduced a modified spheromak flux rope into the solar wind, where the poloidal and toroidal fluxes were set up independently.

$$\vec{b} = \frac{1}{r\sin\theta}\left(\zeta\frac{1}{r}\frac{\partial A}{\partial\theta}\hat{r} - \zeta\frac{\partial A}{\partial r}\hat{\theta} + H\alpha_0 A\hat{\phi}\right),$$

$$A = \frac{4\pi a_1}{\alpha_0^2}\left[\frac{r_0^2}{g(\alpha_0 r)} - r^2\right]\sin^2\theta,$$

$$g(\alpha_0 r) = \frac{\sin(\alpha_0 r)}{\alpha_0 r} - \cos(\alpha_0 r),$$

where $\alpha_0 r_{mc} = 5.763,459$. The origin of the spherical coordinate system was placed at the spheromak center. The flux-rope parameters were introduced into the solar wind as follows:

$$\vec{b}_{final} = \vec{b}_{FR},$$

$$\rho_{final} = \rho_{FR} + \rho_{SW},$$

$$e_{final} = \xi e_{FR} + e_{SW},$$

where ξ was related to plasma β as $\beta \sim (\gamma - 1)(\xi - 1)$ indicated that the magnetic pressure in the flux rope was larger than the pressure in the solar wind. The mass inside the spheromak was a constant with 1.65×10^{15} g. The thermal pressure was assumed to be proportional to the magnetic pressure. The poloidal flux and the toroidal flux were set to be 2×10^{22} Mx and 5×10^{21} Mx, respectively. A positive helicity sign was used. **Figure 14** shows the magnetic field line configuration of the modified spheromak inserted at the inner boundary (Singh et al., 2020).

Table 3 summarizes the various methods to set the CME parameters at lower boundary. Since the ability of a CME to disturb the near-Earth space is largely due to its magnetic fields, the flux-rope model is widely used in recent years because of its internal magnetic fields structure. How to constrain the initial CME parameters by observations is one difficulty. The cone model and GCS model are widely used to estimate the CME kinematic and geometric parameters. The magnetic field parameters, such as handedness, tilt, and toroidal magnetic flux, are deduced from remote-sensing observations in Scolini et al. (2019, 2020). With the launch of the PSP and SO into the inner heliosphere, the ICME will be constrained by more observations.

4 Conclusion and discussion

Overall, the three-dimensional MHD numerical model has become an important tool to report interplanetary solar wind and the CME. The MHD heliospheric model is relatively cheap in terms of calculation, and it takes tens of minutes to hours to simulate the CME propagation from the Sun to Earth. The operational space-weather forecast is usually generated by the heliospheric model. The

resulting forecast errors are largely dependent on the uncertainty of the inner boundary conditions, rather than the heliospheric model.

For the solar wind simulation in the inner heliosphere, the solar wind density and temperature at the inner boundary are always derived as a function of the solar wind speed and magnetic field; thus, the speed and magnetic field can significantly affect the accuracy of the simulated results. The speed specified by the WSA formula given in Eq. 2 seems to be the most efficient formula. However, there are several free parameters in the WSA model, whose uncertainties may result in the deviations between the simulated results and observations. Thus, how to make parameter optimization is a major problem. The dynamic time warping (DTW) algorithm described in Bunting and Morgan (2022) can be used to derive the optimized values for the speed in the WSA formula. DTW algorithm is effective to quantify the agreement between two time series and has recently been used in a variety of fields such as economics, biology, and space weather. Here, DTW distance is used to measure the difference between the model and *in situ* solar wind velocities, and a smaller DTW distance means that the modeled data is in good agreement with the *in situ* data. The initial parameters of the WSA formula were adjusted to minimize the DTW distance, and then, we can obtain the optimized parameters in WSA formula. Also, the IPS observation is a good choice to obtain the velocity distribution, and the results show that IPS data have a high success rate in detecting high-speed solar wind and the space-weather prediction can be enhanced by the combination of MHD simulation and the IPS data. For the magnetic field at the inner boundary, the PFSS model is always used. However, the PFSS model underestimated the magnetic flux in the heliosphere. Shen et al. (2018) only kept the polarity of the magnetic field from the PFSS model and used the observational data at 1 au in the immediate past to limit the value of the magnetic field, and the results showed that the modeled magnetic field at 1 au agrees with the *in situ* observations much better. Additionally, the magnetic fields derived from the PFSS model have no thin HCS or Parker spiral in the interplanetary space; thus, the PFSS + SCS model may be a good choice.

For the CME simulation in the inner heliosphere, CMEs are always initialized by flux-rope models, such as spheromaks or toroidal-like structures, and the CMEs are introduced to the inner radial boundary as a time-dependent boundary. Scolini et al. (2019, 2020) suggested that the prediction of the CME arrival time at Earth was found to be highly dependent on the CME model and CME input parameters used. Thus, how to obtain the CME parameters at the inner boundary is important. At present, CME geometric and kinematic parameters, such as CME position, radius, and velocity, are always derived from the GCS model or the cone model. The density and temperature of the CMEs are assumed to be constant. To improve the prediction of the CME internal magnetic field at Earth, it is necessary to constrain the flux-rope magnetic parameters from the remote-sensing observations of the corona. Scolini et al. (2019, 2020) constrained flux-rope magnetic parameters from remote-sensing observations, and the results showed that the current model can improve the predictions of B_z up to 22–40 percentage points compared to a cone model.

At present, the numerical simulation in the inner heliosphere has a large dependence on the empirical model. How to use the observations to constrain the uncertainties in the empirical models is something we must consider in future. Maybe we can use the *in situ* data in the past CR to correct the uncertainties in the empirical models at the current CR, but this can only be applicable to the solar wind. With the multi-spacecraft observations in the future, such as SO and PSP, the detailed measurements of the inner heliosphere will be provided. All of these will provide us with more realistic inner boundary conditions, and we will have a deeper understanding of the triggering and propagation evolution of solar wind and disturbance.

Author contributions

XF designed the research. MZ and HL participated in data analyses and the writing of the manuscript. MX, FS, LY, XZ, YZ, and XL participated in data interpretation and discussion. All authors read and agreed to the published version of the manuscript.

Funding

The work was jointly supported by the National Natural Science Foundation of China (Grant Nos. 42030204 and 41874202), the National Key R&D Program of China (Grant No. 2022YFF0503900), the Shenzhen Natural Science Fund (the Stable Support Plan Program GXWD20220817152453003), and the Specialized Research Fund for State Key Laboratories.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Altschuler, M. D., and Newkirk, G. J. (1969). Magnetic fields and the structure of the solar corona. *Sol. Phys.* 9, 131–149. doi:10.1007/bf00145734

Arge, C. N., Luhmann, J. G., Odstrcil, D., Schrijver, C. J., and Li, Y. (2004). Stream structure and coronal sources of the solar wind during the may 12th, 1997 cme. *J. Atmos. Solar-Terrestrial Phys.* 66, 1295–1309. doi:10.1016/j.jastp.2004.03.018

Arge, C. N., and Pizzo, V. J. (2000). Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates. *J. Geophys. Res.* 105, 10465–10479. doi:10.1029/1999ja000262

Biondo, R., Bemporad, A., Mignone, A., and Reale, F. (2021). Reconstruction of the parker spiral with the reverse *in situ* data and mhd approach-rimap. *J. Space Weather Space Clim.* 11, 7. doi:10.1051/swsc/2020072

Bochsler, P., Balsiger, H., Gloeckler, G., Burgi, A., Fisk, L., Galvin, A., et al. (1995). The solar wind and suprathermal ion composition investigation on the wind spacecraft. *Space Sci. Rev.* 71, 79–124. doi:10.1007/bf00751327

Bunting, K. A., and Morgan, H. (2022). An inner boundary condition for solar wind models based on coronal density. *J. Space Weather Space Clim.* 12, 30. doi:10.1051/swsc/2022026

Chat, G. L., Issautier, K., and Meyer-Vernet, N. (2012). The solar wind energy flux. *Sol. Phys.* 279, 197–205. doi:10.1007/s11207-012-9967-y

Detman, T., Smith, Z., Dryer, M., Fry, C. D., Arge, C. N., and Pizzo, V. (2006). A hybrid heliospheric modeling system: Background solar wind. *J. Geophys. Res.* 111, A07102. doi:10.1029/2005ja011430

Detman, T. R., Intriligator, D. S., Dryer, M., Sun, W., Deehr, C. S., and Intriligator, J. (2011). The influence of pickup protons, from interstellar neutral hydrogen, on the propagation of interplanetary shocks from the halloween 2003 solar events to ACE and Ulysses: A 3-D MHD modeling study. *J. Geophys. Res.* 116. doi:10.1029/2010ja015803

Elliott, H. A., McComas, D. J., Schwadron, N. A., Gosling, J. T., Skoug, R. M., Gloeckler, G., et al. (2005). An improved expected temperature formula for identifying interplanetary coronal mass ejections. *J. Geophys. Res. Space Phys.* 110, A04103. doi:10.1029/2004ja010794

Feng, X. (2020). *Magnetohydrodynamic modeling of the solar corona and heliosphere*. Singapore: Springer. doi:10.1007/978-981-13-9081-4

Feng, X. S., Zhang, M., and Zhou, Y. F. (2014). A new three-dimensional solar wind model in spherical coordinates with a six-component grid. *Astrophysical J. Suppl.* 214, 6. doi:10.1088/0067-0049/214/1/6

Fox, N. J., Velli, M. C., Bale, S. D., Decker, R., Driesman, A., Howard, R. A., et al. (2016). The solar probe plus mission: Humanity's first visit to our star. *Space Sci. Rev.* 204, 7–48. doi:10.1007/s11214-015-0211-6

Gonzi, S., Weinzierl, M., Bocquet, F., Bisi, M. M., Odstrcil, D., Jackson, B. V., et al. (2021). Impact of inner heliospheric boundary conditions on solar wind predictions at Earth. *Space Weather.* 19, 13. doi:10.1029/2020sw002499

Gopalswamy, N., Yashiro, S., Akiyama, S., and Xie, H. (2017). Estimation of reconnection flux using post-eruption arcades and its relevance to magnetic clouds at 1 au. *Sol. Phys.* 292, 65. doi:10.1007/s11207-017-1080-9

Gopalswamy, N., Yashiro, S., Michalek, G., Xie, H., Makela, P., Vourlidas, A., et al. (2010). A catalog of halo coronal mass ejections from soho. *Am. Geophys. Union* 5, 7–16.

Hayashi, K. (2012). An mhd simulation model of time-dependent co-rotating solar wind. *J. Geophys. Res.* 117, A08105. doi:10.1029/2011ja017490

Hayashi, K. (2003). Mhd tomography using interplanetary scintillation measurement. *J. Geophys. Res. Space Phys.* 108, 1102. doi:10.1029/2002ja009567

Howard, R., Moses, J., Vourlidas, A., Newmark, J., Socker, D., Plunkett, S., et al. (2008). Sun Earth connection coronal and heliospheric investigation (secchi). *Space Sci. Rev.* 136, 67–115. doi:10.1007/s11214-008-9341-4

Isavnin, A. (2016). Fried: A novel three-dimensional model of coronal mass ejections. *Astrophysical J.* 833, 267. doi:10.3847/1538-4357/833/2/267

Kaiser, M., Kucera, T., Davila, J., and St. Cyr, O. C. (2008). The STEREO mission: An introduction. *Space Sci. Rev.* 136, 5–16. doi:10.1007/s11214-007-9277-0

Kilpua, E. K. J., Lugaz, N., Mays, M. L., and Temmer, M. (2019). Forecasting the structure and orientation of earthbound coronal mass ejections. *Space Weather.* 17, 498–526. doi:10.1029/2018sw001944

Kumar, S., Paul, A., and Vaidya, B. (2020). A comparison study of extrapolation models and empirical relations in forecasting solar wind. *Front. Astronomy Space Sci.* 7, 572084. doi:10.3389/fspas.2020.572084

Li, H., Feng, X., Zuo, P., and Wei, F. (2020). Simulation of the interplanetary B_z using a data-driven heliospheric solar wind model. *Astrophysical J.* 900, 76. doi:10.3847/1538-4357/aba61f

Liu, Y., Shen, F., and Yang, Y. (2019). Numerical simulation on the propagation and deflectio n of fast coronal mass ejections (cmes) interacting with a corotating interaction region in interplanetary space. *Astrophysical J.* 887, 150. doi:10.3847/1538-4357/ab543e

Liu, Y., Thernisien, A., Luhmann, J., Vourlidas, A., Davies, J., Lin, R., et al. (2010). Reconstructing coronal mass ejections with coordinated imaging and *in situ* observations: Global structure, kinematics, and implications for space weather forecasting. *Astrophysical J.* 722, 1762–1777. doi:10.1088/0004-637x/722/2/1762

Lopez, R. E., and Freeman, J. W. (1986). Solar wind proton temperature-velocity relationship. *J. Geophys. Res. Space Phys.* 91, 1701–1705. doi:10.1029/ja091ia02p01701

Mackay, D. H., and Yeates, A. R. (1999). The suns global photospheric and coronal magnetic fields: Observations and models. *Living Rev. Sol. Phys.* 9, 1–50. doi:10.12942/lrsp-2012-6

Maharana, A., Isavnin, A., Scolini, C., Wijsen, N., Rodriguez, L., Mierla, M., et al. (2022). Implementation and validation of the FRi3d flux rope model in EUHFORIA. *Adv. Space Res.* 70, 1641–1662. doi:10.1016/j.asr.2022.05.056

Mayank, P., Vaidya, B., and Chakrabarty, D. (2022). SWASTi-SW: Space weather adaptive simulation framework for solar wind and its relevance to the aditya-l1 mission. *Astrophysical J. Suppl. Ser.* 262, 23. doi:10.3847/1538-4365/ac8551

Mcgregor, S. L., Hughes, W. J., Arge, C. N., Owens, M. J., and Odstrcil, D. (2011). The distribution of solar wind speeds during solar minimum: Calibration for numerical solar wind modeling constraints on the source of the slow solar wind. *J. Geophys. Res. Space Phys.* 116, A03101. doi:10.1029/2010ja015881

Merkin, V. G., Lyon, J. G., Lario, D., Arge, C. N., and Henney, C. J. (2016). Time-dependent magnetohydrodynamic simulations of the inner heliosphere. *J. Geophys. Res. Space Phys.* 121, 2866–2890. doi:10.1002/2015ja022200

Merkin, V. G., Lyon, J. G., Mcgregor, S. L., and Pahud, D. M. (2011). Disruption of a heliospheric current sheet fold. *Geophys. Res. Lett.* 38, L14107. doi:10.1029/2011gl047822

Michaek, G. (2006). An asymmetric cone model for halo coronal mass ejections. *Sol. Phys.* 237, 101–118. doi:10.1007/s11207-006-0075-8

Mignone, A., Bodo, G., Massaglia, S., Matsakos, T., Tesileanu, O., Zanni, C., et al. (2007). Pluto: A numerical code for computational astrophysics. *Astrophysical J. Suppl. Ser.* 170, 228–242. doi:10.1086/513316

Müller, D., St Cyr, O. C., Zouganelis, I., Gilbert, H. R., Marsden, R., Nieves-Chinchilla, T., et al. (2020). The Solar Orbiter mission. Science overview. *Astronomy Astrophysics* 642, A1. doi:10.1051/0004-6361/202038467

Nakamizo, A., Tanaka, T., Kubo, Y., Kamei, S., Shimazu, H., and Shinagawa, H. (2009). Development of the 3-D MHD model of the solar corona-solar wind combining system. *J. Geophys. Res.* 114, A07109. doi:10.1029/2008ja013844

Narechania, N. M., Nikolic, L., Freret, L., Sterck, H. D., and Groth, C. (2021). An integrated data-driven solar wind-cme numerical framework for space weather forecasting. *J. Space Weather Space Clim.* 11, 8. doi:10.1051/swsc/2020068

Odstrcil, D., And, P. R., and Zhao, X. P. (2004). Numerical simulation of the 12 may 1997 interplanetary cme event. *J. Geophys. Res. Space Phys.* 109, A02116. doi:10.1029/2003ja010135

Odstrcil, D. (2005). Propagation of the 12 may 1997 interplanetary coronal mass ejection in evolving solar wind structures. *J. Geophys. Res.* 110, A02106. doi:10.1029/2004ja010745

Odstrcil, D. (2003). Modeling 3-D solar wind structure. *Adv. Space Res.* 32, 497–506. doi:10.1016/s0273-1177(03)00332-6

Owens, M., Lang, M., Barnard, L., Riley, P., Gonzi, S., Scott, C. J., et al. (2020). A computationally efficient, time-dependent model of the solar wind for use as a surrogate to three-dimensional numerical magnetohydrodynamic simulations. *Sol. Phys.* 295, 43. doi:10.1007/s11207-020-01605-3

Owens, M. J., Spence, H. E., McGregor, S., Hughes, W. J., Quinn, J. M., Arge, C. N., et al. (2008). Metrics for solar wind prediction models: Comparison of empirical, hybrid, and physics-based schemes with 8 years of L1 observations. *Space Weather.* 6, S08001. doi:10.1029/2007sw000380

Pahud, D. M., Merkin, V. G., Arge, C. N., Hughes, W. J., and Mcgregor, S. M. (2012). An mhd simulation of the inner heliosphere during Carrington rotations 2060 and 2068: Comparison with messenger and ace spacecraft observations. *J. Atmos. Solar-Terrestrial Phys.* 83, 32–38. doi:10.1016/j.jastp.2012.02.012

Pizzo, V. J. (1982). A three-dimensional model of corotating streams in the solar wind: 3. Magnetohydrodynamic streams. *J. Geophys. Res. Space Phys.* 87, 4374. doi:10.1029/ja087ia06p04374

Pomoell, J., and Poedts, S. (2018). Euhforia: European heliospheric forecasting information asset. *J. Space Weather Space Clim.* 8, A35–A204. doi:10.1051/swsc/2018020

Richardson, I. G., and Cane, H. V. (1995). Regions of abnormally low proton temperature in the solar wind (1965–1991) and their association with ejecta. *J. Geophys. Res. Space Phys.* 100, 23397. doi:10.1029/95ja02684

Riley, P., Linker, J. A., Lionello, R., and Mikic, Z. (2012). Corotating interaction regions during the recent solar minimum: The power and limitations of global mhd modeling. *J. Atmos. Solar-Terrestrial Phys.* 83, 1–10. doi:10.1016/j.jastp.2011.12.013

Riley, P., and Lionello, R. (2011). Mapping solar wind streams from the sun to 1 au: A comparison of techniques. *Sol. Phys.* 270, 575–592. doi:10.1007/s11207-011-9766-x

Schatten, K. H. (1971). Current sheet model for the solar corona. *Cosm. Electrodyn.* 2.

Schatten, K. H., Wilcox, J. M., and Ness, N. F. (1969). A model of interplanetary and coronal magnetic fields. *Sol. Phys.* 6, 442–455. doi:10.1007/bf00146478

Scolini, C., Chane, E., Temmer, M., Kilpua, E., Dissauer, K., Veronig, A. M., et al. (2020). Cme-cme interactions as sources of cme geoeffectiveness: The formation of the complex ejecta and intense geomagnetic storm in 2017 early september. *Astrophysical J. Suppl. Ser.* 247, 21. doi:10.3847/1538-4365/ab6216

Scolini, C., Rodriguez, L., Mierla, M., Pomoell, J., and Poedts, S. (2019). Observation-based modelling of magnetised coronal mass ejections with euhforia. *Astronomy Astrophysics* 626, A122. doi:10.1051/0004-6361/201935053

Scolini, C., Verbeke, C., Poedts, S., and Chane, E. (2018). Effect of the initial shape of coronal mass ejections on 3-D MHD simulations and geoeffectiveness predictions. *Space Weather-the Int. J. Res. Appl.* 16, 754–771. doi:10.1029/2018sw001806

Shen, F., Liu, Y., and Yang, Y. (2021). Numerical research on the effect of the initial parameters of a CME flux-rope model on simulation results. *Astrophysical J. Suppl. Ser.* 253, 12. doi:10.3847/1538-4365/abd4d2

Shen, F., Yang, Z., Zhang, J., Wei, W., and Feng, X. (2018). Three-dimensional MHD simulation of solar wind using a new boundary treatment: Comparison with *in situ* data at Earth. *Astrophysical J.* 866, 18. doi:10.3847/1538-4357/aad806

Shiota, D., and Kataoka, R. (2016). Magnetohydrodynamic simulation of interplanetary propagation of multiple coronal mass ejections with internal magnetic flux rope (susanoo-cme). *Space Weather-the Int. J. Res. Appl.* 14, 56–75. doi:10.1002/2015sw001308

Shiota, D., Kataoka, R., Miyoshi, Y., Hara, T., Tao, C., Masunaga, K., et al. (2014). Inner heliosphere mhd modeling system applicable to space weather forecasting for the other planets. *Space Weather.* 12, 187–204. doi:10.1002/2013sw000989

Singh, T., Kim, T. K., Pogorelov, N. V., and Arge, C. N. (2020). Application of a modified spheromak model to simulations of coronal mass ejection in the inner heliosphere. *Space Weather.* 18. doi:10.1029/2019sw002405

Stone, E. C., Frandsen, A. M., Mewaldt, R. A., Christian, E. R., Margolies, D., Ormes, J. F., et al. (1998). The advanced composition explorer. *Space Sci. Rev.* 86, 1–22. doi:10.1023/a:1005082526237

Thernisien, A., Vourlidas, A., and Howard, R. A. (2009). Forward modeling of coronal mass ejections using stereo/secchi data. *Sol. Phys.* 256, 111–130. doi:10.1007/s11207-009-9346-5

Thernisien, A. F. R., Howard, R. A., and Vourlidas, A. (2006). Modeling of flux rope coronal mass ejections. *Astrophysical J.* 652, 763–773. doi:10.1086/508254

Thernisien, A. F. R. (2011). Implementation of the graduated cylindrical shell model for the three-dimensional reconstruction of coronal mass ejections. *Astrophysical J. Suppl. Ser.* 194, 33. doi:10.1088/0067-0049/194/2/33

Usmanov, A. V. (1993). A global numerical 3-d mhd model of the solar wind. *Sol. Phys.* 146, 377–396. doi:10.1007/bf00662021

Verbanac, G., Vrsnak, B., Zivkovic, S., Hojsak, T., Veronig, A., and Temmer, M. (2011). Solar wind high-speed streams and related geomagnetic activity in the declining phase of solar cycle 23. *Astronomy Astrophysics* 533, A49. doi:10.1051/0004-6361/201116615

Verbeke, C., Baratashvili, T., and Poedts, S. (2022). Icarus, a new inner heliospheric model with a flexible grid. *Astronomy Astrophysics* 662, A50. doi:10.1051/0004-6361/202141981

Verbeke, C., Pomoell, J., and Poedts, S. (2019). The evolution of coronal mass ejections in the inner heliosphere: Implementing the spheromak model with euhforia. *Astronomy Astrophysics* 627, A111. doi:10.1051/0004-6361/201834702

Wang, Y. X., Guo, X. C., Wang, C., Florinski, V., Shen, F., Li, H., et al. (2020). Mhd modeling of the background solar wind in the inner heliosphere from 0.1 to 5.5 au: Comparison with *in-situ* observations. *Space Weather.* 18, 23. doi:10.1029/2019sw002262

Wiengarten, T., Kleimann, J., Fichtner, H., Cameron, R., Jiang, J., Kissmann, R., et al. (2013). Mhd simulation of the inner-heliospheric magnetic field. *J. Geophys. Res. Space Phys.* 118, 29–44. doi:10.1029/2012ja018089

Wiengarten, T., Kleimann, J., Fichtner, H., Kühl, P., Kopp, A., Heber, B., et al. (2014). Cosmic ray transport in heliospheric magnetic structures. i. modeling background solar wind using the cronos magnetohydrodynamic code. *Astrophysical J.* 788, 80. doi:10.1088/0004-637x/788/1/80

Wu, C. C., Feng, X. S., Wu, S. T., Dryer, M., and Fry, C. D. (2006). Effects of the interaction and evolution of interplanetary shocks on background solar wind speeds. *J. Geophys. Res. Space Phys.* 111, A12104. doi:10.1029/2006ja011615

Xie, H., Ofman, L., and Lawrence, G. (2004). Cone model for halo cmes: Application to space weather forecasting. *J. Geophys. Res.* 109, A03109. doi:10.1029/2003ja010226

Zhang, M., and Feng, X. (2015). Implicit dual-time stepping method for a solar wind model in spherical coordinates. *Comput. Fluids* 115, 115–123. doi:10.1016/j.compfluid.2015.03.020

Zhang, M., Feng, X., Shen, F., and Yang, L. (2021). Numerical study of two injection methods for the 2007 november 15 coronal mass ejection in the inner heliosphere. *Astrophysical J.* 918, 35. doi:10.3847/1538-4357/ac0b3f

Zhang, M., and Xueshang, F. (2016). A comparative study of divergence cleaning methods of magnetic field in the solar coronal numerical simulation. *Front. Astronomy Space Sci.* 3, 6. doi:10.3389/fspas.2016.00006

Zhao, X., and Hoeksema, J. T. (1995). Predicting the heliospheric magnetic field using the current sheet-source surface model. *Adv. Space Res.* 16, 181–184. doi:10.1016/0273-1177(95)00331-8

Zhao, X. P., Plunkett, S. P., and Liu, W. (2002). Determination of geometrical and kinematical properties of halo coronal mass ejections using the cone model. *J. Geophys. Res. Space Phys.* 107, SSH 13-1–SSH 13-19. doi:10.1029/2001ja009143

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 04 May 2023
DOI 10.3389/fspas.2023.1154760

Check for updates

OPEN ACCESS

EDITED BY
Olga V. Khabarova,
Tel Aviv University, Israel

REVIEWED BY
Laxman Adhikari,
University of Alabama in Huntsville, United States
Jungyeon Cho,
Chungnam National University, Republic of Korea

*CORRESPONDENCE
Alex Lazarian,
lazarian@astro.wisc.edu

RECEIVED 31 January 2023
ACCEPTED 17 April 2023
PUBLISHED 04 May 2023

CITATION
Lazarian A, Xu S and Hu Y (2023), Cosmic ray propagation in turbulent magnetic fields.
Front. Astron. Space Sci. 10:1154760.
doi: 10.3389/fspas.2023.1154760

COPYRIGHT
© 2023 Lazarian, Xu and Hu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Cosmic ray propagation in turbulent magnetic fields

Alex Lazarian[1,2]*, Siyao Xu[3] and Yue Hu[1]

[1]Department of Astronomy, University of Wisconsin, Madison, WI, United States, [2]Centro de Investigación en Astronomía, Universidad Bernardo O'Higgins, Santiago, Chile, [3]Institute for Advanced Study, Princeton, NJ, United States

Propagation of cosmic rays (CRs) in turbulent and magnetized astrophysical media is a long-standing problem that requires both understanding of the properties of turbulent magnetic fields and their interaction with energetic particles. This review focuses on selected recent theoretical findings made based on the progress in understanding and simulating magnetohydrodynamic (MHD) turbulence. In particular, we address the problem of perpendicular and parallel propagation of CRs and identify the conditions when the perpendicular propagation is superdiffusive and diffusive. For the parallel diffusion, we discuss the problems of the traditionally used diffusion mechanism arising from pitch angle scattering and the possible solutions provided by the recently identified "mirror diffusion" in the presence of turbulent magnetic mirrors.

KEYWORDS

cosmic rays, interstellar medium, MHD turbulence, diffusion, magnetic fields

1 Introduction

Nature demonstrates unique abilities to accelerate energetic particles. In astrophysical literature, such particles with energies ranging from 10^7 eV to 10^{20} eV are usually termed cosmic rays (CRs). The processes of acceleration include systematic First Order Fermi and stochastic Second Order Fermi acceleration mechanisms (Melrose, 1969), and shocks, magnetic reconnection, and turbulence are involved in the acceleration processes (Longair, 2011). CRs in the Galaxy have energy density and pressure comparable to those arising from interstellar turbulence and magnetic fields (Parker, 1966). Therefore, they are dynamically important.

The observed CR flux and (an) isotropy in their arrival directions depends on both their acceleration and propagation from their sources to the Earth (Ginzburg and Syrovatskii, 1964; Schlickeiser, 2002). This review focuses on selected recent theoretical findings on CR propagation in turbulent magnetic fields, which are ubiquitous in space and astrophysical environments. CR propagation has numerous implications on various processes in the solar wind, heliosphere, interstellar media, and intracluster media (see monograph by Longair (2011) and references therein). It is a crucial part of CR acceleration process (e.g., Jokipii, 1982; Perri and Zimbardo, 2009; Lazarian and Yan, 2014; Demidem et al., 2020; Marcowith et al., 2020; Xu and Lazarian, 2022). The diffusion in the vicinity of CR acceleration sites, e.g., supernova remnants, and pulsar wind nebulae, is important for understanding gamma-ray observations (Di Mauro et al., 2020; Xu, 2021). Other implications include the solar modulation of Galactic CRs, space weather forecasting (Parker, 1965; Jokipii, 1971; Singer et al., 2001), the origin and chemical composition of CRs, ionization in molecular gas and circumstellar discs (e.g., Schlickeiser et al., 2016;

Padovani et al., 2018), Fermi Bubble emission (Anjos and Catalani, 2020), galactic winds (e.g., Ipavich, 1975; Holguin et al., 2019; Hopkins et al., 2020; Quataert et al., 2022), feedback heating in clusters of galaxies (e.g., Guo and Oh, 2008; Brunetti and Jones, 2014), modeling the synchrotron foreground emission for cosmic microwave background (CMB) radiation and redshifted 21 cm radiation (e.g., Cho and Lazarian, 2002b; Cho et al., 2012).

In the propagation process, CRs interact with the pre-existing magnetohydrodynamic (MHD) fluctuations and the magnetic fluctuations created by themselves. The most notable example for the latter is the perturbations created by the streaming instability (see Kulsrud and Pearce, 1969). The suppression of streaming instability by various damping effects in the multi-phase interstellar medium (Plotnikov et al., 2021; Xu and Lazarian, 2022; Sampson et al., 2023) and by MHD turbulence (Yan and Lazarian, 2002; Farmer and Goldreich, 2004; Lazarian, 2016) is an important effect that modifies the CR propagation (see Krumholz et al., 2020 and review by Lazarian and Xu, 2022). Below we discuss the CR propagation in the pre-existing MHD turbulence.

The interaction of CRs with magnetic turbulence has been investigated for decades (Jokipii, 1966; Kulsrud and Pearce, 1969; Schlickeiser and Miller, 1998; Giacalone and Jokipii, 1999), mostly with *ad hoc* models of MHD turbulence, including superposition of MHD waves (Giacalone and Jokipii, 1999), isotropic MHD turbulence (see Schlickeiser, 2002 and references therein), and 2D + slab superposition model for solar wind turbulence (Matthaeus et al., 1990; Zank et al., 2017). More recently, both MHD simulations with driven turbulence (Cho and Vishniac, 2000; Maron and Goldreich, 2001; Cho et al., 2002; Cho and Lazarian, 2003; Kowal and Lazarian, 2010; Beresnyak, 2014) and solar wind observations (e.g., Horbury et al., 2008; Luo and Wu, 2010; Forman et al., 2011) show evidence for the statistical properties of MHD turbulence corresponding to theoretical expectations (Goldreich and Sridhar, 1995; Lazarian and Vishniac, 1999, see also Beresnyak and Lazarian, 2019 and references therein). As we discuss in the review, the progress in the understanding of MHD turbulence brings significant changes in understanding CR propagation[1].

In what follows, in §2 we briefly discuss the basic properties and scaling relations of turbulent magnetic fields. §3 deals with CR superdiffusion and diffusion in the direction perpendicular to the mean magnetic field that arise from the perpendicular superdiffusion and diffusion of turbulent magnetic fields. In §4, we discuss different mechanisms leading to the parallel diffusion of CRs, including a recently identified new mechanism, the mirror diffusion, that can significantly suppress the CR diffusion in the vicinity of CR sources. A summary is provided in §5.

1 For the sake of simplicity, we discuss only the so-called "balanced" MHD turbulence, i.e., MHD turbulence with equal fluxes of Alfvén energy in opposite directions. This is not the case for Solar wind over distances less than 1 AU. However, the turbulence gets balanced at larger scales. The theory of imbalanced MHD turbulence is described in the book by Beresnyak and Lazarian 2019.

2 Properties of MHD turbulence that affect CR propagation

2.1 Compressible MHD turbulence and fast modes

We consider turbulence as a result of non-linear interactions inducing the energy cascade from large to small scales. Therefore, other types of excitations, including the streaming instability (see Kulsrud and Pearce, 1969; Farmer and Goldreich, 2004; Lazarian, 2016), gyro-resonance instabilities (see Lazarian and Beresnyak, 2006), are not classified as turbulence. The changes in plasma conditions affect MHD turbulence and, through this, the CR propagation. For instance, partial ionization of astrophysical media can significantly affect the dynamics of CRs (Xu et al., 2016).

The turbulence in magnetized media in typical astrophysical settings is injected at a large scale L, i.e., the injection scale, with the injected turbulent velocity at L equal to V_L. The values of L and V_L depend on the driving mechanism of turbulence. For instance, for the interstellar turbulence driven by supernova explosions, the typical values of L and V_L are ~100 pc and ~10 km s^{-1} (Chamandy and Shukurov, 2020). We note that, unlike the bulk fluid velocity, the turbulent velocity is the averaged velocity difference between two points in space separated by a given length scale. For MHD turbulence, if the injection of turbulence happens with $V_L > V_A$, where V_A is the Alfvén speed, this is the case of super-Alfvénic turbulence[2]. If $V_L < V_A$, the turbulence is sub-Alfvénic. The ratio $M_A = V_L/V_A$ is the Alfvén Mach number, which is an important parameter characterizing the properties of MHD turbulence. The special case with $M_A = 1$ corresponds to the trans-Alfvénic turbulence, with energy equipartition between turbulence and magnetic fields at L. The trans-Alfvénic turbulence was originally studied in Goldreich and Sridhar (1995). Other turbulence regimes were covered in Lazarian and Vishniac (1999). **Figure 1** illustrates the different magnetic field structures in sub- and super-Alfvénic turbulence. The turbulent energy then cascades down to a small scale l_d that in most cases is determined by microphysical processes. The range of scales between L and l_d is the inertial range, and it is usually a very extended range in typical astrophysical settings. For instance, **Figure 2** illustrates the Galactic electron density spectrum (Chepurnov and Lazarian, 2010) induced by the interstellar turbulence in the warm ionized medium, and the velocity spectrum towards Taurus molecular cloud (Yuen et al., 2022). The properties of turbulence over the inertial range are the most relevant for the CR physics that we deal with in this review. All the scaling relations that we will discuss are within the inertial range of turbulence.

The current understanding of MHD turbulence differs significantly from the earlier models adopted for CR studies in the literature (e.g., Matthaeus et al., 1990; Kóta and Jokipii, 2000; Qin et al., 2002). MHD turbulence in a compressible medium (see Beresnyak and Lazarian, 2019) can be approximated as a

2 Super-Alfvénic turbulence should be distinguished from super-Alfvénic (e.g., solar wind) flow. In super-Alfvénic turbulence, it is the averaged relative velocity at L, rather than the bulk flow velocity, that exceeds V_A.

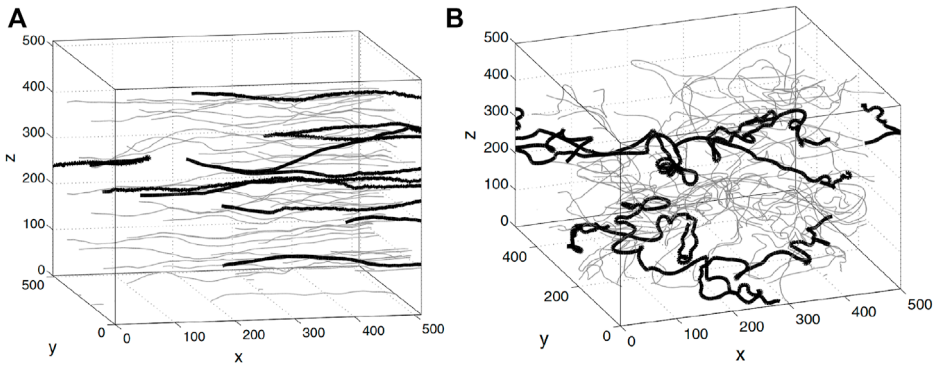


FIGURE 1
Magnetic field lines (thin lines) and trajectories of CRs (thick lines) in **(A)** sub-Alfvénic turbulence with $M_A = 0.3$ and **(B)** super-Alfvénic turbulence with $M_A = 1.5$. From Xu and Yan (2013).

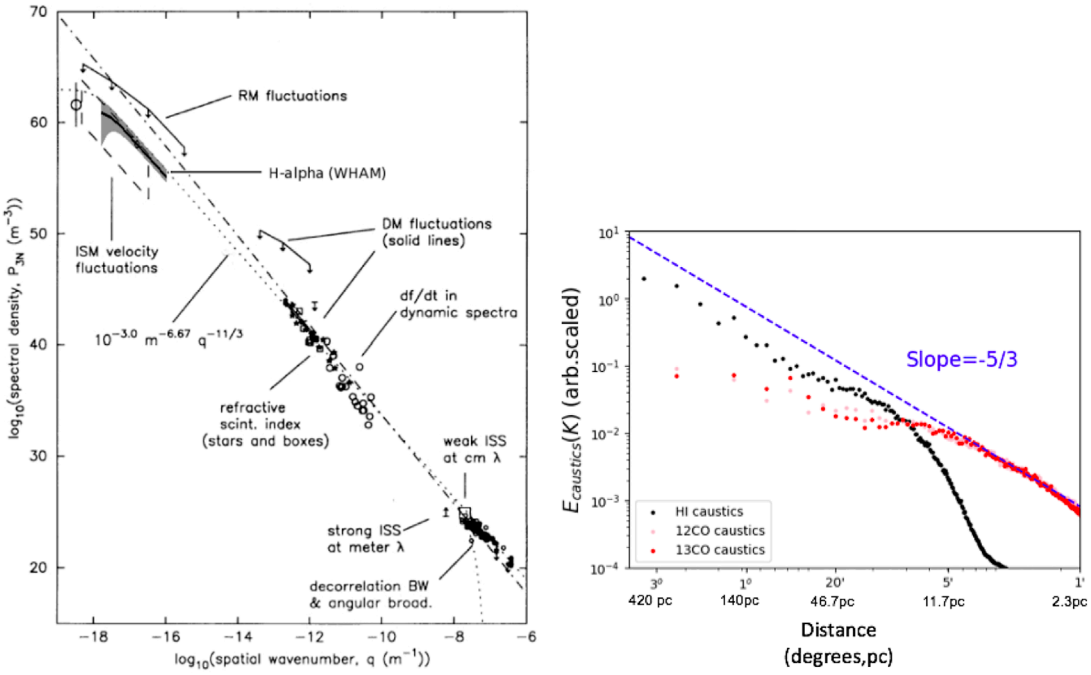


FIGURE 2
Left panel: "Extended Big Power Law in the Sky" of Galactic electron density fluctuations obtained by combining the scattering measurements of nearby pulsars in Armstrong et al. (1995) and H_α measurements from WHAM data in Chepurnov and Lazarian (2010). Right panel: Power law of velocity fluctuations measured toward the Taurus molecular cloud. The horizontal axis corresponds to $\log_{10}$(spatial wavenumber) and the slope in the right panel is consistent with that in the left panel. From Yuen et al. (2022).

superposition of three cascades of fundamental modes, i.e., Alfvén, slow, and fast modes. The Alfvén modes play a dominant role in regulating the dynamics of MHD turbulence and shaping turbulence anisotropy. Unlike long-lived wave oscillations about a fixed equilibrium point in space, the reconnection diffusion of turbulent magnetic fields enables their turbulent motions in the direction perpendicular to the local magnetic field (Lazarian and Vishniac, 1999), with a limited lifetime determined by the turbulent eddy turnover time in the strong MHD turbulence regime. Therefore, we use the term "modes" rather than "waves" for describing MHD turbulence. Another important property of MHD turbulence is its scale-dependent anisotropy. Matthaeus et al. (1996) reported the anisotropy being greater for smaller wavenumber. By contrast, larger anisotropy toward smaller length scales (larger wavenumber) is seen in MHD turbulence simulations in the reference frame with respect to the local magnetic field (Cho and Vishniac, 2000; Cho and Lazarian, 2002a; Cho and Lazarian, 2003). The latter is well expected by the MHD turbulence theory (Goldreich and Sridhar,

1995; henceforth GS95) based on the critical balance relation of Alfvén modes[3].

The numerical decomposition of MHD turbulence into different modes has demonstrated that the interaction between fast modes and other modes, i.e., Alfvén and slow modes, is relatively weak for sub-Alfvénic non-relativistic MHD turbulence (Cho and Lazarian, 2002a). Therefore, the cascade of fast modes can be assumed independent of the cascades of Alfvén and slow modes. The cascade of fast modes is similar to the acoustic one in a high β medium, where β is the ratio of the plasma to magnetic pressure. In this regime, fast modes are mostly compressions of plasmas that propagate with the sound speed c_s. It is also shown by Cho and Lazarian (2002a) that in the opposite limit of a low β medium, fast modes are expected to form a cascade similar to the acoustic one, even though the fluctuations are compressions of magnetic fields that propagate with $\sim V_A$. The numerical simulations by Cho and Lazarian (2002a), Cho and Lazarian (2003) support that the cascade of fast modes is very similar to the acoustic cascade for various values of β.

The acoustic turbulence in non-magnetized gas has the energy spectrum $E_s(k) \sim k^{-3/2}$ for low-amplitude perturbations. The spectrum tends to steepen with $E_s(k) \sim k^{-2}$ as the amplitude of perturbations increases. For fast modes in a high β medium, the perturbation amplitude increases with increasing sonic Mach number $M_s = V_L/c_s$. For fast modes in a low-β medium, it increases with the increase of M_A. The numerical results in Cho and Lazarian (2003) are consistent with $E_f \sim k^{-3/2}$, while those in Kowal and Lazarian (2010); Hu et al. (2022b) are better fitted by $E_f \sim k^{-2}$ (see **Figure 3**). The difference may be accounted for by appealing to the analogy with the acoustic turbulent cascade mentioned above. However, the issue has not been settled. In any case, similar to the acoustic cascade, fast modes have isotropic energy distribution.

2.2 Properties of Alfvén and slow modes

Slow modes are cascaded by Alfvén modes, and their anisotropy is imposed by that of Alfvén modes (Goldreich and Sridhar, 1995; Lithwick and Goldreich, 2001; Cho and Lazarian, 2002a; Cho and Lazarian, 2003). **Figure 3** shows the numerically measured turbulent kinetic energy spectra of decomposed MHD modes. As expected, the spectral scalings of Alfvén and slow modes in both compressible and incompressible media are similar (Cho and Lazarian, 2003). This allows appealing to high-resolution simulations of incompressible MHD turbulence (Beresnyak, 2014) to test their spectral scalings.

The turbulent motions, rather than wave oscillations, of magnetic fields in Alfvén modes are enabled by the turbulent reconnection of magnetic fields (Lazarian and Vishniac, 1999) (henceforth LV99), which is an intrinsic part of MHD turbulence. The LV99 theory treats Alfvén modes as a collection of turbulent eddies whose axes of rotation are aligned with the local magnetic field surrounding them. Due to turbulent reconnection, the fluid turbulent motions perpendicular to the local magnetic field are not constrained. As a result, an eddy with scale $l_\perp$ perpendicular to the local magnetic field induces Alfvénic perturbation of scale $l_\parallel$ that propagates along the magnetic field with speed V_A. The timescale of this perturbation $l_\parallel/V_A$ should be equal to the eddy turnover time $l_\perp/v_l$. The corresponding relation between the parallel and perpendicular sizes of the eddy:

$$\frac{l_\parallel}{V_A} \approx \frac{l_\perp}{v_l}, \tag{1}$$

constitutes the modern understanding of the critical balance for Alfvénic turbulence (GS95). It follows that the GS95 theory is only valid in the so-called local system of reference defined by the local mean magnetic field of the eddies. This fact is confirmed by numerical simulations (Cho and Vishniac, 2000; Maron and Goldreich, 2001; Cho et al., 2002). The direction of the local magnetic field at a given length scale can differ significantly from the global mean magnetic field direction resulting from the large-scale averaging. For CR propagation, it is important that the local small-scale magnetic field is the magnetic field sampled by CRs. This justifies the perturbative approach for describing the propagation of low-energy CRs, even though the large-scale variations of the magnetic field can be significant.

As the turbulent reconnection allows the turbulent cascade in the direction perpendicular to the local magnetic field, this cascade is not affected by the back-reaction of the magnetic field and remains Kolmogorov-like. For trans-Alfvénic turbulence with $V_L = V_A$, this means that:

$$v_l \approx V_A \left(\frac{l_\perp}{L}\right)^{1/3}, \tag{2}$$

where v_l is the turbulent velocity at $l_\perp$, and $l_\perp$ is the perpendicular size of a turbulent eddy. By combining Eqs **1**, **2**, it is easy to obtain the scale-dependent anisotropy of trans-Alfvénic turbulence in the local system of reference:

$$l_\parallel \approx L \left(\frac{l_\perp}{L}\right)^{2/3}. \tag{3}$$

From Eq. **3**, it is evident that smaller eddies are more elongated along the local magnetic field. We note, however, that Eqs **2**, **3** should be understood in the statistical sense. They represent the scaling relations between the most probable values of the quantities involved rather than the properties of individual eddies. The eddies have a distribution of scales. Using MHD turbulence simulations, Cho et al. (2002) provided an analytical fit to the probability distribution function of $l_\perp$ at a given $l_\parallel$. This distribution was adopted later in Yan and Lazarian (2002), Yan and Lazarian (2004) as well as the subsequent studies on CR scattering.

In super-Alfvénic turbulence, the magnetic field is of marginal importance at scales near L. Therefore, super-Alfvénic turbulence has an isotropic Kolmogorov energy spectrum at large scales. However, as the turbulent velocity decreases along the energy cascade, i.e., $v_l \sim V_L\,(l/L)^{1/3}$, the effect of the magnetic field becomes more and more manifested. At the Alfvénic scale (Lazarian, 2006)

$$l_A = LM_A^{-3}, \tag{4}$$

v_l becomes equal to V_A, and the turbulence becomes fully magnetohydrodynamic. Its properties at scales smaller than l_A can

3 However, in the global reference frame with respect to the mean magnetic field adopted in Goldreich and Sridhar (1995), the scale-dependent anisotropy is not expected. The local reference frame, i.e., the only reference frame for one to see the scale-dependent anisotropy, was introduced in Lazarian and Vishniac (1999).

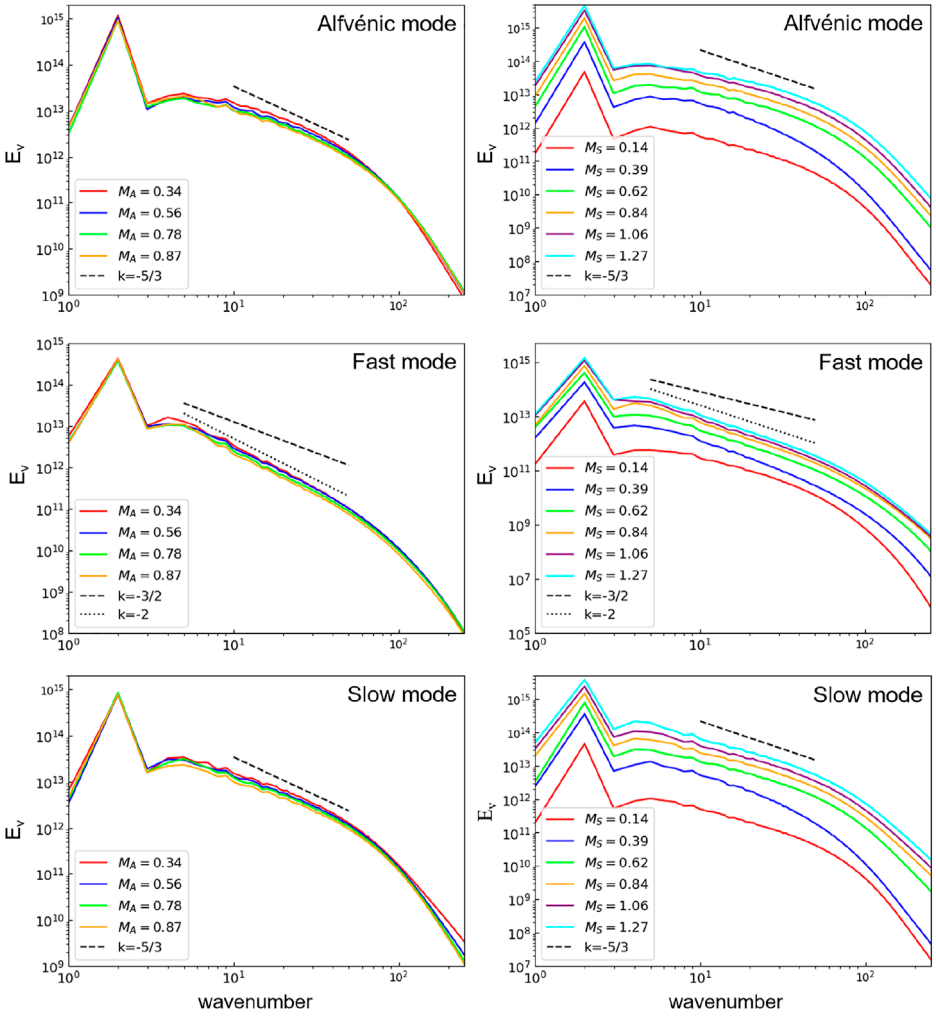


FIGURE 3
Left: The turbulent kinetic energy spectra for Alfvén (top), fast (middle), and slow (bottom) modes for various M_A values and $M_s \approx 0.6$. Right: Same as left panels but for various M_s values and $M_A \approx 0.5$. From Hu et al. (2022b).

be described by trans-Alfvénic scaling, provided that L in Eqs 2, 3 is replaced by l_A.

In sub-Alfvénic turbulence, it was shown in LV99 that below L, there is a range of scales where the cascade is in the weak turbulence regime. In this regime, the parallel scale of wave packets remains unchanged, i.e., $l_\parallel = L$, and oppositely directed wave packets have to interact multiple times to get cascaded. The scaling obtained in LV99 for the weak turbulence[4] under the assumption of the isotropic turbulence driving at L is

$$v_l \approx V_L \left(\frac{l_\perp}{L}\right)^{1/2}, \tag{5}$$

and this result was supported by the subsequent analytical study by Galtier et al. (2000). With the decrease of $l_\perp$, the intensity of interactions of Alfvénic perturbations increases despite the decrease in the turbulence amplitude. It was shown in LV99 that at the transition scale

$$l_{\rm tran} \approx L M_A^2, \tag{6}$$

where $M_A < 1$, the turbulence gets into the strong turbulence regime. For the sub-Alfvénic turbulence at $l < l_{\rm tran}$, there are

$$v_l \approx V_L \left(\frac{l_\perp}{L}\right)^{1/3} M_A^{1/3}, \tag{7}$$

and:

$$l_\parallel \approx L \left(\frac{l_\perp}{L}\right)^{2/3} M_A^{-4/3}. \tag{8}$$

The above relations derived in LV99 differ from Eqs 2, 3 for trans-Alfvénic turbulence by the additional dependence on M_A.

To unify the formalism of scalings of super- and sub-Alfvénic turbulence, Lazarian and Xu (2021) (henceforth LX21) proposed to introduce an effective injection scale:

$$L_{\rm eff} = L M_A^{-4}. \tag{9}$$

4 Weak turbulence is weak in terms of the non-linear interactions of oppositely directed wave packets.

Using $L_{\rm eff}$, Eqs 7, 8 can be rewritten as:

$$v_l \approx V_A\left(\frac{l_\perp}{L_{\rm eff}}\right)^{1/3}, \quad (10)$$

and:

$$l_\parallel \approx L_{\rm eff}\left(\frac{l_\perp}{L_{\rm eff}}\right)^{2/3}, \quad (11)$$

which take the same forms as Eqs 2, 3 with L replaced by $L_{\rm eff}$. We note that unlike l_A for super-Alfvénic turbulence, $L_{\rm eff}$ does not have a particular physical meaning. LX21 introduced $L_{\rm eff}$ to present the turbulent scaling on scales less than $l_{\rm tran}$ for sub-Alfvénic turbulence in a more convenient way.

2.3 Interactions of CRs with MHD turbulence

Without magnetic fields, CRs would propagate ballistically and easily escape from cosmic accelerators and diffuse astrophysical media. It is MHD turbulence that induces pitch-angle scattering, stochastic acceleration, and spatial diffusion of CRs along the magnetic field (e.g., Chandran, 2000b; Yan and Lazarian, 2002; Yan and Lazarian, 2004; Brunetti and Lazarian, 2007; Strong et al., 2007; Yan and Lazarian, 2008; Blasi and Amato, 2012; Lynn et al., 2012; Xu and Yan, 2013; Xu and Lazarian, 2018; Lemoine and Malkov, 2020; Sioulas et al., 2020; Kempski and Quataert, 2022), as well as superdiffusion and diffusion perpendicular to the mean magnetic field (Yan and Lazarian, 2008; Xu and Yan, 2013; Lazarian and Yan, 2014; Hu et al., 2022b; Maiti et al., 2022).

Traditionally, the transport of CRs is separated in two categories, the transport along a magnetic field and the transport perpendicular to the magnetic field. In general, parallel and perpendicular transport take place simultaneously. Magnetic fluctuations are assumed to be small and therefore the magnetic field direction approximately coincides with that of the mean magnetic field. This is a very idealized presentation of the actual CR propagation in realistic astrophysical environments. In reality, the magnetic fluctuations are comparable to the mean magnetic field for trans-Alfvénic turbulence and are larger than the mean field for super-Alfvénic turbulence. Moreover, LX21 demonstrated that the processes of parallel and perpendicular diffusion can be interdependent.

The field line wandering causes the diffusion of CRs perpendicular to the mean magnetic field. This effect was discussed extensively in the literature (Jokipii, 1966). However, perpendicular diffusion only applies on scales larger than L. On scales less than L, LV99 showed that the magnetic field lines are superdiffusive and exhibit fast dispersion, which is also numerically demonstrated by Lazarian et al. (2004); Beresnyak (2013). This fast separation of magnetic field lines induces superdiffusive transport of CRs perpendicular to the mean magnetic field direction (Lazarian and Yan, 2014). As we will discuss further, the superdiffusion of CRs is an essential but frequently ignored aspect of CR propagation.

As CRs move along magnetic field lines, they interact with magnetic fluctuations, including resonant (e.g., scattering, transit-time damping) and non-resonant (e.g., mirroring) interactions. These interactions affect the parallel transport of CRs. The well-known interaction of CRs with magnetic fluctuations is gyroresonant scattering (i.e., pitch-angle scattering). It occurs as magnetic fluctuations induce fluctuations of the electric field in the reference frame of CRs with a frequency equal to the Larmor frequency of CRs. All MHD modes induce pitch angle scattering. In addition, a special type of interaction is related to compressible fluctuations, i.e., slow and fast modes, and is termed Transit-time Damping (TTD) interaction (Schlickeiser, 2002; Xu and Lazarian, 2018). This sort of interaction is associated with the surfing of CRs on the front of an oblique compressible wave. It changes the CR momentum parallel to the local magnetic field.

In super-Alfvénic turbulence, when the parallel mean free path of CRs is larger than l_A, the tangling of field lines acts to confine CRs, with an effective mean free path equal to l_A (Brunetti and Lazarian, 2007), and the CR diffusion becomes isotropic. Super-Alfvénic turbulence in molecular clouds and starburst galaxies plays an important role in determining the diffusion of CRs (Krumholz et al., 2020; Xu, 2021).

The new understanding of the statistics and dynamics of turbulent magnetic fields can shed light on some long-standing problems and observational puzzles (e.g., Palmer, 1982; Evoli and Yan, 2014; López-Barquero et al., 2016; Krumholz et al., 2020; Xu, 2021). It was found that the numerically tested model of MHD turbulence, rather than isotropic Kolmogorov turbulence, can satisfactorily interpret the high-precision AMS-02 measurements of CRs (Fornieri et al., 2021). Updated multifrequency observations and direct CR measurements (e.g., Nava and Gabici, 2013; Orlando, 2018; Gabici et al., 2019; Amato and Casanova, 2021) request improved understanding on interactions of CRs with MHD turbulence, as well as measurements on the properties of astrophysical turbulent magnetic fields (Lazarian et al., 2018; Hu et al., 2021; Hu et al., 2022a; Xu and Hu, 2021).

3 CR propagation perpendicular to the mean magnetic field

The perpendicular (super) diffusion of CRs arises from that of turbulent magnetic fields. In the direction perpendicular to the mean magnetic field, the mean squared displacement $\langle y^2\rangle$ of CRs is proportional to s^α, where s is the distance traveled by CRs along the magnetic field. $\alpha = 1$ corresponds to normal diffusion, $\alpha > 1$ corresponds to superdiffusion, and $\alpha < 1$ corresponds to subdiffusion. Subdiffusion, i.e., compound diffusion, was proposed in, e.g., Kóta and Jokipii (2000) under the assumption that CRs are able to retrace the same magnetic field line. This is not possible due to the stochasticity of turbulent magnetic fields (Yan and Lazarian, 2008; Lazarian and Yan, 2014; Lazarian and Xu, 2021). Subdiffusion of CRs is not observed in MHD turbulence simulations (Xu and Yan, 2013; Hu et al., 2022b).

3.1 Perpendicular superdiffusion of CRs

Based on the kinetic theory, a CR particle is approximately shifted one Larmor radius perpendicular to the magnetic field during one scattering mean free path $\lambda_\parallel$ (Jokipii, 1987). The resulting perpendicular diffusion is usually negligible with $r_L \ll \lambda_\parallel$, where r_L is the Larmor radius. Even in the limited case of strong scattering,

i.e., in the case of the so-called Bohm diffusion, the perpendicular diffusion coefficient does not exceed $r_L c$, where c is the speed of light. When the scattering is less efficient, CRs follow magnetic field lines during the time between the scattering events. As a result, the CR dynamics are very much affected by the stochasticity and (super) diffusion of turbulent magnetic fields. On scales smaller than L, the superdiffusion of magnetic fields was identified in LV99, which is a natural consequence of turbulent energy cascade and reconnection diffusion of magnetic fields (Lazarian, 2014).

The LV99 scaling for magnetic field superdiffusion can be obtained by considering the energy cascade of Alfvén modes. It is natural to assume that when one follows the magnetic field line over the parallel size $l_\parallel$ of a turbulent eddy, the magnetic field line undergoes perpendicular displacement equal to the transverse size $l_\perp$ of the eddy. The displacement $l_\perp$ can be either positive or negative, and therefore the dispersion $\langle y^2\rangle$ of magnetic field lines increases in a random-walk manner with

$$d\langle y^2\rangle \approx l_\perp^2 \frac{ds}{l_\parallel}, \tag{12}$$

where ds is the distance measured along the magnetic field line, and the $\langle\ldots\rangle$ denotes an ensemble average. Using the above relation and the scaling relation between $l_\parallel$ and $l_\perp$ for sub-Alfvénic turbulence given by Eq. **8** and associating $l_\perp^2$ with $\langle y^2\rangle$, one gets (LV99):

$$d\langle y^2\rangle \approx \langle y^2\rangle^{2/3} M_A^{4/3} L^{-1/3} ds, \tag{13}$$

which indicates an accelerated separation of field lines. Superdiffusion of magnetic fields in MHD turbulence is analogous to Richardson dispersion (Richardson, 1926) in hydrodynamic turbulence. Numerically, the superdiffusion of turbulent magnetic fields was demonstrated in Lazarian et al. (2004); Beresnyak (2013). The accelerated separation of magnetic field lines manifests that the "frozen-in" condition is grossly violated in turbulent media. More discussion on the relation between superdiffusion and flux-freezing breakdown can be found in Eyink et al. (2011). The latter has been numerically confirmed by Eyink et al. (2013).

The essence of superdiffusion is easy to understand. As one follows neighboring magnetic field lines over the distance s, the divergence rate of field lines increases as larger and larger turbulent eddies contribute to the dispersion of field line separations. In terms of CRs dynamics, Eq. **13** also applies to the dispersion of separations of CRs that move ballistically along magnetic field lines. The resulting perpendicular divergence of CR trajectories was studied in Yan and Lazarian (2008); Lazarian and Yan (2014), with the analytical predictions numerically confirmed in Xu and Yan (2013). The dispersion of CR separations in the direction perpendicular to the mean magnetic field in sub-Alfvénic turbulence is:

$$\langle y_{cr}^2\rangle \approx \frac{s^3}{L} M_A^4, \quad M_A<1,\ s<\lambda_\parallel,\ s<L, \tag{14}$$

where s is the distance traveled by CRs along the magnetic field line, and $\lambda_\parallel$ is the CR parallel mean free path. In super-Alfvénic turbulence, there is

$$\langle y_{cr}^2\rangle \approx \frac{s^3}{L} M_A^3, \quad M_A>1,\ s<\lambda_\parallel,\ s<l_A. \tag{15}$$

In the case with efficient scattering, the arguments above can be generalized to describe the perpendicular superdiffusion of CRs while they propagate diffusively along the field lines. We have

$$s^2 \approx D_\parallel t \approx \lambda_\parallel v_{cr} t, \tag{16}$$

where v_{cr} is the CR velocity, and $D_\parallel$ is the parallel diffusion coefficient. The average angle of a given magnetic field line with respect to the mean magnetic field is approximately

$$\alpha(s) \approx \frac{\langle y^2(s)\rangle^{1/2}}{s} \approx \frac{s^{1/2}}{L^{1/2}} M_A^2, \tag{17}$$

where Eq. **14** is used. It is easy to see that in the system of reference with respect to the mean magnetic field, the perpendicular displacement of CRs is $\lambda_{cr,\perp} \approx \alpha(s)\lambda_\parallel$ as they move over $\lambda_\parallel$ along the mean magnetic field. The perpendicular to the mean magnetic field component of CR velocity is $v_{cr,\perp} \approx \alpha(s) v_{cr,\parallel}$, where $v_{cr,\parallel}$ is the CR velocity along the mean magnetic field. The perpendicular propagation is a random walk with the step size $\lambda_{cr,\perp}$ and the velocity $v_{cr,\perp}$ that increase with s. Therefore, for a given s the perpendicular diffusion coefficient is $\sim v_{cr,\perp}\lambda_{cr,\perp} \sim \alpha^2(s) D_\parallel$. The motion of CRs perpendicular to the mean magnetic field can be described as

$$\langle y_{cr}^2\rangle \approx \alpha^2 D_\parallel t \approx \frac{s^3}{L} M_A^4, \quad M_A<1,\ \ \lambda_\parallel \ll s, \tag{18}$$

It shows a similar superdiffusion as the case of inefficient scattering (Eq. **14**).

We note that the concept of perpendicular superdiffusion of CRs contradicts some existing theories on small-scale CR transport, e.g., the Non-Linear Guiding Center Theory (NLGC) (Matthaeus et al., 2003). There it is assumed that in both parallel and perpendicular directions with respect to the mean magnetic field, the propagation of CRs is diffusive. However, test particle simulations in MHD turbulence support perpendicular superdiffusion of CRs on scales less than L in both cases with ballistic and diffusive motions of CRs along magnetic field lines (Xu and Yan, 2013; Hu et al., 2022b).

3.2 Perpendicular diffusion of CRs

The perpendicular diffusion of magnetic field lines in MHD turbulence takes place on scales larger than l_A for super-Alfvénic turbulence and $l_{\rm tran}$ for sub-Alfvénic turbulence (see Table 1 in Lazarian and Yan, 2019). For the former, magnetic field lines get entangled on the scale l_A and this induces the random walk with the step l_A, i.e.,:

$$\langle y^2\rangle \approx l_A^2 \frac{s}{l_A} = s l_A = s L M_A^{-3}, \quad M_A>1, \quad s \gg l_A. \tag{19}$$

For sub-Alfvénic turbulence, when perpendicular scales are larger than $l_{\rm tran}$, turbulence is in the weak turbulence regime, and the growth of magnetic field separation is diffusive with the step in the perpendicular direction LM_A^2 for every transposition along the magnetic field by L. Thus for $s \gg L$, there is

$$\langle y^2\rangle \approx (LM_A^2)^2 \frac{s}{L} = s L M_A^4, \quad M_A<1, \quad s \gg L. \tag{20}$$

For the CRs moving ballistically along magnetic field lines with $s<\lambda_\parallel$, the perpendicular diffusion of CRs is also given by Eqs **19**, **20**. If the CR propagation along the magnetic field is diffusive with $s \gg \lambda_\parallel$, the perpendicular diffusion has dependence on the parallel diffusion.

In super-Alfvénic turbulence at scales $s \gg l_A$, the CR diffusion is isotropic. The characteristic diffusion coefficient is

$$D \approx v_{cr} X, \quad X = min\left[\lambda_\parallel, l_A\right]. \tag{21}$$

In sub-Alfvénic turbulence, if $L \ll s < \lambda_\parallel$, Eq. **20** leads to the perpendicular diffusion coefficient of CRs (Yan and Lazarian, 2008)

$$D_\perp \approx v_{cr} L M_A^4. \tag{22}$$

If $\lambda_\parallel \ll L \ll s$, CRs propagate diffusively over L along the magnetic field line. Below, we reproduce the derivation of the diffusion coefficient in Lazarian (2006)[5]. According to Eq. **20** the magnetic field lines that guide the perpendicular diffusion undergo a random walk. As the magnetic field line is traced over L, its perpendicular transposition is LM_A^2. To cover the distance $\langle y_{cr}^2\rangle^{1/2}$ with the step size LM_A^2, it requires $N = (\langle y_{cr}^2\rangle^{1/2}/LM_A^2)^2$ steps. The time required for each step is $\delta t_{step} \approx L^2/D_\parallel$, and the total time is $\delta t \approx N\delta t_{step}$. Substituting this into the expression for the diffusion coefficient one gets

$$D_\perp \approx \frac{\langle y_{cr}^2\rangle}{\delta t} \approx D_\parallel M_A^4. \tag{23}$$

This agrees with the CR perpendicular diffusion coefficient in (Yan and Lazarian, 2008). This diffusion coefficient differs from that in the literature (see Jokipii, 1966) by having a dependence on M_A^4 rather than M_A^2. The M_A^4 dependence was numerically confirmed in Xu and Yan (2013).

4 Propagation of CRs along the magnetic field

4.1 Gyroresonance scattering

Gyroresonance scattering, i.e., pitch angle scattering, requires that the Doppler-shifted wave frequency is equal to the gyrofrequency of the particle or its harmonics. Through the gyroresonant interaction with magnetic fluctuations, CRs experience diffusion in their pitch angles (the angle between the particle velocity and magnetic field direction) while moving along the local magnetic field. The distance that CRs travel along the magnetic field corresponding to the change of pitch angles by 90° is $\lambda_\parallel$.

For theoretical studies, the Quasilinear Theory (QLT) (Jokipii, 1966) is frequently adopted in the literature. In QLT, particles are assumed to propagate along the magnetic field with infinitesimal fluctuations. Given $r_L \ll L$ and the turbulent energy cascade, the magnetic fluctuations that satisfy the gyroresonance condition can be sufficiently small for the above assumption to be valid. However, the QLT faces the so-called 90° problem, with vanishing scattering close to 90° and thus infinitely large $\lambda_\parallel$. The resonance broadening induced by, e.g., magnetic fluctuations, can help alleviate but not fully resolve the problem (Xu and Lazarian, 2018).

A remarkable consequence of the modern MHD turbulence theory is the inefficiency of gyroresonance scattering of CR with $r_L \ll L$ by Alfvén and slow modes. This effect was studied by Chandran (2000b); Yan and Lazarian (2002, 2003). Yan and Lazarian (2002, 2003) identified fast modes as the main scattering agent of CRs with $r_L \ll L$ in MHD turbulence. Yan and Lazarian (2002), Yan and Lazarian (2003), Xu and Lazarian (2020) adopted the anisotropic distribution of magnetic fluctuations measured from MHD turbulence simulations (Cho et al., 2002). This significantly shifted the paradigm of CR propagation and turbulent acceleration, as in earlier studies, Alfvén modes were mostly considered as the source of gyroresonance scattering.

The inefficiency of gyroresonance scattering by Alfvén and slow modes for CRs with $r_L \ll L$ arises from the scale-dependent anisotropy. The gyroresonance condition requires that r_L is comparable to $l_\parallel$. However, with scale-dependent anisotropy, there is $l_\perp \ll l_\parallel$ for small-scale magnetic fluctuations. As a result, a CR with $r_L \sim l_\parallel$ samples many uncorrelated turbulent eddies within one gyro-orbit. In addition, the energy cascade of Alfvén and slow modes is mainly in the direction perpendicular to the local magnetic field. With a steep parallel energy spectrum $E(k_\parallel) \sim k_\parallel^{-2}$ (Beresnyak, 2015), the power that induces the gyroresonance scattering falls relatively fast with the decrease of r_L. Both effects cause the inefficiency of gyroresonance scattering by Alfvén and slow modes (Chandran, 2000b; Yan and Lazarian, 2002; Yan and Lazarian, 2003; Xu and Lazarian, 2020).

As suggested by numerical simulations (Cho and Lazarian, 2003), fast modes have isotropic energy distribution. As a result, their interaction with CRs is not reduced due to geometrical factors. However, compared with Alfvén and slow modes, fast modes are more subject to damping effects because of their slower cascading rate. The anisotropic nature of collisionless damping induces the preferential suppression of the modes with wavevectors perpendicular to the magnetic field, creating mostly slab-type structures of scattering fluctuations (Yan and Lazarian, 2003; Beresnyak and Lazarian, 2007). This mitigates the supression of scattering by fast modes in the presence of significant damping (Yan and Lazarian, 2003). In a weakly ionized medium, fast modes are severely damped due to ion-neutral collisional damping (Xu et al., 2016). Only CRs with r_L larger than the damping scale can be efficiently scattered.

4.2 Transit-time damping (TTD) interaction

TTD interaction in plasma physics is usually associated with the damping of waves in collisionless plasmas. For CR studies, the TTD is the process of stochastically accelerating CRs by compressible MHD waves (Schlickeiser, 2002). The process is easy to understand if one considers compressible waves oblique to the magnetic field direction. If the phase speed of waves is $v_{\rm ph}$, and the angle between the wavefront and the magnetic field is α, then the intersection point between the wavefront and the magnetic field moves with the speed $v_{\rm inter} = v_{\rm ph}/\sin\alpha \gg v_{\rm ph}$ if α is small. If a CR moves along the magnetic field in the same direction with the speed close to $v_{\rm inter}$, it can surf the wave, gaining or losing energy. With a linear resonance function, the limitation of the process is that only waves in the limited range of oblique angles can interact with fast-moving CRs. Due to magnetic fluctuations and nonlinear decorrelation of turbulence, the resonance broadening effects in MHD turbulence plays an important role in determining the TTD efficiency by slow

5 The initial derivation was performed for the problem of thermal electron diffusion in clusters of galaxies. However, the physics of the diffusion of non-relativistic and relativistic particles are identical.

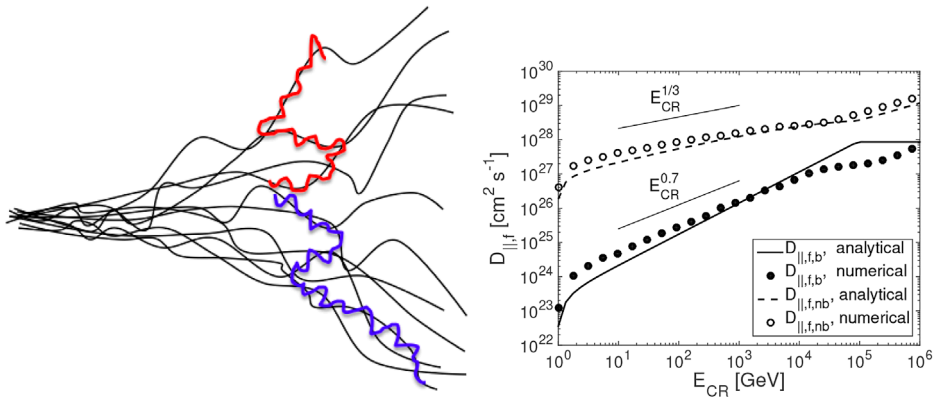


FIGURE 4
Left: Illustration of mirror diffusion. Thin lines represent turbulent magnetic field lines. Thick lines represent the trajectories of two CR particles whose initial separation is small. Right: Parallel diffusion coefficients $D_{\|,f,b}$ of mirror diffusion and $D_{\|,f,nb}$ of scattering diffusion induced by fast modes. From LX21.

and fast modes beyond the threshold of the linear resonance (Xu and Lazarian, 2018).

The change of pitch angles caused by TTD is due to the particle acceleration. TTD interaction increases the parallel component of CR momentum in a stochastic manner. Compared with gyroresonance scattering, the advantage of TTD is that it is much less subject to damping effects that cut off the turbulent energy cascade. Thus the process can act in settings where the gyroresonance scattering is inefficient (Yan and Lazarian, 2003; Beresnyak and Lazarian, 2007; Xu et al., 2016).

4.3 Mirror diffusion of CRs

In addition to the pitch-angle scattering, it has long been known that CR propagation can also be affected by magnetic mirror reflection (Fermi, 1949; Noerdlinger, 1968; Cesarsky and Kulsrud, 1973; Klepach and Ptuskin, 1995; Chandran, 2000a). In particular, the magnetic mirroring effect was explored to resolve the 90° problem that arises in the QLT describing the pitch-angle scattering (see **Section 4.1**). In these studies, the mirroring effect was invoked for trapping CRs that bounce back and forth between two mirror points. On the basis of improved understanding of MHD turbulence theory, Lazarian and Xu (2021) identified a new effect associated with CRs interacting stochastically with different mirrors, which is termed mirror diffusion. It serves as a new diffusion mechanism that can effectively confine CRs. In MHD turbulence, compressions of magnetic fields, which arise from slow and fast modes in a compressible medium and pseudo Alfvén modes in an incompressible medium, naturally, induce the mirroring effect over a range of length scales. The properties of the magnetic mirrors are determined by the scaling properties of slow (pseudo Alfvén) and fast modes. Combined with the intrinsic perpendicular superdiffusion of turbulent magnetic fields arising from Alfvénic modes (Lazarian and Vishniac, 1999; Eyink et al., 2013; Lazarian and Yan, 2014), this results in mirror diffusion of CRs along magnetic field lines. In other words, CRs are not trapped between mirrors, but exhibit a new type of diffusive propagation.

Magnetic compressions arising from slow and fast modes create magnetic mirrors that result in the reflecting of CRs. The mirroring effect caused by static magnetic bottles has been well studied in plasma physics (e.g., Post, 1958; Budker, 1959; Noerdlinger, 1968; Kulsrud and Pearce, 1969). A CR with r_L smaller than the variation scale of the magnetic field preserves its first adiabatic invariant, i.e., $p_\perp^2/B$ =const. Therefore, as a particle moves along the magnetic field with the field strength increasing from B_0 to $B_0+\delta b$, its perpendicular momentum $p_\perp$ increases. It implies that with pitch angle cosine μ

$$\mu < \mu_{lc} = \sqrt{\frac{\delta b}{B_0+\delta b}}, \tag{24}$$

the particle can be reflected at the mirror point, while particles with larger μ's, i.e., smaller pitch angles, can escape from the mirror. Earlier studies considered that the magnetic mirrors trap the CRs until the gyroresonance scattering allows the particles to escape the mirrors (Cesarsky and Kulsrud, 1973). CRs with $\mu < \mu_{lc}$ were considered "trapped" in magnetic bottles and thus unable to diffuse. However, in Lazarian and Xu (2021) it was shown that this is not true in realistic MHD turbulence. During the perpendicular superdiffusion, CRs cannot trace back the same magnetic field line. Instead, after each mirroring interaction, they always encounter a different mirror, leading to their diffusion along magnetic field lines (see **Figure 4**).

For mirroring interaction with fast modes, the corresponding parallel diffusion coefficient is (LX21).

$$D_{\|,b}(\mu) = \begin{cases} \nu\mu k^{-1} = \nu L\left(\dfrac{\delta B_f}{B_0}\right)^{-4}\mu^9, & \mu_{\min,f} < \mu < \mu_c, \\ \nu\mu r_L, & \mu < \mu_{\min,f}, \end{cases} \tag{25}$$

Where

$$\mu_c \approx \left[\frac{14}{\pi}\frac{\delta B_f^2}{B_0^2}\left(\frac{\nu}{L\Omega}\right)^{\frac{1}{2}}\right]^{\frac{2}{11}} \tag{26}$$

is the critical μ at the balance between pitch-angle scattering and mirroring,

$$\mu_{\rm min,f} = \sqrt{\frac{\delta B_f}{B_0}}\left(\frac{r_L}{L}\right)^{\frac{1}{8}}, \quad (27)$$

v is the particle velocity, Ω is the gyrofrequency, B_0 is the mean magnetic field strength, and δB_f is the magnetic fluctuation of fast modes at L. For $\mu > \mu_c$ the diffusion is determined by scattering, and the scattering diffusion coefficient is typically much larger than that of mirror diffusion (see **Figure 4**). In the vicinity of CR sources, e.g., supernova remnants, the mirror diffusion of CRs with $\mu < \mu_c$ can prevent fast escape of CRs (Xu, 2021). With pitch angle change by pitch-angle scattering, CRs undergo periods of slow mirror diffusion separated by periods of fast scattering diffusion, showing a Lévy-flight-like characteristic.

5 Summary

The progress in understanding CR propagation was hindered for decades due to an inadequate understanding of MHD turbulence. Recent development in MHD turbulence theories, MHD turbulence simulations, as well as new observational techniques in measuring turbulence and magnetic fields, bring us new physical insight into the interaction of CRs with MHD turbulence and CR diffusion. The perpendicular (super) diffusion of CRs arises from that of turbulent magnetic fields regulated by Alfvén modes of MHD turbulence. The highly tangled magnetic field lines in super-Alfvénic turbulence provide additional confinement of CRs with an effective mean free path determined by the Alfvénic scale l_A. For parallel diffusion, the widely-used gyroresonance scattering, i.e., pitch-angle scattering, is inefficient in scattering CRs with large pitch angles and low-energy CRs in the presence of severe damping of fast modes, e.g., in a weakly ionized medium. With resonance broadening effects in MHD turbulence taken into account, the transit-time damping interaction can efficiently cause the change of pitch angles via stochastic particle acceleration. With both magnetic compressions and superdiffusion of magnetic field lines in MHD turbulence, CRs interacting with turbulent magnetic mirrors undergo mirror diffusion. It results in a slow diffusion of CRs in the vicinity of their sources and a Lévy-flight-like propagation of CRs.

The perpendicular (super) diffusion and parallel diffusion of CRs strongly depend on the regime and properties of MHD turbulence. Combining the updated theoretical understanding of CR propagation and new observational techniques for mapping the characteristic parameters, e.g., M_A, in diverse astrophysical conditions (e.g., Lazarian et al., 2018; Hu et al., 2019; Hu et al., 2020; Zhang et al., 2019; Xu and Hu, 2021; Hu et al., 2022a) holds a great promise in realistic modeling of CR propagation and solving long-standing observational puzzles related to CRs.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

The research of AL and YH is supported by NASA ATP AAH7546. SX acknowledges the support for this work provided by NASA through the NASA Hubble Fellowship grant # HST-HF2-51473.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. YH is supported by NASA through award 09_0231 issued by the Universities Space Research Association, Inc. (USRA).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amato, E., and Casanova, S. (2021). On particle acceleration and transport in plasmas in the galaxy: Theory and observations. *J. Plasma Phys.* 87, 845870101. doi:10.1017/S0022377821000064

Anjos, R. C., and Catalani, F. (2020). Galactic Center as an efficient source of cosmic rays. *Phys. Rev. D.* 101, 123015. doi:10.1103/PhysRevD.101.123015

Armstrong, J. W., Rickett, B. J., and Spangler, S. R. (1995). Electron density power spectrum in the local interstellar medium. *ApJ* 443, 209–221. doi:10.1086/175515

Beresnyak, A. (2013). Asymmetric diffusion of magnetic field lines. *ApJ* 767, L39. doi:10.1088/2041-8205/767/2/L39

Beresnyak, A., and Lazarian, A. (2007). *Strong imbalanced turbulence*. B3: AGU Fall Meeting Abstracts.

Beresnyak, A., and Lazarian, A. (2019). *Turbulence in magnetohydrodynamics*.

Beresnyak, A. (2015). On the parallel spectrum in magnetohydrodynamic turbulence. *ApJ* 801, L9. doi:10.1088/2041-8205/801/1/L9

Beresnyak, A. (2014). Spectra of strong magnetohydrodynamic turbulence from high-resolution simulations. *ApJ* 784, L20. doi:10.1088/2041-8205/784/2/L20

Blasi, P., and Amato, E. (2012). Diffusive propagation of cosmic rays from supernova remnants in the galaxy. I: Spectrum and chemical composition. *J. Cosmol. Astropart. Phys.* 2012, 010. doi:10.1088/1475-7516/2012/01/010

Brunetti, G., and Jones, T. W. (2014). Cosmic rays in galaxy clusters and their nonthermal emission. *Int. J. Mod. Phys. D* 23, 1430007–1430098. doi:10.1142/S0218271814300079

Brunetti, G., and Lazarian, A. (2007). Compressible turbulence in galaxy clusters: Physics and stochastic particle re-acceleration. *MNRAS* 378, 245–275. doi:10.1111/j.1365-2966.2007.11771.x

Budker, G. I. (1959). in *Plasma physics and the problem of controlled thermonuclear reactions*, Ed. by M. A. Leontovich (Pergamon, New York), Vol. 1

Cesarsky, C. J., and Kulsrud, R. M. (1973). Role of hydromagnetic waves in cosmic-ray confinement in the disk. Theory of behavior in general wave spectra. *ApJ* 185, 153. doi:10.1086/152405

Chamandy, L., and Shukurov, A. (2020). Parameters of the supernova-driven interstellar turbulence. *Galaxies* 8, 56. doi:10.3390/galaxies8030056

Chandran, B. D. G. (2000a). Confinement and isotropization of galactic cosmic rays by molecular-cloud magnetic mirrors when turbulent scattering is weak. *ApJ* 529, 513–535. doi:10.1086/308232

Chandran, B. D. G. (2000b). Scattering of energetic particles by anisotropic magnetohydrodynamic turbulence with a goldreich-sridhar power spectrum. *Phys. Rev. Lett.* 85, 4656–4659. doi:10.1103/PhysRevLett.85.4656

Chepurnov, A., and Lazarian, A. (2010). Extending the Big power law in the Sky with turbulence spectra from Wisconsin Hα mapper data. *ApJ* 710, 853–858. doi:10.1088/0004-637X/710/1/853

Cho, J., and Lazarian, A. (2003). Compressible magnetohydrodynamic turbulence: Mode coupling, scaling relations, anisotropy, viscosity-damped regime and astrophysical implications. *MNRAS* 345, 325–339. doi:10.1046/j.1365-8711.2003.06941.x

Cho, J., and Lazarian, A. (2002a). Compressible sub-alfvénic MHD turbulence in low-β plasmas. *Phys. Rev. Lett.* 88, 245001. doi:10.1103/PhysRevLett.88.245001

Cho, J., and Lazarian, A. (2002b). Magnetohydrodynamic turbulence as a foreground for cosmic microwave background studies. *ApJ* 575, L63–L66. doi:10.1086/342722

Cho, J., Lazarian, A., and Timbie, P. T. (2012). A technique for foreground subtraction in redshifted 21 cm observations. *ApJ* 749, 164. doi:10.1088/0004-637X/749/2/164

Cho, J., Lazarian, A., and Vishniac, E. T. (2002). Simulations of magnetohydrodynamic turbulence in a strongly magnetized medium. *ApJ* 564, 291–301. doi:10.1086/324186

Cho, J., and Vishniac, E. T. (2000). The anisotropy of magnetohydrodynamic alfvénic turbulence. *ApJ* 539, 273–282. doi:10.1086/309213

Demidem, C., Lemoine, M., and Casse, F. (2020). Particle acceleration in relativistic turbulence: A theoretical appraisal. *Phys. Rev. D.* 102, 023003. doi:10.1103/PhysRevD.102.023003

Di Mauro, M., Manconi, S., and Donato, F. (2020). Evidences of low-diffusion bubbles around Galactic pulsars. *Phys. Rev. D.* 101, 103035. doi:10.1103/PhysRevD.101.103035

Evoli, C., and Yan, H. (2014). Cosmic ray propagation in galactic turbulence. *ApJ* 782, 36. doi:10.1088/0004-637X/782/1/36

Eyink, G. L., Lazarian, A., and Vishniac, E. T. (2011). Fast magnetic reconnection and spontaneous stochasticity. *ApJ* 743, 51. doi:10.1088/0004-637X/743/1/51

Eyink, G., Vishniac, E., Lalescu, C., Aluie, H., Kanov, K., Bürger, K., et al. (2013). Flux-freezing breakdown in high-conductivity magnetohydrodynamic turbulence. *Nature* 497, 466–469. doi:10.1038/nature12128

Farmer, A. J., and Goldreich, P. (2004). Wave damping by magnetohydrodynamic turbulence and its effect on cosmic-ray propagation in the interstellar medium. *ApJ* 604, 671–674. doi:10.1086/382040

Fermi, E. (1949). On the origin of the cosmic radiation. *Phys. Rev.* 75, 1169–1174. doi:10.1103/PhysRev.75.1169

Forman, M. A., Wicks, R. T., and Horbury, T. S. (2011). Detailed fit of "critical balance" theory to solar wind turbulence measurements. *ApJ* 733, 76. doi:10.1088/0004-637X/733/2/76

Fornieri, O., Gaggero, D., Cerri, S. S., De La Torre Luque, P., and Gabici, S. (2021). The theory of cosmic ray scattering on pre-existing MHD modes meets data. *MNRAS* 502, 5821–5838. doi:10.1093/mnras/stab355

Gabici, S., Evoli, C., Gaggero, D., Lipari, P., Mertsch, P., Orlando, E., et al. (2019). *The origin of galactic cosmic rays: Challenges to the standard paradigm*. arXiv e-prints , arXiv:1903.11584.

Galtier, S., Nazarenko, S. V., Newell, A. C., and Pouquet, A. (2000). A weak turbulence theory for incompressible magnetohydrodynamics. *J. Plasma Phys.* 63, 447–488. doi:10.1017/S0022377899008284

Giacalone, J., and Jokipii, J. R. (1999). The transport of cosmic rays across a turbulent magnetic field. *ApJ* 520, 204–214. doi:10.1086/307452

Ginzburg, V. L., and Syrovatskii, S. I. (1964). *The origin of cosmic rays*.

Goldreich, P., and Sridhar, S. (1995). Toward a theory of interstellar turbulence. 2: Strong alfvenic turbulence. *ApJ* 438, 763–775. doi:10.1086/175121

Guo, F., and Oh, S. P. (2008). Feedback heating by cosmic rays in clusters of galaxies. *MNRAS* 384, 251–266. doi:10.1111/j.1365-2966.2007.12692.x

Holguin, F., Ruszkowski, M., Lazarian, A., Farber, R., and Yang, H. Y. K. (2019). Role of cosmic-ray streaming and turbulent damping in driving galactic winds. *MNRAS* 490, 1271–1282. doi:10.1093/mnras/stz2568

Hopkins, P. F., Chan, T. K., Garrison-Kimmel, S., Ji, S., Su, K.-Y., Hummels, C. B., et al. (2020). But what about.: Cosmic rays, magnetic fields, conduction, and viscosity in galaxy formation. *MNRAS* 492, 3465–3498. doi:10.1093/mnras/stz3321

Horbury, T. S., Forman, M., and Oughton, S. (2008). Anisotropic scaling of magnetohydrodynamic turbulence. *Phys. Rev. Lett.* 101, 175005. doi:10.1103/PhysRevLett.101.175005

Hu, Y., Lazarian, A., Beck, R., and Xu, S. (2022a). Role of magnetic fields in fueling seyfert nuclei. *ApJ* 941, 92. doi:10.3847/1538-4357/ac9df0

Hu, Y., Lazarian, A., Li, Y., Zhuravleva, I., and Gendron-Marsolais, M.-L. (2020). Probing magnetic field morphology in galaxy clusters with the gradient technique. *ApJ* 901, 162. doi:10.3847/1538-4357/abb1c3

Hu, Y., Lazarian, A., and Xu, S. (2022b). Superdiffusion of cosmic rays in compressible magnetized turbulence. *MNRAS* 512, 2111–2124. doi:10.1093/mnras/stac319

Hu, Y., Xu, S., and Lazarian, A. (2021). Anisotropies in compressible MHD turbulence: Probing magnetic fields and measuring magnetization. *ApJ* 911, 37. doi:10.3847/1538-4357/abea18

Hu, Y., Yuen, K. H., Lazarian, V., Ho, K. W., Benjamin, R. A., Hill, A. S., et al. (2019). Magnetic field morphology in interstellar clouds with the velocity gradient technique. *Nat. Astron.* 3, 776–782. doi:10.1038/s41550-019-0769-0

Ipavich, F. M. (1975). Galactic winds driven by cosmic rays. *ApJ* 196, 107–120. doi:10.1086/153397

Jokipii, J. R. (1966). Cosmic-ray propagation. I. Charged particles in a random magnetic field. *ApJ* 146, 480. doi:10.1086/148912

Jokipii, J. R. (1982). Particle drift, diffusion, and acceleration at shocks. *ApJ* 255, 716–720. doi:10.1086/159870

Jokipii, J. R. (1971). Propagation of cosmic rays in the solar wind. *Rev. Geophys. Space Phys.* 9, 27–87. doi:10.1029/RG009i001p00027

Jokipii, J. R. (1987). Rate of energy gain and maximum energy in diffusive shock acceleration. *ApJ* 313, 842. doi:10.1086/165022

Kempski, P., and Quataert, E. (2022). Reconciling cosmic ray transport theory with phenomenological models motivated by Milky-Way data. *MNRAS* 514, 657–674. doi:10.1093/mnras/stac1240

Klepach, E. G., and Ptuskin, V. S. (1995). Propagation of cosmic rays in a medium with magnetic traps. *Astron. Lett.* 21, 411–417.

Kóta, J., and Jokipii, J. R. (2000). Velocity correlation and the spatial diffusion coefficients of cosmic rays: Compound diffusion. *ApJ* 531, 1067–1070. doi:10.1086/308492

Kowal, G., and Lazarian, A. (2010). Velocity field of compressible magnetohydrodynamic turbulence: Wavelet decomposition and mode scalings. *ApJ* 720, 742–756. doi:10.1088/0004-637X/720/1/742

Krumholz, M. R., Crocker, R. M., Xu, S., Lazarian, A., Rosevear, M. T., and Bedwell-Wilson, J. (2020). Cosmic ray transport in starburst galaxies. *MNRAS* 493, 2817–2833. doi:10.1093/mnras/staa493

Kulsrud, R., and Pearce, W. P. (1969). The effect of wave-particle interactions on the propagation of cosmic rays. *ApJ* 156, 445. doi:10.1086/149981

Lazarian, A., and Beresnyak, A. (2006). Cosmic ray scattering in compressible turbulence. *MNRAS* 373, 1195–1202. doi:10.1111/j.1365-2966.2006.11093.x

Lazarian, A. (2016). Damping of Alfvén waves by turbulence and its consequences: From cosmic-ray streaming to launching winds. *ApJ* 833, 131. doi:10.3847/1538-4357/833/2/131

Lazarian, A. (2006). Enhancement and suppression of heat transfer by MHD turbulence. *ApJ* 645, L25–L28. doi:10.1086/505796

Lazarian, A. (2014). Reconnection diffusion in turbulent fluids and its implications for star formation. *Space Sci. Rev.* 181, 1–59. doi:10.1007/s11214-013-0031-5

Lazarian, A., Vishniac, E. T., and Cho, J. (2004). Magnetic field structure and stochastic reconnection in a partially ionized gas. *ApJ* 603, 180–197. doi:10.1086/381383

Lazarian, A., and Vishniac, E. T. (1999). Reconnection in a weakly stochastic field. *ApJ* 517, 700–718. doi:10.1086/307233

Lazarian, A., and Xu, S. (2022). Damping of Alfvén waves in MHD turbulence and implications for cosmic ray streaming instability and galactic winds. *Front. Phys.* 10, 702799. doi:10.3389/fphy.2022.702799

Lazarian, A., and Xu, S. (2021). Diffusion of cosmic rays in MHD turbulence with magnetic mirrors. *ApJ* 923, 53. doi:10.3847/1538-4357/ac2de9

Lazarian, A., and Yan, H. (2019). Erratum: "Superdiffusion of cosmic rays: Implications for cosmic ray acceleration". *ApJ* 885, 170. doi:10.3847/1538-4357/ab50ba

Lazarian, A., and Yan, H. (2014). Superdiffusion of cosmic rays: Implications for cosmic ray acceleration. *ApJ* 784, 38. doi:10.1088/0004-637X/784/1/38

Lazarian, A., Yuen, K. H., Ho, K. W., Chen, J., Lazarian, V., Lu, Z., et al. (2018). Distribution of velocity gradient orientations: Mapping magnetization with the velocity gradient technique. *ApJ* 865, 46. doi:10.3847/1538-4357/aad7ff

Lemoine, M., and Malkov, M. A. (2020). Power-law spectra from stochastic acceleration. *MNRAS* 499, 4972–4983. doi:10.1093/mnras/staa3131

Lithwick, Y., and Goldreich, P. (2001). Compressible magnetohydrodynamic turbulence in interstellar plasmas. *ApJ* 562, 279–296. doi:10.1086/323470

Longair, M. S. (2011). *High energy astrophysics*.

López-Barquero, V., Farber, R., Xu, S., Desiati, P., and Lazarian, A. (2016). Cosmic-ray small-scale anisotropies and local turbulent magnetic fields. *ApJ* 830, 19. doi:10.3847/0004-637X/830/1/19

Luo, Q. Y., and Wu, D. J. (2010). Observations of anisotropic scaling of solar wind turbulence. *ApJ* 714, L138–L141. doi:10.1088/2041-8205/714/1/L138

Lynn, J. W., Parrish, I. J., Quataert, E., and Chandran, B. D. G. (2012). Resonance broadening and heating of charged particles in magnetohydrodynamic turbulence. *ApJ* 758, 78. doi:10.1088/0004-637X/758/2/78

Maiti, S., Makwana, K., Zhang, H., and Yan, H. (2022). Cosmic-ray transport in magnetohydrodynamic turbulence. *ApJ* 926, 94. doi:10.3847/1538-4357/ac46c8

Marcowith, A., Ferrand, G., Grech, M., Meliani, Z., Plotnikov, I., and Walder, R. (2020). Multi-scale simulations of particle acceleration in astrophysical systems. *Living Rev. Comput. Astrophysics* 6, 1. doi:10.1007/s41115-020-0007-6

Maron, J., and Goldreich, P. (2001). Simulations of incompressible magnetohydrodynamic turbulence. *ApJ* 554, 1175–1196. doi:10.1086/321413

Matthaeus, W. H., Ghosh, S., Oughton, S., and Roberts, D. A. (1996). Anisotropic three-dimensional MHD turbulence. *J. Geophys. Res.* 101, 7619–7629. doi:10.1029/95JA03830

Matthaeus, W. H., Goldstein, M. L., and Roberts, D. A. (1990). Evidence for the presence of quasi-two-dimensional nearly incompressible fluctuations in the solar wind. *J. Geophys. Res.* 95, 20673–20683. doi:10.1029/JA095iA12p20673

Matthaeus, W. H., Qin, G., Bieber, J. W., and Zank, G. P. (2003). Nonlinear collisionless perpendicular diffusion of charged particles. *ApJ* 590, L53–L56. doi:10.1086/376613

Melrose, D. B. (1969). On the formation of energy spectra in synchrotron sources. *Ap&SS* 5, 131–149. doi:10.1007/BF00650288

Nava, L., and Gabici, S. (2013). Anisotropic cosmic ray diffusion and gamma-ray production close to supernova remnants, with an application to W28. *MNRAS* 429, 1643–1651. doi:10.1093/mnras/sts450

Noerdlinger, P. D. (1968). An improved model for cosmic-ray propagation. *Phys. Rev. Lett.* 20, 1513–1516. doi:10.1103/PhysRevLett.20.1513

Orlando, E. (2018). Imprints of cosmic rays in multifrequency observations of the interstellar emission. *MNRAS* 475, 2724–2742. doi:10.1093/mnras/stx3280

Post, R. F. (1958). "Proc. Of Second U.N," in Int. Conf. on Peaceful Uses of Atomic Energy, 1 September- 13 September (Geneva, 245–265. Paper A/Conf. 15/P/377.32

Padovani, M., Ivlev, A. V., Galli, D., and Caselli, P. (2018). Cosmic-ray ionisation in circumstellar discs. *A&A* 614, A111. doi:10.1051/0004-6361/201732202

Palmer, I. D. (1982). Transport coefficients of low-energy cosmic rays in interplanetary space. *Rev. Geophys. Space Phys.* 20, 335–351. doi:10.1029/RG020i002p00335

Parker, E. N. (1966). The kinetic properties of the galactic cosmic-ray gas. *ApJ* 144, 916. doi:10.1086/148689

Parker, E. N. (1965). The passage of energetic charged particles through interplanetary space. *Planet. Space Sci.* 13, 9–49. doi:10.1016/0032-0633(65)90131-5

Perri, S., and Zimbardo, G. (2009). Ion superdiffusion at the solar wind termination shock. *ApJ* 693, L118–L121. doi:10.1088/0004-637X/693/2/L118

Plotnikov, I., Ostriker, E. C., and Bai, X.-N. (2021). Influence of ion-neutral damping on the cosmic-ray streaming instability: Magnetohydrodynamic particle-in-cell simulations. *ApJ* 914, 3. doi:10.3847/1538-4357/abf7b3

Qin, G., Matthaeus, W. H., and Bieber, J. W. (2002). Perpendicular transport of charged particles in composite model turbulence: Recovery of diffusion. *ApJ* 578, L117–L120. doi:10.1086/344687

Quataert, E., Thompson, T. A., and Jiang, Y.-F. (2022). The physics of galactic winds driven by cosmic rays I: Diffusion. *MNRAS* 510, 1184–1203. doi:10.1093/mnras/stab3273

Richardson, L. F. (1926). Atmospheric diffusion shown on a distance-neighbour graph. *Proc. R. Soc. Lond. Ser. A* 110, 709–737. doi:10.1098/rspa.1926.0043

Sampson, M. L., Beattie, J. R., Krumholz, M. R., Crocker, R. M., Federrath, C., and Seta, A. (2023). Turbulent diffusion of streaming cosmic rays in compressible, partially ionized plasma. *MNRAS* 519, 1503–1525. doi:10.1093/mnras/stac3207

Schlickeiser, R., Caglar, M., and Lazarian, A. (2016). Cosmic rays and MHD turbulence generation in interstellar giant molecular clouds. *ApJ* 824, 89. doi:10.3847/0004-637X/824/2/89

Schlickeiser, R. (2002). *Cosmic ray astrophysics*.

Schlickeiser, R., and Miller, J. A. (1998). Quasi-linear theory of cosmic-ray transport and acceleration: The role of oblique magnetohydrodynamic waves and transit-time damping. *ApJ* 492, 352–378. doi:10.1086/305023

Singer, H. J., Heckman, G. R., and Hirman, J. W. (2001). "Space weather forecasting: A grand challenge," in *Geophysical monograph series* (Washington DC: American Geophysical Union), 125, 23–29. doi:10.1029/GM125p0023

Sioulas, N., Isliker, H., Vlahos, L., Koumtzis, A., and Pisokas, T. (2020). Superdiffusive stochastic Fermi acceleration in space and energy. *MNRAS* 491, 3860–3869. doi:10.1093/mnras/stz3259

Strong, A. W., Moskalenko, I. V., and Ptuskin, V. S. (2007). Cosmic-ray propagation and interactions in the galaxy. *Annu. Rev. Nucl. Part. Sci.* 57, 285–327. doi:10.1146/annurev.nucl.57.090506.123011

Xu, S., and Hu, Y. (2021). Measuring magnetization with rotation measures and velocity centroids in supersonic MHD turbulence. *ApJ* 910, 88. doi:10.3847/1538-4357/abe403

Xu, S., and Lazarian, A. (2018). Resonance-broadened transit time damping of particles in MHD turbulence. *ApJ* 868, 36. doi:10.3847/1538-4357/aae840

Xu, S., and Lazarian, A. (2022). Shock acceleration with oblique and turbulent magnetic fields. *ApJ* 925, 48. doi:10.3847/1538-4357/ac3824

Xu, S., and Lazarian, A. (2020). Trapping of cosmic rays in MHD turbulence. *ApJ* 894, 63. doi:10.3847/1538-4357/ab8465

Xu, S. (2021). Mirror diffusion of cosmic rays in highly compressible turbulence near supernova remnants. *Astrophysical J.* 922. doi:10.3847/1538-4357/ac2d8f

Xu, S., and Yan, H. (2013). Cosmic-ray parallel and perpendicular transport in turbulent magnetic fields. *ApJ* 779, 140. doi:10.1088/0004-637X/779/2/140

Xu, S., Yan, H., and Lazarian, A. (2016). Damping of magnetohydrodynamic turbulence in partially ionized plasma: Implications for cosmic ray propagation. *ApJ* 826, 166. doi:10.3847/0004-637X/826/2/166

Yan, H., and Lazarian, A. (2008). Cosmic-ray propagation: Nonlinear diffusion parallel and perpendicular to mean magnetic field. *ApJ* 673, 942–953. doi:10.1086/524771

Yan, H., and Lazarian, A. (2004). Cosmic-ray scattering and streaming in compressible magnetohydrodynamic turbulence. *ApJ* 614, 757–769. doi:10.1086/423733

Yan, H., and Lazarian, A. (2003). Grain acceleration by magnetohydrodynamic turbulence: Gyroresonance mechanism. *ApJ* 592, L33–L36. doi:10.1086/377487

Yan, H., and Lazarian, A. (2002). Scattering of cosmic rays by magnetohydrodynamic interstellar turbulence. *Phys. Rev. Lett.* 89, 281102. doi:10.1103/PhysRevLett.89.281102

Yuen, K. H., Ho, K. W., Law, C. Y., Chen, A., and Lazarian, A. (2022). *Turbulent universal galactic Kolmogorov velocity cascade over 6 decades*. arXiv:2204.13760.

Zank, G. P., Adhikari, L., Hunana, P., Shiota, D., Bruno, R., and Telloni, D. (2017). Theory and transport of nearly incompressible magnetohydrodynamic turbulence. *ApJ* 835, 147. doi:10.3847/1538-4357/835/2/147

Zhang, J.-F., Liu, Q., and Lazarian, A. (2019). Tracing magnetic fields by the synergies of synchrotron emission gradients. *ApJ* 886, 63. doi:10.3847/1538-4357/ab4b4a

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 09 May 2023
DOI 10.3389/fspas.2023.1167540

Check for updates

OPEN ACCESS

EDITED BY
Lucia Kleint,
University of Geneva, Switzerland

REVIEWED BY
Sergio Javier González Manrique,
Leibniz-Institut für Sonnenphysik (KIS),
Freiburg, Germany
Feng Wang,
Guangzhou University, China

*CORRESPONDENCE
David Orozco Suárez,
orozco@iaa.es

RECEIVED 16 February 2023
ACCEPTED 05 April 2023
PUBLISHED 09 May 2023

CITATION
Orozco Suárez D, Álvarez García D, López Jiménez AC, Balaguer Jiménez M, Hernández Expósito D, Labrousse P, Bailén FJ, Bustamante Díaz I, Bailón Martínez E, Aparicio del Moral B, Morales Fernández JM, Sánchez Gómez A, Tobaruela Abarca Á, Moreno Mantas AJ, Ramos Más JL, Pérez Grande I, Piqueras Carreño J, Katsukawa Y, Kubo M, Kawabata Y, Oba T, Rodríguez Valido M, Magdaleno Castelló E and Del Toro Iniesta JC (2023), SPGCam: A specifically tailored camera for solar observations. *Front. Astron. Space Sci.* 10:1167540. doi: 10.3389/fspas.2023.1167540

COPYRIGHT
© 2023 Orozco Suárez, Álvarez García, López Jiménez, Balaguer Jiménez, Hernández Expósito, Labrousse, Bailén, Bustamante Díaz, Bailón Martínez, Aparicio del Moral, Morales Fernández, Sánchez Gómez, Tobaruela Abarca, Moreno Mantas, Ramos Más, Pérez Grande, Piqueras Carreño, Katsukawa, Kubo, Kawabata, Oba, Rodríguez Valido, Magdaleno Castelló and Del Toro Iniesta. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

SPGCam: A specifically tailored camera for solar observations

David Orozco Suárez[1,2]*, Daniel Álvarez García[1,2], Antonio C. López Jiménez[1,2], María Balaguer Jiménez[1,2], David Hernández Expósito[2,3], Pierre Labrousse[1,2], Francisco J. Bailén[1,2], Isabel Bustamante Díaz[1,2], Eduardo Bailón Martínez[1,2], Beatriz Aparicio del Moral[1,2], José M. Morales Fernández[1,2], Antonio Sánchez Gómez[1,2], Ángel Tobaruela Abarca[1,2], Antonio J. Moreno Mantas[1,2], José L. Ramos Más[1,2], Isabel Pérez Grande[2,4], Javier Piqueras Carreño[2,4], Yukio Katsukawa[5], Masahito Kubo[5], Yusuke Kawabata[5], Takayoshi Oba[5], Manuel Rodríguez Valido[6], Eduardo Magdaleno Castelló[6] and Jose Carlos Del Toro Iniesta[1,2]

[1]Solar Physics Group, Instituto de Astrofísica de Andalucía, Consejo Superior de Investigaciones Científicas (IAA-CSIC), Granada, Spain, [2]Spanish Space Solar Physics Consortium (S^{3}PC), Spain, [3]Instituto de Astrofísica de Canarias (IAC), La Laguna, Spain, [4]Instituto de Microgravedad "Ignacio da Riva", Universidad Politécnica de Madrid (IDR-UPM), Madrid, Spain, [5]National Astronomical Observatory of Japan, Mitaka, Japan, [6]Department of Industrial Engineering, University of La Laguna, San Cristóbal de La Laguna, Spain

Designing a new astronomical instrument typically challenges the available cameras on the market. In many cases, no camera can fulfill the requirements of the instrument in terms of photon budget, speed, and even interfaces with the rest of the instrument. In this situation, the only options are to either downgrade the performance of the instrument or design new cameras from scratch, provided it is possible to identify a compliant detector. The latter is the case of the SPGCams, the cameras developed to be used with the Tunable Magnetograph (TuMag) and the Sunrise Chromospheric Infrared spectroPolarimeter (SCIP) for the Sunrise III mission. SPGCams have been designed, developed, and built entirely in-house by the Solar Physics Group (SPG) at the Instituto de Astrofísica de Andalucía (IAA-CSIC). We report here on the scientific rationale and system engineering requirements set by the two instruments that drove the development, as well as on the technical details and trade-offs used to fulfill the specifications. The cameras were fully verified before the flight, and results from the assembly and verification campaign are presented as well. SPGCams share the design, although some parametric features differentiate the visible cameras (for TuMag) and the IR ones (for SCIP). Even though they were specifically developed for the Sunrise III mission, the robust and careful design makes them suitable for different applications in other astronomical instruments.

KEYWORDS

scientific camera, image sensor, CoaXPress, spectropolarimetry, SUNRISE mission

1 Introduction

A large fraction of astronomical information comes from photon counting. Many astronomical measurements are photometric and, hence, the paramount importance of detectors and cameras. Astronomers measure light; they do not measure temperatures, velocities, magnetic fields, and other physical quantities of heavenly bodies. These physical quantities are inferred with uncertainties which directly rely on the photometric errors. Proper detectors and cameras are, therefore, key elements in observational astronomy. Indeed, most of the current instrument designs start from the end, from the camera, in order to fulfill the scientific requirements and goals that scientists have set. Cameras must ensure the required photon budget, the required signal-to-noise ratio (S/N), the needed exposure time, and cadence to cope with the dynamical properties of the target astronomical (solar) features and with the seeing (in ground-based instruments) or spacecraft jitter (in the case of aerospace-borne instruments). They should be free of automatic cosmetic procedures that many available commercial cameras incorporate for everyday usage in civil applications. They may have special requirements about its behavior under harsh environmental conditions of temperature and pressure. They cannot afford other sources of noise than, typically, readout and photon noise. In addition to these photometric requirements, mechanical alignment and/or special tools to be adjusted in the final optical path may call for other special features.

Special attention must be paid to differential photometry. The combination of images taken asynchronously is a clear source of potential uncertainties. This is particularly true in the case of polarimetry, a singular way of differential photometry: the so-called Stokes parameters, I, Q, U, and V, which determine the polarization state of light, cannot be measured independently, although they have units of energy (i.e., photons). Rather, only linear combinations of all four can be recorded in single shots. Therefore, at least four independent measurements of these linear combinations are mandatory to retrieve the four Stokes parameters. It is easy to understand that any evolutionary change in the object or differential seeing or jitter conditions among the (at least) four measurements can corrupt the result, a well-known effect called polarization crosstalk among the Stokes parameters in the specific jargon (Lites, 1987; see also, e.g., Del Toro Iniesta, 2003). Intensity, Stokes I, can enter the measured Q, U, and V. Circular polarization of light, represented by Stokes V, can be measured as linear polarization, represented by Stokes Q and U, and *vice versa*. Stokes Q and U can also be intermingled.[1] All this crossed contamination naturally imposes specific requirements to the cameras in instruments like TuMag (Tunable Magnetograph; Del Toro Iniesta et al., 2023) and SCIP (Sunrise Chromospheric Infrared spectroPolarimeter; Katsukawa et al., 2023) that are spectropolarimeters.

A market screening is often insufficient to find a compliant camera. Scientific needs are usually far from commercial needs (e.g., Zhang et al., 2020; Zhu et al., 2022). This has indeed been the case when starting the conceptual design of our two instruments for the Sunrise III mission (Solanki et al., 2023). We then decided instead to screen the sensor market and, after identifying one that we could adapt to our needs, develop the cameras from scratch. Our own custom design could exploit the large commonalities between the two instruments that are indeed different, at the same time of flavoring the final versions for TuMag and SCIP. TuMag is a dual-beam, tunable imaging spectropolarimeter based on the use of liquid crystal variable retarders for the polarization modulation and a solid $LiNbO_3$ Fabry–Pérot etalon as a spectrometer. These two main elements were also fundamental to Flare Genesis (Rust, 1994) and IMaX (Imaging Magnetograph eXperiment; Martínez Pillet et al., 2011). TuMag produces quasi-monochromatic (FWHM of 6 pm) images of the aforementioned linear combinations of the Stokes parameters. The spectral information is then obtained by scanning a spectral line and its nearby continuum at discrete samples centered at given wavelengths. It operates at visible (green) wavelengths. On the other hand, SCIP is a slit-spectrograph-based spectropolarimeter which uses a rotating waveplate for the polarization modulation and a classical spectrograph for the spectral analysis. SCIP produces spectrograms where the full spectrum is recorded in one of the directions and the other is a one-dimensional coverage of the Sun. The spatial information is then completed by scanning the Sun along a perpendicular direction to the spectrograph slit. It operates in near-infrared wavelengths.

TuMag incorporates two cameras, one for each of the two (orthogonally polarized) beams coming out from the polarization analyzer. SCIP includes three cameras: two for each of the two wavelength range arms of the instrument and a third, so-called slit-jaw camera, which images the solar light that does not enter the spectrograph that serves as a context image for the observation. The SPGCam project, then, had to comply with the requirements of the two instruments, which are summarized in **Section 2** along with the selected sensor. A discussion on the behavior of sensors and the various sources of noise follows in **Section 3**. The results of that analysis in the case of our sensor help decide how to fill the full well of the pixels according to our requirements of the signal-to-noise ratio, which are directly associated with our polarimetric sensitivity. Then, a description of the electronic hardware follows in **Section 4**. The camera mechanics is described in **Section 5**. **Section 6** summarizes the thermo-vacuum characterization of the cameras, with all the tests performed before integration into the instruments.

2 Scientific and system engineering requirements

TuMag is a dual-beam, tunable imaging spectropolarimeter that uses two cameras and should image a solar field of view (FoV) of $63'' \times 63''$. SCIP is a slit-spectrograph-based spectropolarimeter that images the (~ 8.5 nm wide) spectrum of a one-dimensional, $58''$-long region of the Sun. Its slit-jaw camera has the same FoV

1 Correction from this contamination can be carried out in practice by using point-to-point correlation between the two related Stokes parameter images. Typically, continuum wavelength samples are used since they are supposed to be unpolarized. In any case, however, the less crosstalk the raw images have, the better the results.

as that of TuMag.[2] The camera electronics should host, drive, and control the temperature of the image sensor. A shared design of a camera for the two instruments thus asks for a common sensor that fulfills the requirements for both with the same electronic components and, perhaps, some parametric differences. Similar drivers apply for the mechanical design. The envelope, materials, coatings, etc., are determined by the most demanding instrument and should almost be the same for the two instruments. A specially tailored window protects the sensor from dust, humidity, and other volatile materials while preventing any extra aberrations which might degrade the final optical quality. The window may be removed for flight depending on the needs of the specific instrument. The scientific and system engineering requirements are summarized in **Table 1**.

After a review of the sensor market, we found that the rolling shutter, backside-illuminated GPIXEL SENSE GSENSE400BSI[3] 2 k × 2 k pixel detector is compliant for both the TuMag and SCIP requirements. Thus, the SPGCams have been designed to use this detector. While physically identical, the two cameras need a few different parameters in the driving electronics. The main driver for selecting this sensor was twofold: first, the pixel lateral size of 11 μm is fairly close to the foreseen plate scale in the initial development phase of both instruments. Second, the BSI sensor has very good visible and infrared sensitivity without the need of cooling it down. Usually, infrared sensors need to be cooled down to boost their sensitivity; however, this sensor is among the most sensitive sensors at room temperature. Operation at room temperature is very important for a stratospheric balloon mission like Sunrise III.

3 Noise and its sources

Camera measurements can be broken down into three clearly distinct stages, namely, the conversion from photons to electrons, the conversion from electrons to voltage, and, finally, the digitization of such a voltage. Each stage contributes to the measurement with noise, whose origin has to be clearly understood in order to correct or mitigate its effects.

As a consequence of the photo-electric effect being a Poissonian process, some of the noises have a statistical dependence on the signal itself. There are, however, sources that are independent of the signal so that they are always present. Among those sources independent of the signal, there are some that depend on environmental factors like temperature or pressure. The characterization of noise in a camera is, thus, a significant challenge. This section is devoted to summarizing all sources of noise and explain ways for calculating them in the laboratory. Many of the concepts may already be familiar to the reader, but we offer a comprehensive review for non-expert users, e.g., astronomers. **Table 2** summarizes all the noise variables and their meaning.

Let $n_{\rm ph}$ be the number of photons arriving at a given pixel of the detector. The photon emission process is random in nature, and it produces so-called photon shot noise (or simply photon noise), whose variance is equal to $n_{\rm ph}$:

$$\sigma^2_{\rm ph} = n_{\rm ph}. \tag{1}$$

The detected photons are converted to a current of n_{e^-} electrons with a quantum efficiency, Q: $n_{e^-} = Qn_{\rm ph}$. Then, the signal

$$S = n_{e^-} = Qn_{\rm ph}. \tag{2}$$

Since the production of an electric current from individual electrons is a Poissonian process associated with the quantum nature of charges, in much the same way as photon emission is, the corresponding noise is also a shot noise and hence equal to the square root of the signal. In the astronomical jargon, this electron noise continues being called photon noise:

$$N_{\rm ph} = \sigma_{e^-} = \sqrt{Qn_{\rm ph}}. \tag{3}$$

Not every pixel responds to photons with exactly the same efficiency under uniform illumination. Indeed, the pixel-to-pixel sensitivity differences result in a photon-response non-uniformity[4], with an associated noise proportional to the number of photons:[5]

$$N_{\rm prnu} = f_{\rm prnu} n_{\rm ph}. \tag{4}$$

This response non-uniformity has the same driver as what astronomers know as flat field: The individual pixel sensitivity introduces modifications in the pixel signal and generates an individual pixel noise. Indeed, the two-dimensional distribution across the detector of the $f_{\rm prnu}$ factor equals the flat-field response to a uniform illumination of the camera.

Since $N_{\rm prnu}$ is signal-dependent and indeed proportional to $n_{\rm ph}$, it can easily dominate over the photon shot noise, which is not a recommended circumstance in any kind of photometry. When the ratio equals 1, photon-response non-uniformity is the more important noise source.

$$\frac{N_{\rm prnu}}{N_{\rm ph}} = f_{\rm prnu}\sqrt{\frac{n_{\rm ph}}{Q}}$$

There still remains a noise in this first process of conversion of photons into electrons (or current) that is not dependent on the signal. It can be considered a toll for reading the camera and is called the read noise (or readout noise), $N_{\rm r}$.

Each pixel can produce, however, a current without the presence of light. It is known as dark current, which generates 1) a dark-current shot noise, due to a spontaneous random generation of $n_{\rm dc}$ electrons, and 2) a fixed-pattern noise, resulting from pixel-to-pixel differences on manufacturing materials and alike.

$$N_{\rm dc} = \sqrt{n_{\rm dc}}; \tag{5}$$

$$N_{\rm dc,fpn} = f_{\rm dc,fpn} n_{\rm dc}. \tag{6}$$

2 A slit-jaw camera is devoted to imaging the focal plane of the instrument, exactly where the spectrograph slit is. This way, all the light that does not enter the spectrograph is redirected to the slit-jaw camera in order to get a context image of the solar scene.

3 https://www.gpixel.com/products/area-scan-en/gsense/gsense400bsi-11-µm-4mp-rolling-shutter-image-sensor/

4 Its appearance is largely associated with the sensor manufacturing process itself and is time-invariant.

5 Note that $f_{\rm prnu}$ should have Q included as an internal factor for dimensional reasons. ($N_{\rm prnu}$ should also be expressed in number of electrons.) Eq. **4** is written like this to follow customary nomenclature.

TABLE 1 Camera scientific and system engineering requirements.

Name	Symbol	Value TuMag/SCIP	Units (type)[a]
Sensor's quantum efficiency at [515, 527]/854 nm	Q	≥0.8/≥0.5	
Pixel full well		$\geq 6 \cdot 10^4$	e^-
Photon flux budget	F	$\geq 3 \cdot 10^4$	e^-
Photon-response non-uniformity	f_{prnu}	<0.02	
Digital depth		12	bit
Pixel area		11×11	μm^2
Effective collecting area		2048×2048	pixels
Sensor's built-in processing		No	
Configurable readout speed		≤48	frames s^{-1}
Configurable exposure time		Yes	
Configurable pixel readout gain		Yes	
Readout noise	N_{r}	<10	e^-
Fixed-pattern noise[b]	$N_{\mathrm{dc,fpn}}$	<100	e^-
Dark current at 20°C	N_{dc}	<50	e^-
Configurable ROI area[c]		Yes	
Externally triggered image capture		Yes	CoaXPress[d]
Maximum camera envelope		$100 \times 80 \times 73$	mm^3
Optical entrance window		Yes	Fixed/removable
Housing material		Black anodized	Al 6082-TL
Thermal interface		Yes	Cold finger
Overall camera weight		<1	kg
Housing electrical insulation[e]		>1	MΩ
Sensor's surface roll[f]		<1°	
Sensor's surface tip/tilt[f]		<1°	
Adjustable sensor's tip/tilt		Yes	
Camera's FPGA programming port		Yes	
Camera's FPGA debugging port		Yes	
CoaXPress coaxial connector		Yes	DIN 1.0/2.3 jack
Thermal control power connector		Yes	
Near to vacuum operations		Yes	
Camera operational temperature[f]		[−20, 40]	°C
Sensor's configurable thermal control		Yes	
Default sensor's operational temperature		20.0 ± 0.5	°C
Power consumption		<6.9	W

[a]Units are per pixel where necessary.
[b]For the specified photon flux.
[c]ROI stands for region of interest.
[d]Digital interface for high-speed data transmission.
[e]From camera electronics.
[f]Reference point at the housing.

These two noises associated with dark current are temperature-dependent. Therefore, characterization of the camera should be carried out at operating temperatures. Moreover, n_{dc} has an implicit dependence on time: it varies linearly with integration time. Hence, the longer the exposure time, the more the dark current generated.

The conversion of charge to voltage, characterized by a gain g, implies that we have to multiply both S and all the aforementioned noises by g. (We will keep, however, the same symbol for the variables, as it is simply a change in units from e^- to V.) This is not a perfect process either. Gain fluctuations and non-linearities

TABLE 2 Noise-related variables and their definition.

Variable	Definition
n_{ph}	Number of photons
n_{e^-}	Number of electrons
n_{dc}	Number of dark-current electrons
σ^2_{ph}	Variance of the photon distribution
N_{ph}	Photon noise
N_{prnu}	Photon-response non-uniformity noise
f_{prnu}	Photon-response non-uniformity factor
N_r	Readout noise
N_{dc}	Dark-current shot noise
$N_{dc,fpn}$	(Dark-current) Fixed-pattern noise
$f_{dc,fpn}$	(Dark-current) Fixed-pattern factor
N_{ofp6n}	(Dark-current) Offset fixed pattern noise
N_q	Quantization noise

appear in the various pixel and/or column amplifiers, which produce so-called offset fixed-pattern noise, N_{ofpn}, which is one-dimensional in itself as it varies from column to column in the detector. Typically, a few specifically devised, physically darkened rows of pixels are used for the determination of, at least, the contribution of the column amplifiers to N_{ofpn}. Indeed, our detector has 2,048 × 2,050 pixels. Hence, two rows are used to measure N_{ofpn}.

Finally, the digitization of voltage to data numbers (DNs) of n_{bit} bits introduces a quantization noise, given by

$$N_q = \frac{\Delta}{12}, \tag{7}$$

where the quantization interval $\Delta = D_r/2^{n_{bit}}$ with D_r being the dynamic range (the pixel full well expressed in DN).

In summary, besides possible non-linearities that can be hardly modeled, we have a time-invariant noise or fixed-pattern noise (coming from photon-response non-uniformity and offset fixed pattern), a time-dependent noise (coming from photon shot noise, dark-current fixed pattern, and dark current), a read noise which is independent of the signal, and a quantization noise. We can see the various contributions in yet another way, namely, assuming that all sources of noise are uncorrelated,

$$N^2 = \left[N_r^2 + N_{ofpn}^2 + \frac{\Delta^2}{12^2}\right] + \left(1 + f_{dc,fpn}^2 S_{dc}\right) S_{dc} + \left(1 + f_{prnu}^2 S\right) S, \tag{8}$$

that is, an independent term (in brackets), a quadratic term on the dark-current signal, $S_{dc} = gn_{dc}$, and a quadratic term on the signal, S. Still another physical interpretation is possible: noise is produced by the generation of electrons from photons (with an offset given by the readout N_r), from the spontaneous generation of electrons by the detector (with an offset given by N_{ofpn}), and by quantization.

Since in practice dark current (with its offset fixed-pattern noise) can be calculated separately and subtracted from the signal, the effective noise of a measurement can be given by

$$N = \sqrt{\left[N_r^2 + \Delta^2/12^2\right] + f_{dc,fpn}^2 S_{dc}^2 + \left(1 + f_{prnu}^2 S\right) S}. \tag{9}$$

Note that the dark-current, fixed-pattern noise remains after subtracting S_{dc} along with N_{ofpn}, since it is proportional to S_{dc}.

From Eqs 2, 3, written in V units, the gain can be obtained as $g = N_{ph}^2/S$. If we take two subsequent images with equal, uniform (flat) illumination, the variance of their difference, σ^2_{diff}, can be considered an approximation of N_{ph}^2. Indeed, $N_{ph}^2 \simeq \sqrt{2}\,\sigma^2_{diff}$. S can be taken as the average of the images. For a statistically significant estimate, one can take a series of equally illuminated flat images and perform a linear least squares fit of N_{ph}^2 vs. S. The slope gives the gain, and the ordinate in the origin is a good estimate of N_r. A measure of the dispersion of the fits is given by so-called linearity,

$$L = \frac{\max\left(S - S_{fit}\right) + \min\left(|S - S_{fit}|\right)}{\max\left(S\right)} \times 100. \tag{10}$$

The dynamic range is, by definition, the ratio of the electron full well of the sensor to the read noise. Since the full well is an intrinsic parameter of the sensor, the calculation of D_r is straightforward. From it, the quantization noise follows from Eq. 7.

The photon-response non-uniformity factor is calculated from the ratio of the rms to the average of an image which has been uniformly illuminated up to half its full well, once dark current is subtracted, directly from Eq. 4.

$f_{dc,fpn}$ can be calculated similarly as g but with a series of dark images, each with a different exposure time.

When the camera is used in imaging photometry, the flat-fielding process helps get rid of all those noises except the photon noise: an average of uniformly illuminated images (subtracted from dark current and normalized to their means), *the flat-field image*, is used to divide all the individual measurement images, which thus provides results corrected from pixel-to-pixel gain differences. The larger the number of flat images we use to obtain the flat-field image, the closer its photon noise is to zero, and then the better. When cameras are used for slit spectroscopy, corrections are less simple because of the large changes in the signal along the wavelength direction. In this case, an extra step is necessary, which consists in removing the spectral line information after being useful for flat-fielding the images. **Figure 1** illustrates the differences between these two types of images for those readers who are not familiar with solar astronomy. On the left, a regular (polarized) image is shown. The image corresponds to the spatial variation of the solar scene at a quasi-monochromatic wavelength of the visible solar spectrum. In this case, this image was taken during sunlight calibrations of the TuMag instrument, and, although it does not show solar structure due to the atmospheric conditions under which it was taken, we have introduced a grid pattern for the purpose of performing alignment tests between the cameras. On the right, a spectrogram is shown: the Y-axis corresponds to the (one-dimensional) solar scene; the X-axis corresponds to wavelength. Conspicuous spectral lines can be observed. The left and right spectrograms correspond to two distinct, orthogonally polarized beams.

Figure 2 presents so-called photon transfer curves (PTCs) generated by the GSENSE400BSI sensor integrated into a prototype

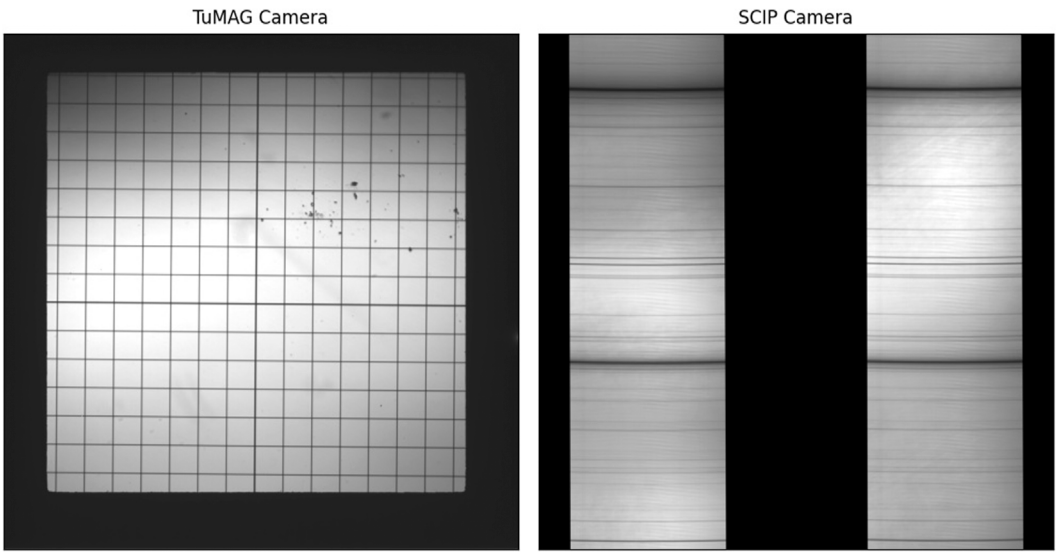


FIGURE 1
Left: polarized image coming out from TuMag's solar light tests. Right: spectrogram coming out from the SCIP's solar light tests. In both images, black areas correspond to masked regions within the instruments.

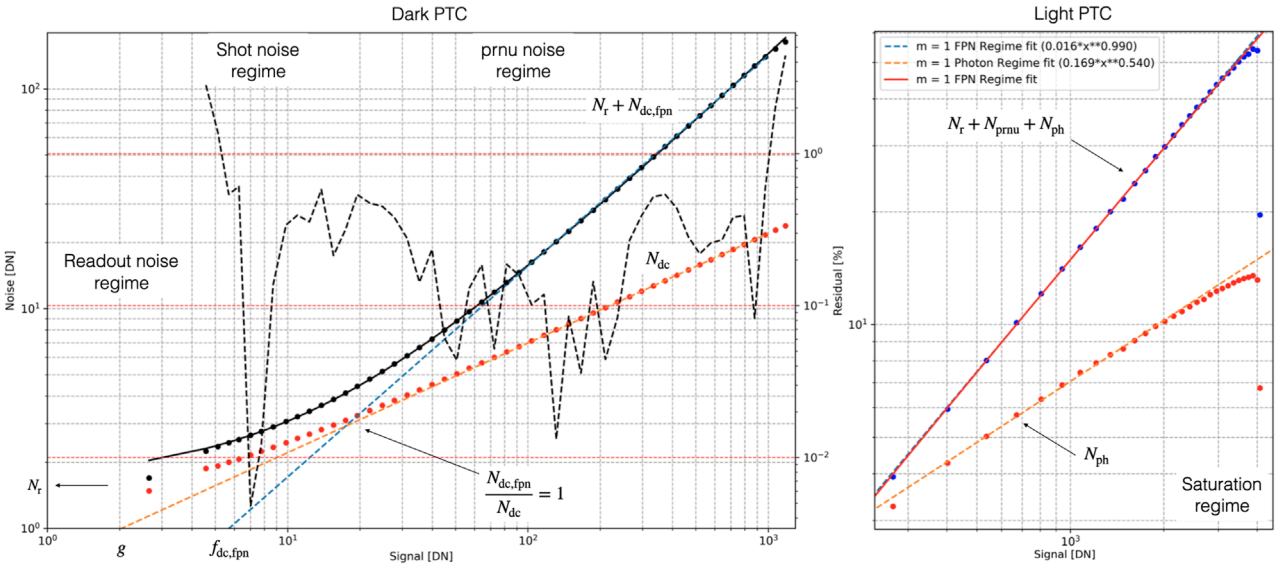


FIGURE 2
Variation in noise vs. signal, in a log–log plot, without light (dark PTC; left) and with external illumination (light PTC; right).

SPGCam under ambient conditions at 25°C. The left panel represents the PTC curve in the absence of light. It has been built by varying the exposure time, which gives rise to larger dark currents. Hence, in this dark PTC, only the two first terms of Eq. 8 contribute to the measured signal and noise. The right panel, on the other hand, has been obtained by illuminating uniformly the sensor and, again, changing the exposure time. Hence, it contains the contribution of all three terms in Eq. 8. Light PTCs are usually obtained by varying the light source, but in such a case, special care must be taken to ensure the stability of the light source, especially for experimental repeatability purposes. In both panels, dots indicate measurements, and dashed lines stand for slope 1/2 (orange) and slope 1 (blue) behaviors. In both PTCs, the black dots represent the standard deviation calculated across the image, while the red dots show the standard deviation of the frame difference of two identical frames with the same exposure time. The solid line stands for a linear fit to

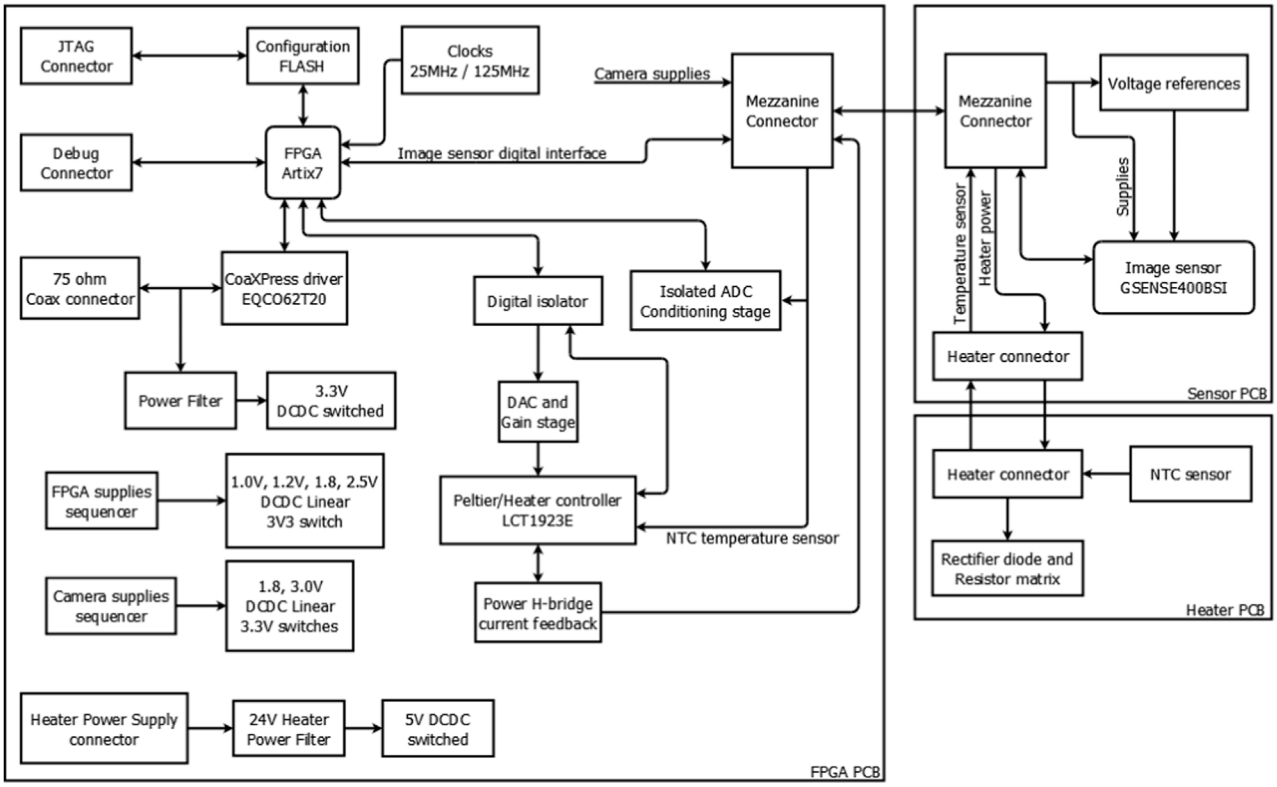


FIGURE 3
Block diagram of three PCBs in the camera.

the black dots, and the black dashed line in the dark PTC represents the percentage difference between measurements and the fit (scale on the right). From any of the PTCs, $-\log g$ (in DN/e$^-$) and $\log f_{\mathrm{prnu}}$ (or $\log f_{\mathrm{dc,fpn}}$; non-dimensional) are directly given by the crossing points of the fits with the X-axis. The extrapolation of the black fit to $S = 0$ gives N_{r}. Four horizontal regimes are labeled, namely, a readout noise regime (where N_{r} dominates), a shot noise regime (where N_{ph} or N_{dc} is the most important source), a photon-response non-uniformity regime (where N_{prnu} and $N_{\mathrm{dc,fpn}}$ are the biggest sources), and the saturation regime where the signal saturates because it is close to the maximum of the sensor dynamic range. It is important to remark that, in both PTCs, the takeover between N_{dc} and $N_{\mathrm{dc,fpn}}$ occurs very soon, when the signal is still not that large. This takeover shifts to the right, i.e., $N_{\mathrm{dc,fpn}}$ becomes less important as soon as the sensor temperature decreases. Ironically, and contrary to what other astronomers do, the SPGCams work at room temperature in the Sunrise III mission. For this reason, the cut-off point is located right at the beginning.

4 Camera electronic hardware

4.1 Overall view

The camera electronics consist of three printed circuit boards (PCBs), namely, one hosting the sensor, a second hosting the control field-programmable gate array (FPGA), and a third hosting an array of resistors which operate as a heater for temperature control. A block diagram of the three PCBs is shown in Figure 3.

The Sensor-PCB (upper right of Figure 3) contains the image sensor. It includes the sensor decoupling capacitors, its voltage reference generator, and the heater connector for facilitating the camera assembly. The Sensor-PCB is designed with an opening area for accessing the bottom of the image sensor with a cold finger. The selected GSENSE400BSI sensor is scientific-CMOS (sCMOS), compatible with vacuum environments, low power, low readout noise, high dynamic range, and high quantum efficiency peak at 560 nm (95%), among other features.

The FPGA-PCB (left of Figure 3) includes the external interfaces of the camera, the power management, the FPGA itself, and its thermal controller. There are four external interfaces: CoaXPress, the thermal controller power supply, a debug port, and the programming port (JTAG). The last two interfaces cannot be accessed by the user as they are intended for development purposes only. The system engineering choice of a CoaXPress interface has a great impact on this project. From an electrical point of view, it is simpler than other alternatives such as USB3 or Camera Link, since CoaXPress consists of a single 75 Ω coaxial cable to power the camera and downlink and uplink the data at the same time. This provides a smart mechanical and physical layer, which makes routing the harness within the optical bench easier, thanks to its small bending radius. It also allows long harness distances, reduces risks during harness manufacturing, and simplifies the PCB design

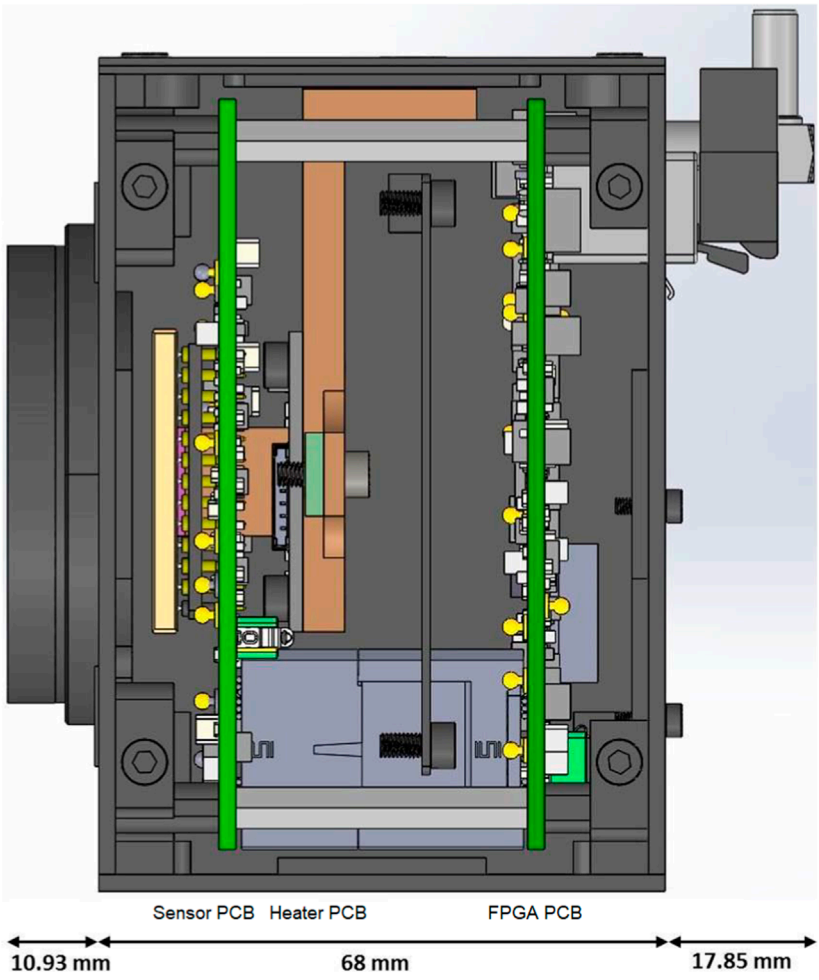


FIGURE 4
Cross section of the TuMag camera (including the entrance window and connectors) showing the position of the three PCBs. On the left is the sensor-PCB in which the image sensor appears plugged. In the center, the heater-PCB appears attached to a cold finger (orange piece). On the right is the FPGA-PCB, with the CoaXPress and heater supply connectors of the harness already connected. The mezzanine connector (Hirose IT5 family), as well as a shielding screen between the sensor-PCB and FPGA-PCB, is also shown.

as it saves space by considerably reducing the number of lines involved.

The FPGA-PCB and sensor-PCB are in a mezzanine disposition, as shown in Figure 4. There are two reasons for having such a configuration. First, due to camera size restrictions, not all components fit on a single PCB. Second, we should minimize the thermal coupling between the image sensor and the rest of the electronics.

The heater-PCB (lower right of Figure 3) is a metallic substrate PCB attached to the cold finger for stabilizing its temperature. The temperature sensor is a pigtailed, 10 kΩ negative temperature coefficient thermistor. It is placed on the cold finger at the base of the image sensor, and its terminals are soldered to the heater-PCB. The heater-PCB is then located on the cold finger touching the image sensor but isolated from the other two PCBs. The heater-PCB also includes a rectifier diode. The rationale for that diode is that the thermal controller, included in the sensor-PCB, is designed to drive a Peltier module. When the thermal module needs to cool down the sensor, it reverses the current sense, but the rectifier diode plays the role of canceling it.

All PCBs are painted black (solder mask) to minimize light reflections inside the camera.

4.2 FPGA printed circuit board

The purpose of this PCB is to include most of the camera electronics. It is a 10-layer FR4 PCB with a very compact size (90 mm × 70 mm; 2.11 mm high), as shown in Figure 5. When required, to ensure the signal integrity, the tracks are routed in a controlled impedance way and on dedicated layers. This PCB includes four user interface connectors and one mezzanine connector:

- Right DIN 1.0/2.3 75 Ω for CoaXPress. The choice of this type of connector provides compact size and matting simplicity as compared to a BNC type.

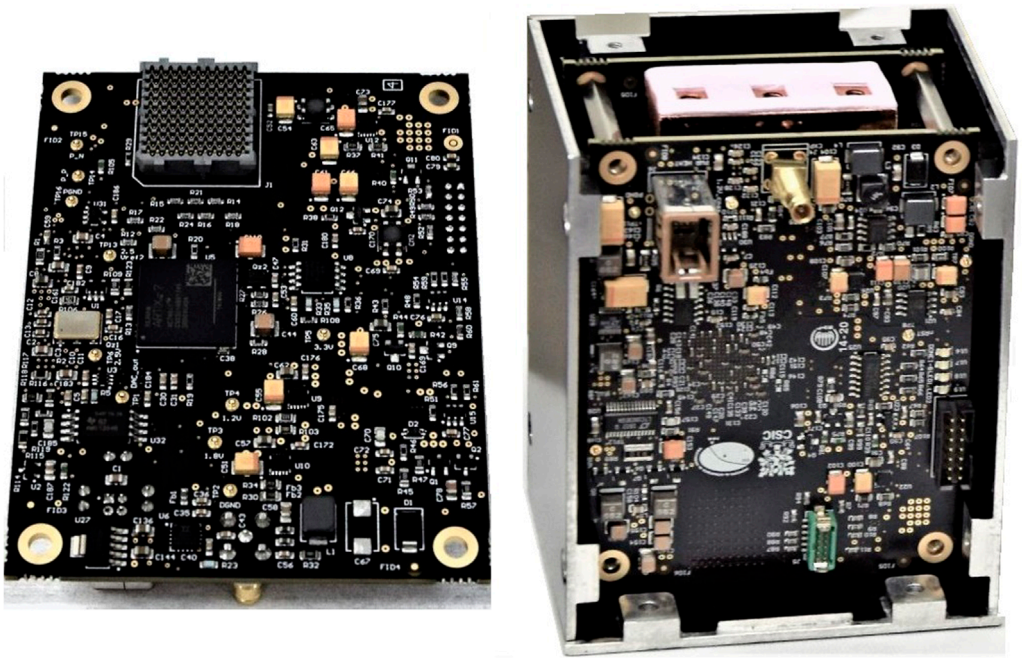


FIGURE 5
Outlook of FPGA-PCB. On the left is the top view of the PCB, highlighting the FPGA (in the middle) and the mezzanine receptacle at the top left. On the right, the camera is partially assembled, showing the bottomside of the PCB with the four user interface connectors.

- Four wires, Hirose GT17 family, for powering (24 V) the camera heater controller. This supply is isolated from the CoaXPress supply. That choice provides robustness, shield construction, harness retention, and different counterpart orientation capabilities.
- Standard Xilinx JTAG (14-wire) connector to program the flash memory that stores the FPGA firmware. This connector is not accessible in normal operation (by removing a protection cover, the user can update the camera firmware).
- Twelve-wire Harwin Gecko family connector, for FPGA debugging. This connector is only accessible when the camera is disassembled. It is a compact, polarized, surface mount device connector with a retainer.
- One hundred positions Hirose IT5 receptacle. This is one part of the three-piece mezzanine system (consisting of two polarized receptacles and an interposer that modulates the stacking height, set at 35 mm). The IT5 connector system is capable of transmitting, in a very compact size, single-ended signals, high-speed differential signals (with 100 Ω controlled impedance), and power lines between two boards. The interface with the sensor-PCB consists of 10 differential lines, 42 control lines, and 12 power lines.

The camera-FPGA and image sensor are fed by the CoaXPress carrier voltage (24 V). To perform this, the CoaXPress voltage is first decoupled for isolating the signal data from the 24 V source, following the recommendation of the CoaXPress driver manufacturer (Microchip EQCO62T20.3). It then passes through an additional L-C filter for extra noise attenuation to supply switching DC/DC of 2.15 MHz that gives 3.3 V. The purpose of that additional filter is to accommodate the solution proposed by Microchip[6] on the PCB in order to avoid situations where the crosstalk is still high on the uplink channel after the first filter. We have followed a conservative design in order to minimize risks and verified that transmission errors do not appear with the current design. The FPGA and image sensor supplies are generated sequentially from this 3.3 V, following the Xilinx and GPIXEL specifications. The chosen sequencer is the Texas Instruments LM3880MF, one for each device. Since the FPGA and image sensor require power supplies with relatively low current consumption, they have been designed using DC/DC low dropout (LDO) regulators. Furthermore, in the case of the image sensor, this is a safe design choice to prevent voltage fluctuations (to the extent that the LDO supplies are cleaner than the switched supplies) in the power supplies related to the image acquisition stage from degrading quality or introducing artifacts into the image. This is a safe option for the FPGA high-speed interface-related supplies (VMGTAVTT and VMGTAVCC).

The chosen FPGA is a Xilinx XC7A50T-2CSG325I (belonging to the Artix7 family). This version includes so-called Gigabit Transceiver Port (GTP) transceiver, whose speed rates allow the FPGA to communicate with CoaXPress devices. Among other properties, this FPGA has the advantage of being reconfigurable, providing very good performance vs. power, and has a small package, features a MicroBlaze processor, and includes Vivado, a highly optimized development suite. The tasks of the FPGA are:

6 https://ww1.microchip.com/downloads/aemDocuments/documents/OTH/ProductDocuments/DataSheets/60001301B.pdf

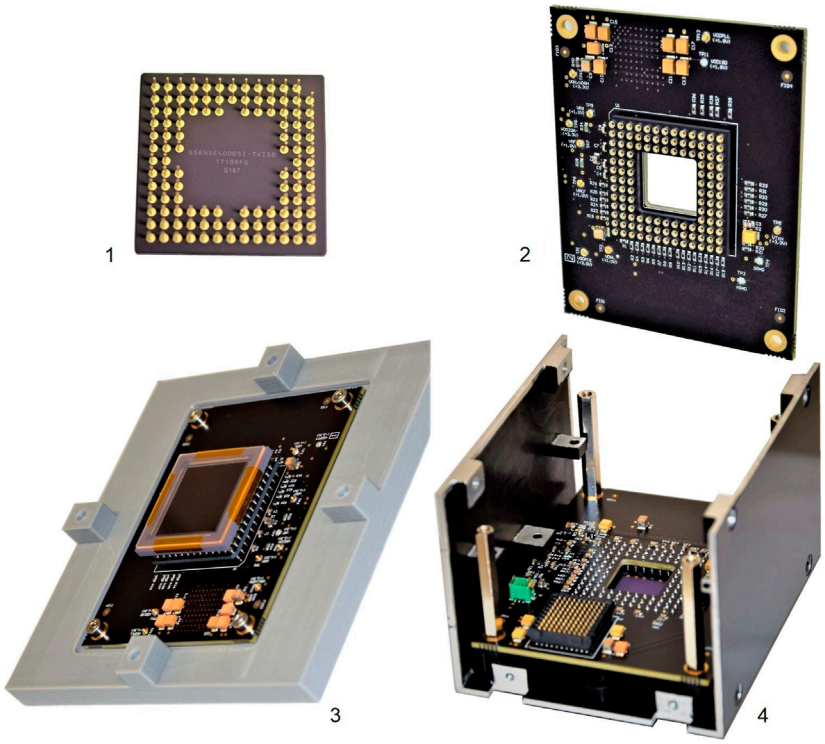


FIGURE 6
Outlook of sensor-PCB: 1) GPIXEL GSENSE400BSI bottomside; 2) top view of the sensor-PCB without the image sensor mounted. The 120-pin PGA socket and the PCB window are shown; 3) top view of the sensor-PCB with the GSENSE400BSI already inserted (with its protective glass cover fixed with Kapton). The sensor-PCB mounted on the image sensor assembling tool; 4) camera partially assembled, PCB bottomside view. It can be distinguished by the GSENSE400BSI mounted, the mezzanine connector, and the heater connector (green).

- Configuring and reading the image sensor. The image sensor includes several interfaces that are managed by the FPGA: a serial peripheral interface (SPI) to configure its operational status, several lines (control, decoding, and timing) dedicated to handle the image acquisition, two reference input clocks (high and low speed), and, finally, nine LVDS output channels involved in image reading.
- Communicating with the instrument data processing unit (DPU) through the CoaXPress protocol. Downlink transmission (3.125 Gbps) is achieved by capacitive coupling between the GTP transceiver and the CoaXPress driver. The uplink transmission (20.83 Mbps) is a low-voltage transistor–transistor logic (LVTTL) line between the FPGA and the CoaXPress driver.
- Managing the thermal controller. To carry out this, the FPGA shall enable the controller (analog LTC1923E devices) and drive a digital-to-analog converter (Texas Instruments ADC8311; three-wire SPI protocol) to configure the temperature setpoint. Additionally, the FPGA recognizes the status of the thermal controller by reading two signals (error and heating/cooling status).
- Acquiring telemetry parameters from the camera to transfer them to the instrument DPU. The telemetry parameters include the FPGA internal supplies, the CoaXPress carrier voltage, the temperatures of the FPGA (built-in sensor), the cold finger, and the image sensor (built-in sensor).
- Including a universal asynchronous receiver–transmitter and input/output ports for firmware debugging purposes.
- Enabling/disabling the image sensor power stage.

The camera hardware includes a “master-serial” design to interface the FPGA with the SPI flash memory in which the FPGA firmware is stored. With that master-serial configuration, the user can reprogram the flash memory by plugging a Xilinx-compatible programmer into the JTAG connector. The firmware is autonomously loaded into the FPGA from the flash memory right after the camera is powered on (when the 24 V appears in the CoaXPress cable). There is an external power-on-reset circuit that triggers the FPGA just as its configuration is over.

Two clocks are provided to the FPGA: a 25 MHz, low-voltage complementary metal oxide semiconductor as the main system clock and a capacitively coupled, 125 MHz, low-voltage differential signaling clock reference to perform the CoaXPress protocol, related to the GTP transceiver.

The FPGA voltage supplies of the input/output banks and their pin assignment are chosen in the hardware to allow all the previous interface protocols. The heater power and electronics are separated from the rest of the camera to minimize noise couplings that could degrade the digital interface communications (CoaXPress, camera-FPGA data transfer) and acquired image quality. Consequently, the FPGA drives and controls the heater electronics through devices

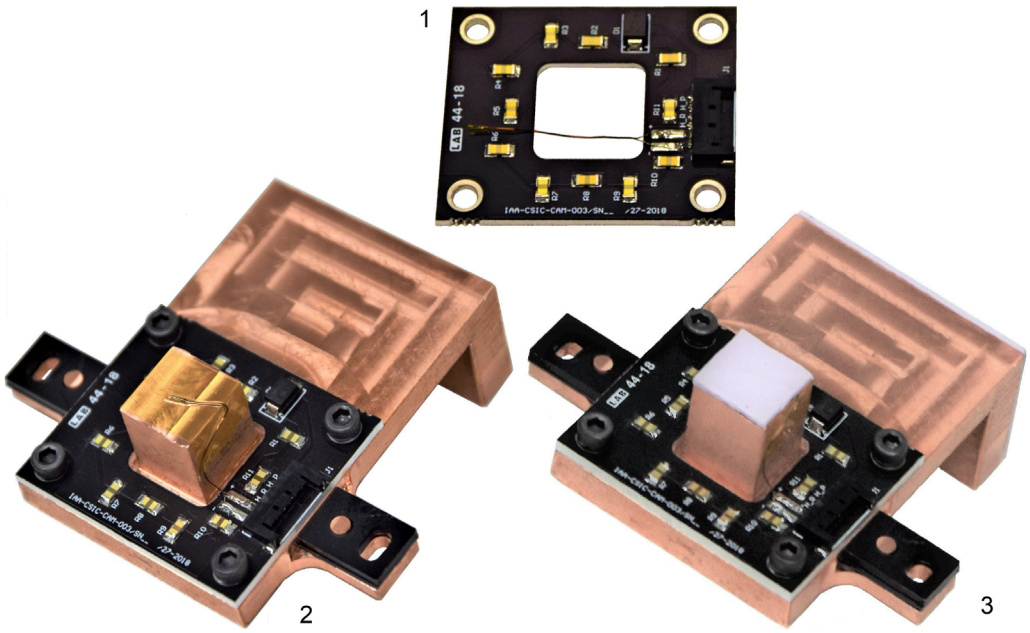


FIGURE 7
Heater-PCB mounted. 1) Before assembling on the cold finger. 2) Screwed into the cold finger (machined on a copper block). The thermocouple is fixed in place with Kapton (before the thermal pads were added). Apart from that, a piece of black plastic (Delrin) is included at the cold finger attachment points to prevent thermal coupling between the cold finger and the rest of the housing. 3) The thermal pads are placed on one side at the bottom of the image sensor and on the other at the top cover of the camera housing (thermally isolated from the rest of the housing).

such as a digital input/output isolator and an isolated analog-to-digital converter.

4.3 Sensor printed circuit board

Like the FPGA-PCB, this is an FR4, 10-layer PCB with a very compact size (90 mm × 70 mm; 2.12 mm high), as shown in Figure 6. It includes the image sensor socket, decoupling capacitors, local voltage references, and a connector for heater control. The reason for having as few circuits as possible on that PCB is to avoid transferring heat to the image sensor. The tracks are routed in the controlled impedance format and on dedicated layers to ensure signal integrity.

The sensor-PCB interfaces with the FPGA-PCB via the 100-position, Hirose IT5 receptacle, as explained in the previous section, and with the heater-PCB through a 6-wire Harwin Gecko family connector.

The image sensor requires four 1.0 V voltage references, each generated by buffering a common dedicated voltage reference. Having the references on different branches prevents coupling and interference between them. This circuitry is placed on the sensor-PCB because the cleanliness of these references is improved by being closer to the sensor and the power dissipated is negligible.

The image sensor is plugged into a socket soldered on the PCB, which prevents GSENSE400BSI from getting dirty and stressed as if it were soldered. A tool has been designed to insert the sensor into the socket uniformly up to a known distance so as not to damage the sensor during that process. The sensor has a 115-pin PGA package, where all pins are located on its perimeter (panel 1 in Figure 6). Therefore, the sensor-PCB includes a window at the bottom of the sensor (panel 2 in Figure 6) to allow the cold finger/heater to come into contact with the sensor in order to regulate its temperature.

4.4 Heater printed circuit board and cold finger assembly

This is an even more compact PCB (43 mm × 36 mm; 1.6 mm high), painted black, with a single layer on an aluminum substrate to facilitate a homogeneous temperature distribution (panel 1 in Figure 7). It includes an array of power resistors, a rectifier diode, the pads for soldering a remote thermistor, and the connector.

The heater-PCB is attached with four M3 screws (0.5 Nm of torque) to a cold finger that touches the base of the image sensor (panel 2 in Figure 7). The heater-PCB has a window through which the part of the cold finger that touches the image sensor passes. So the resistor array covers all four sides of the PCB. In this way, the temperature reaches the base of the image sensor evenly.

The heater-PCB connector belongs to the Molex Pico-Lock wire-to-board, surface-mounted device family. This choice provides a secure matting (pin retention and connection lock) with a very low profile and compact result so as not to interfere with other parts of the camera.

Thermal control feedback is provided by a negative temperature coefficient (NTC) thermistor probe. It is placed on the cold finger,

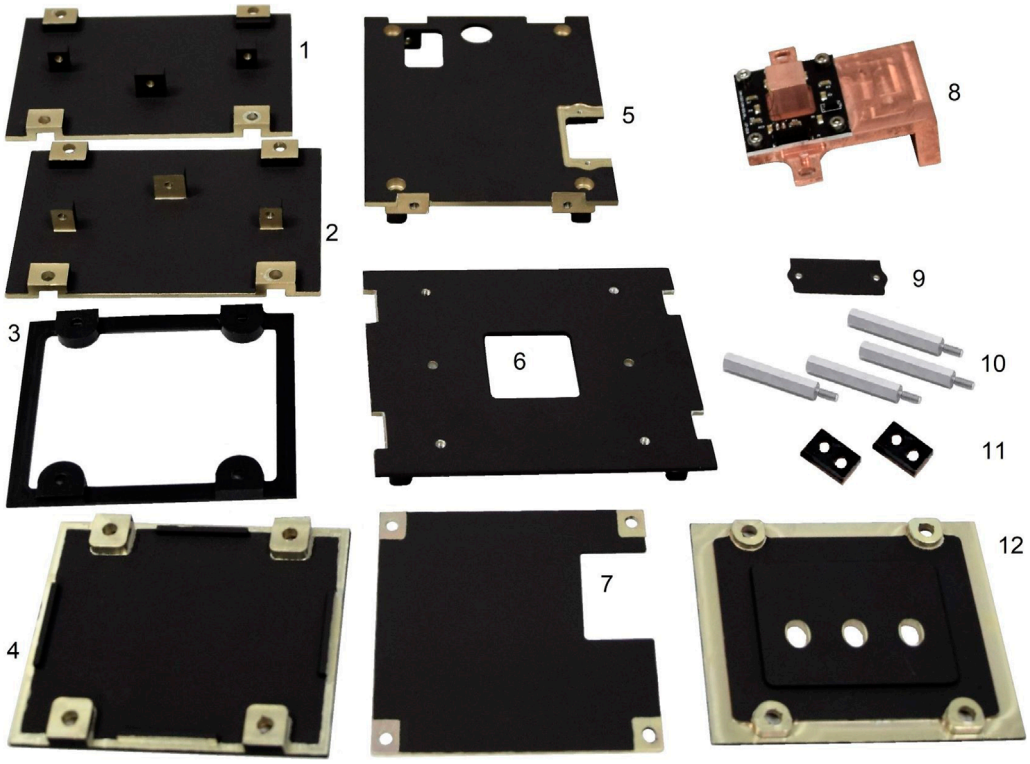


FIGURE 8
Mechanical parts of the camera. 1, 2) Housing side covers; 3) top cover thermal insulator; 4) bottom cover; 5) back cover; 6) SCIP front cover; 7) EM protection layer; 8) cold finger; 9) JTAG connector cover; 10) 35 mm M3 male/female spacers; 11) cold finger thermal insulator parts; and 12) top cover.

just below the image sensor, and affixed there with a small piece of Kapton (panel 2 in Figure 7). Hence, a miniature probe (Tyco GA10K3MCD1) has been chosen so that its size does not interfere with the mechanics. That temperature sensor includes a pigtail whose terminals are soldered to the sensor-PCB. Between the NTC probe (cold finger) and the bottom of the image sensor, there is a soft thermal pad (Bergquist GapPad HC5.0; panel 3 in Figure 7) to absorb mechanical tolerances and provide a good thermal contact. The cold finger also includes a thermal pad of the same type that interfaces with the top cover of the housing, where an external radiator is screwed on. Initially, the camera was designed so that a Peltier module stabilized the temperature of the image sensor. It was later found that a cold finger, in conjunction with a passive heater, was sufficient to stabilize the temperature and meet the scientific requirements. However, to make the camera hardware design versatile and to have the future possibility of including a Peltier module, this initial controller was kept. The Peltier module is controlled by modulating the amplitude of the current. The sense of that current causes the module to cool down or heat up. As the camera heater is driven by the Peltier controller, in case it needs to cool down the sensor, the heater has to include a rectifier to prevent the current sense shift from continuing to heat up. By adding that rectifier, the heater does not dissipate power when the controller needs to cool down, so the cooling happens passively through the cold finger. The system has been designed so that the heater allocates a maximum of 5 W and achieves a thermal stability of ±0.1°C.

5 Camera mechanics

5.1 Individual parts

The mechanical parts of the camera are divided into nine pieces of aluminum (6982-T6), one piece of copper (oxygen-free Cu-HPC CW021A), and three pieces of Delrin (plastic), as shown in Figure 8.

All aluminum parts are coated with black anodizing treatment. The coating applied is S/MIL A 8625F TIPO II CL2 (black). The reason is to minimize spurious reflections that could affect the optics of the instrument. The interfaces between pieces are not treated to allow a better fit. To minimize the position tolerances of the image sensor, the sensor-PCB is attached to the front cover. In addition, the optical interface with the camera mount is on the outer of the cover surface. Its design depends on whether the camera is to be used for TuMag or SCIP.

In the case of TuMag (Figure 9), it has a circular aperture (36 mm in diameter) for the illumination of the image sensor, with the optical center just in the middle. In this part, TuMag requires assembling an entrance window composed of a fused silica window (Schott FS7980 IF, 5 mm thick, 40 mm diameter, and

10.3389/fspas.2023.1167540

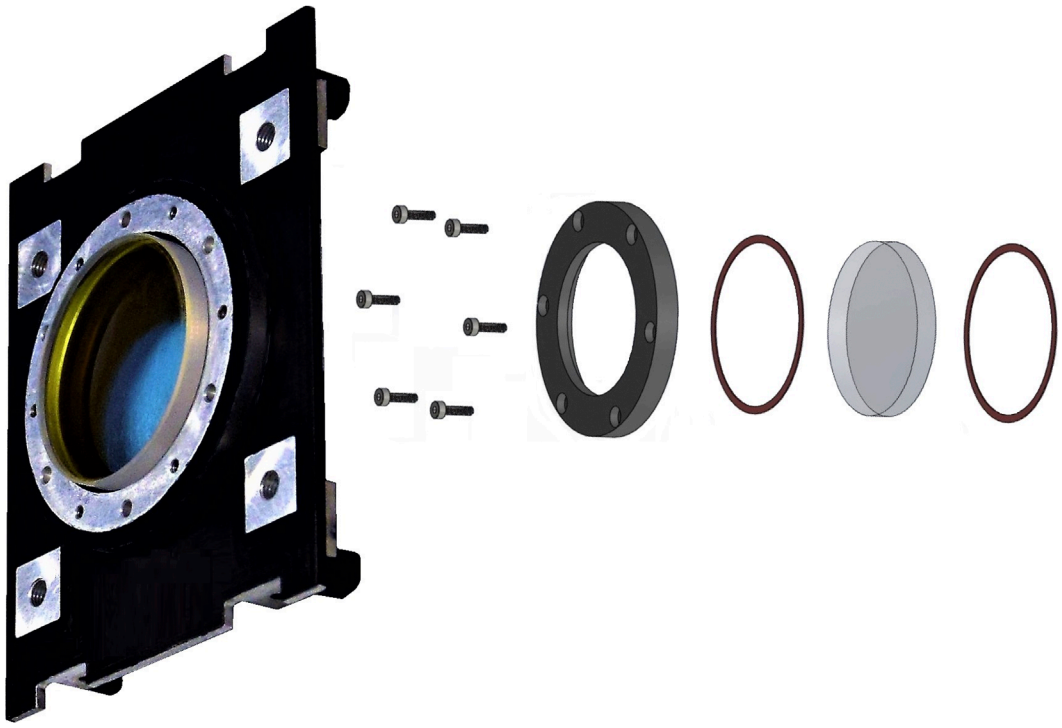


FIGURE 9
TuMag camera entrance window. On the left, parts are partially assembled on the front cover. On the right, individual parts of the entrance window (screws, fixing cover, fused-silica window, and O-rings) are shown.

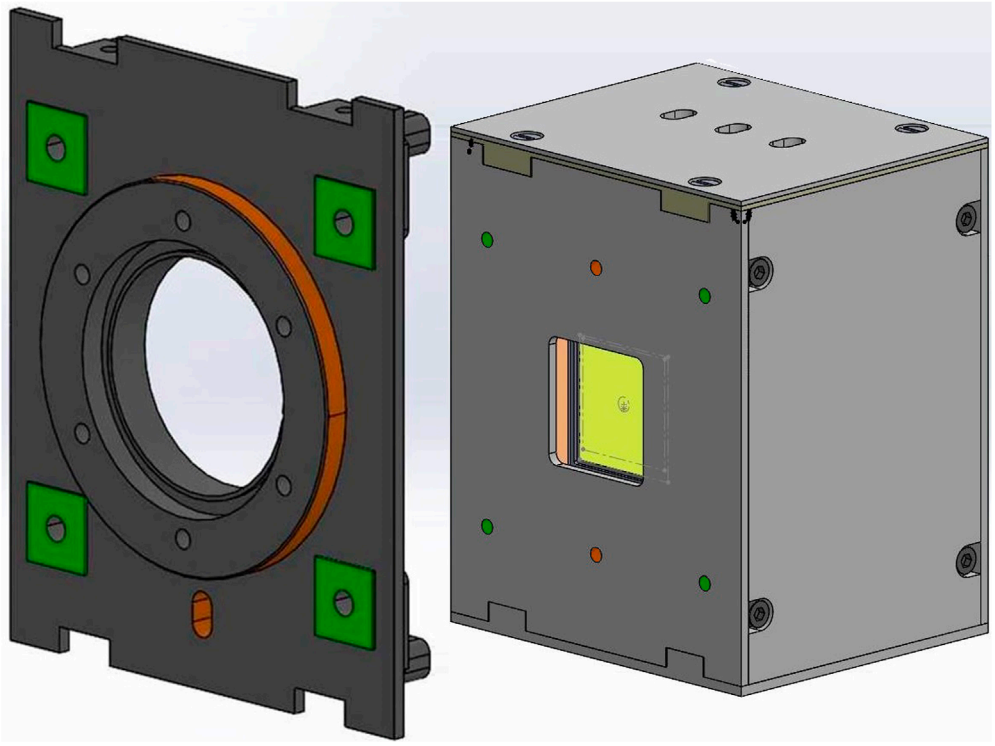


FIGURE 10
Camera interface with the optical bench. Left: TuMag interface. Right: SCIP interface.

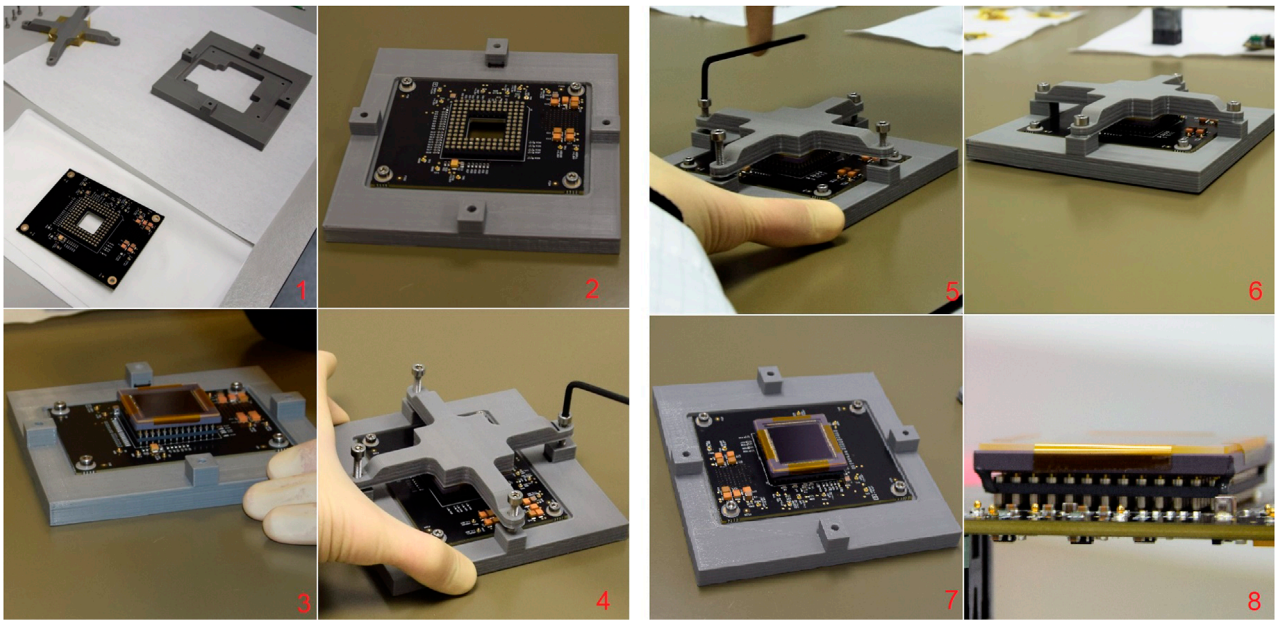


FIGURE 11
Image sensor assembly. 1) Sensor-PCB and assembly tool; 2) sensor-PCB attached to the assembly tool with screws; 3) image sensor smoothly inserted into the PCB socket; 4) upper part of the assembly tool placed on top of the sensor and four M4 screws gradually inserted; 5) tightening the screws (in a "cross-way") until the sensor reaches its final position; 6) sensor inserted in its final position; 7) removing the upper part and the sensor-PCB from the assembling tool; 8) detail of the adhesive drops on the corners of the image sensor/socket.

37 mm aperture), two O-rings of Viton TH -1270 75 Sh (38.25 mm internal diameter and 41.75 mm external diameter, 0.5 mm width, located on both sides of the window), and an aluminum cover with six screws DIN 912 M2 (8 mm length) for fixing the entrance window sub-assembly to the camera housing (applying 0.34 Nm of torque on each screw). The optical window has been designed to protect and isolate the sensor from the environment. To avoid potential interference of ghost images due to quasi-monochromatic illumination of the instrument (517 nm and 525 nm), it includes a highly anti-reflective coating[7] $R < 0.01\%$ at working wavelengths. The former reflectivity has been set to reduce the visibility of possible interfering rays below 10^{-3}, the required (S/N) of the instrument. In the case of SCIP (panel 6 in **Figure 8**), it has a 26 mm × 26 mm aperture for the image sensor illumination, with the optical center right in the middle, but no entrance window.

There are two side covers (panels 1 and 2 in **Figure 8**). These parts mirror each other. They have the fixation interfaces for the electromagnetic protection layer and cold finger (panel 8 in the figure). The electromagnetic protection layer is used to shield the image sensor from any noise radiated by the FPGA-PCB. The cold finger is attached to the side covers using two M3 nylon screws and two Delrin pieces (panel 11 in **Figure 8**) that thermally insulate the junctions.

The back cover (panel 5) has three apertures for the harness connections (CoaXPress, heater power supply, and JTAG). The JTAG connector is used only to update the camera firmware. A protection cover (panel 9 of **Figure 8**) saves this connector from user mishandling (electrostatic discharge issues); it is fixed with two M2 screws on the back cover of the camera.

The top cover insulation is made of Delrin and allows the upper part of the camera (panel 3 in **Figure 8**) to be thermally insulated from the rest of the housing. Attached to the top cover are the cold finger and a thermal strap (all three are fixed with three M4 screws) to transfer the heat from the image sensor/heater to the radiator.

The sensor-PCB is fixed to the front cover with four 35-mm, brass–nickel male/female spacers (the male side is represented in panel 8 in **Figure 8**). This spacer also provides the separation distance with the FPGA-PCB in accordance with the Hirose IT5 mezzanine connectors.

Finally, a bottom cover (panel 4 in **Figure 8**) completes the housing. All these parts are assembled with black-coated stainless steel DIN 912 M3 screws.

5.2 Mechanical interface with the optical bench

TuMag and SCIP have different optical mounts for the cameras, which is the reason why the cameras have different front covers for each of them. In the case of TuMag, the interface between the camera mount and the front cover of the housing has the following features in order to ensure precise positioning on the optical bench:

7 Measured in the laboratory and at 0° incidence, it fulfills the R < 0.05 % TuMag requirement.

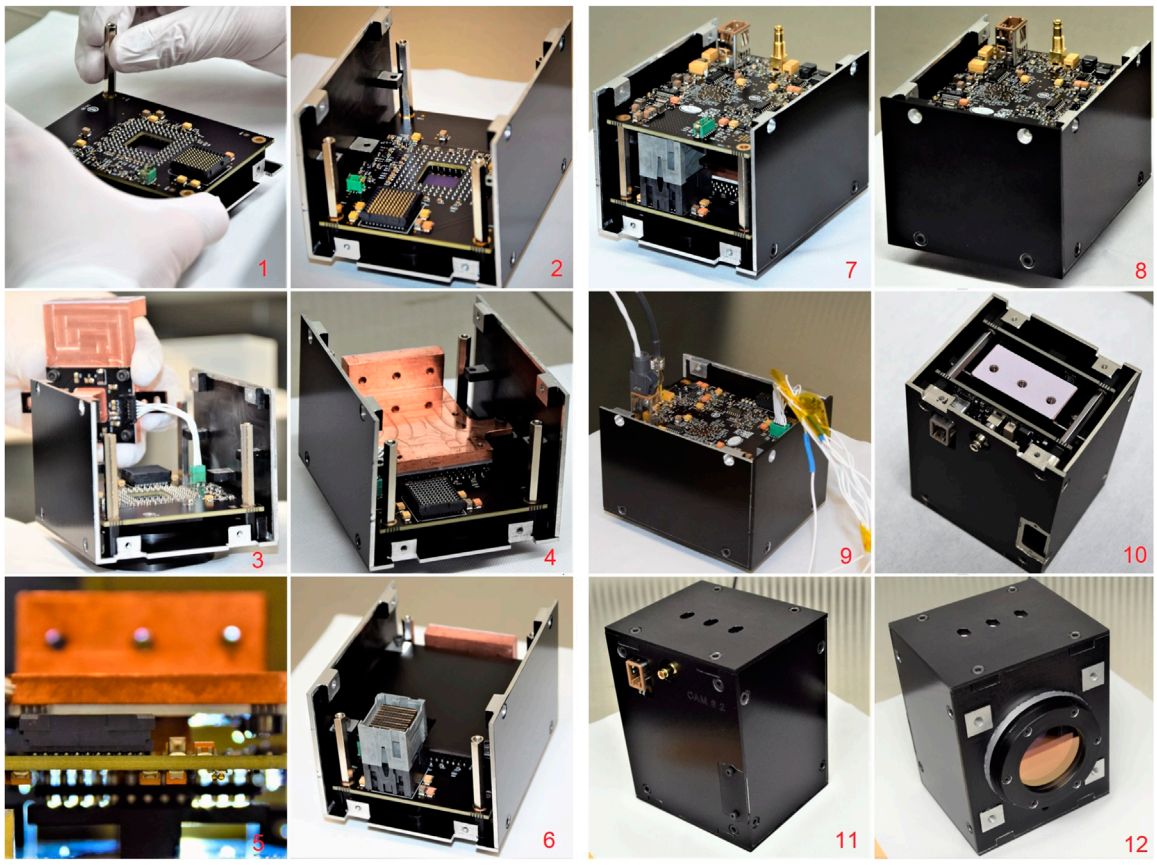


FIGURE 12
Camera assembly. 1) Sensor-PCB being assembled on the front cover; 2) side covers assembled; 3) cold finger harness plugged; 4) cold finger assembled to the side covers; 5) visually checking if the cold finger comes into contact with the image sensor. If so, no light appears through the PGA pins and image sensor; 6) EM shield placed on its position and Hirose IT5 interposed connected to the sensor-PCB; 7) FPGA-PCB placed; 8) bottom cover placed; 9) performing an electrical test to discard problems; 10) back cover assembled; 11) top cover and JTAG connector cap fitted; rearview of the camera assembled; 12) front view of the assembled camera (TuMag).

- Four coplanar platforms (0.5 ± 0.025 mm height) that define the reference contact plane with the mount. These provide good parallelism with the image sensor surface (see the four green areas in **Figure 10**).
- A 60-mm-diameter cylinder (tolerance 60 h7, 4 mm height) together with a 7 mm × 4 mm elliptical slot (tolerance 4F7) allows the rotation to be adjusted along the optical axis (coinciding with the center of the cylinder and the optical center of the image sensor) (see the orange areas in **Figure 10**).
- Once the camera is in the correct position, four M4 screws are used to fix it to the mount. The holes are located in the center of the four coplanar platforms.

In the case of SCIP, these coplanar platforms are not necessary because the camera mount absorbs the tolerances within its own adjustment/refocus system. SCIP cameras do not have an entrance window either. In this way, a simpler design is adopted in which the front of the camera has:

- Two blind positioning holes of 3 mm diameter (tolerance 3H7) and 6 mm depth (orange areas of the right panel in **Figure 10**).
- Once the camera is in the correct position, it is fixed to the mount with four M3 screws (8 mm depth) (green positions of the right panel in **Figure 10**.). The coplanar platforms and slots are masked during the coating to ensure correct dimensions and tolerances, whilst the electrical conductivity is maintained.

5.3 Camera assembly

Before being assembled into the camera, the electronics parts (FPGA-PCB, sensor-PCB, heater-PCB, and the camera heater harness) were tested and characterized to verify that they met the requirements. All the processes were carried out within a clean room in a laminar flow cabinet to prevent dust particles from entering the camera, an antistatic mat to prevent electrostatic discharges, and with all the due quality measures for the operator.

The image sensor is inserted into a socket on the sensor-PCB. An assembly tool has been designed to perform this in a safe, homogeneous, and repeatable way (see **Figure 11**). It consists of

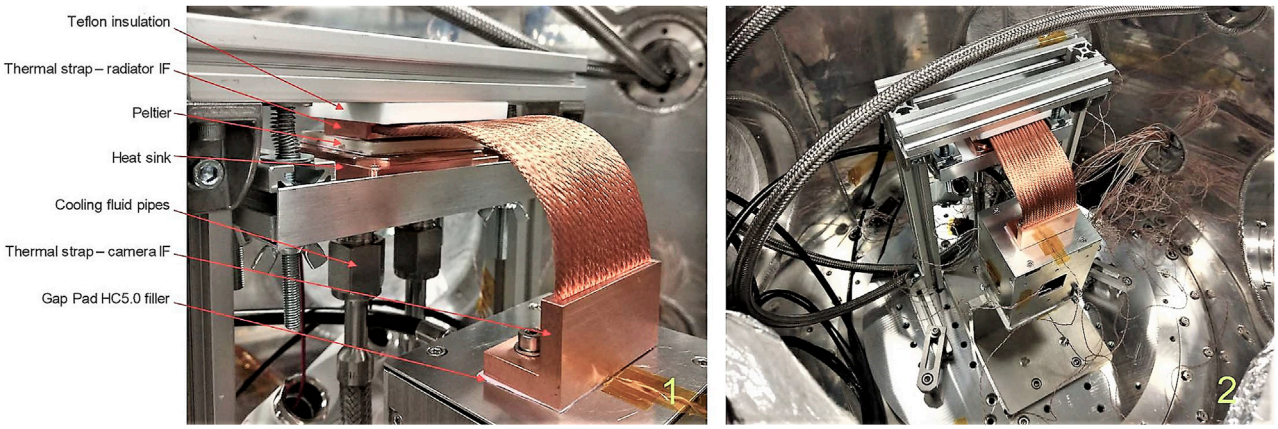


FIGURE 13
TVT setup. 1) Test setup to cool the prototype camera radiator; 2) overall view of the test setup in the vacuum chamber. The bundle of fine wires are thermocouples to acquire the temperature of the camera at several points on its electronics and cover.

two pieces: one serves to house the sensor-PCB and the other to uniformly press the image sensor by tightening four screws until that piece reaches a physical stop. Note that the image sensor includes a protective glass window so that no optical damage occurs during the insertion process. After removing the PCB from the assembling tool and testing the image sensor, a drop of Scotch-Weld 2216 adhesive is applied to each corner and cured for 24 h. Once this process is finished, before assembling the sensor-PCB on the front cover of the camera, the protective glass window is removed. To clean any possible dust particles on the bare sensor, an air-spray, dust remover was applied in several of the following steps of the camera assembly, which are illustrated in **Figure 12**.

The sensor-PCB is screwed to the front cover (in the case of TuMag, the entrance window is already mounted on the front cover) with four spacers (M3, 35 mm long), and 0.7 Nm of torque is applied on each. The two sides of the housing (panels 1 and 2 in **Figure 8**) are fastened to the front cover with two M3 screws (8 mm long). At the moment, no torque is applied.

The heater harness is connected to the sensor-PCB and the cold finger assembly. The cold finger assembly is then positioned in front of the sensor and attached to the camera housing sides with two M3 (8 mm long) nylon screws with 0.5 Nm of torque. A visual inspection confirms that the cold finger touches the image sensor (no visible gap).

Then, the Hirose IT5 interposer is connected to the sensor-PCB receptacle. The electromagnetic shield is then screwed to the housing sides with four M3 screws (6 mm long) and 0.5 Nm of torque. The FPGA-PCB is now placed on the camera and temporally screwed into the turrets (without applying torque). The bottom cover is fastened with two M3 screws (8 mm long). Their torque is applied later. An electrical functional test is then performed to check if the camera behaves correctly. If positive, the camera assembly continues.

Before attaching the back cover of the camera, the screws that temporally hold the FPGA-PCB in place are removed. Then, the back cover is fastened with four M3 screws (16 mm long), although their torque is applied later. Two additional M3 screws (8 mm long) are fitted to the bottom of the camera without applying torque either. Finally, the top cover is tightened to the camera with four M3 screws (8 mm long). These last four and all screws without applied torque are finally tightened with 0.5 Nm.

5.4 Mass budget

The mass of the SPGCams is measured as 825 g (TuMag) and 778 g (SCIP). The breakdown of item masses is as follows:

- Housing assembly: 413 g.
- Cold finger assembly: 208 g.
- Entrance window assembly (TuMag): 47 g.
- Mezzanine PCBs: 157 g.

6 Camera thermo-vacuum characterization

The nominal operation of the camera is under near-vacuum conditions (5 mbar). The only electronic device that, under specification, is designed to operate under these conditions is the image sensor. The remaining components are extended-range industrial components. The reasons for this choice are three-fold. First, some components have no easy alternative. Second, the cameras are not intended to work for long periods in a vacuum, as the duration of the Sunrise iii mission is less than a week (the continuous stress is short). Third and last, the cost of the components is much lower than that of the components of higher grade. Therefore, the behavior of the camera has to be characterized and verified under environmental working conditions.

6.1 Camera temperature characterization

During the development process, a prototype camera has been electrically and thermally characterized in a thermo-vacuum test to

TABLE 3 Summary of inferred noise parameters of a camera taken as an example.

Gain code	G (DN/e$^-$)	N_r(DN)	Δ/12 (DN)	f_{prnu} (%)	$N_{dc,fpn}$ (%)	L (%)	S_{DC} (DN)
1 (1.85 ×)	0.0520	1.77	50,711	1.67	0.49	0.20	131.17 ± 0.68
2 (2.49 ×)	0.0707	2.13	42,236	1.74	0.43	0.22	131.23 ± 0.70
3 (3.68 ×)	0.1034	2.37	37,899	1.87	0.18	0.34	131.11 ± 0.82
4 (1.29 ×)	0.0369	1.61	55,943	1.59	0.70	0.10	130.35 ± 0.74
5 (3.70 ×)	0.1028	2.66	33,828	1.86	0.40	0.39	131.14 ± 1.10

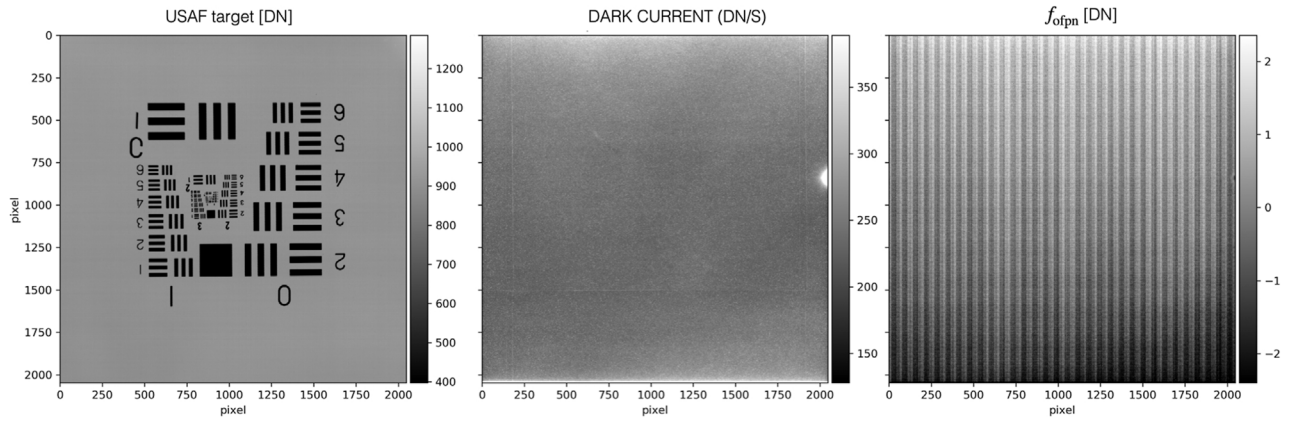


FIGURE 14
USAF target image (left), dark current (center), and offset fix pattern noise (right) derived from dark current measurements. Data were taken in the laboratory under ambient conditions (~ 25°C and 1 bar). The bright spot appearing at the right edge of the dark current image is associated with a hot spot in the sensor circuitry, especially with one of the internal signal clocks that make that area of the sensor slightly hotter than the rest of this specific sample chip.

obtain the information to correlate its thermal model. Under near-vacuum conditions (5 mbar), three environmental temperatures have been used for the camera radiator (screwed into the cold finger of the camera): hot (15°C), nominal (0°C), and cold (−10°C). In the test, the temperature of the radiator is adjusted with a Peltier module, whose hot side is cooled with fluid pipes (see panel 1 in **Figure 13**). A total of 14 temperature sensors have been placed inside and outside the camera to measure its thermal behavior (panel 2 in **Figure 13**).

The thermo-vacuum resulted in a good correlation between the thermal model predictions and the measured temperature response, with a maximum discrepancy of 1.1°C. In all cases, the camera electronic components were within a safe temperature range. This test also confirms that the camera can stabilize the image sensor at a programmable setpoint with ±0.1°C stability.

6.2 Flight camera characterization

Each flight and spare camera is later subjected to the following test campaign to ensure its functionality:

1. Optical characterization under ambient conditions (~ 25°C and 1 bar).
2. Optical and electrical characterization under near-vacuum conditions and three temperatures (hot, nominal, and cold). This provides confidence that the camera behaves correctly if there is no degradation with respect to the first step. No early component failures ("infant mortality") have arisen from that first stress.
3. A bake-out to degas all particles that could degrade the optics of the instrument. (The camera is not supplied with power yet but maintained at 60°C for 72 h and 5 mbar of pressure). Note: SCIP cameras were baked out after delivery.
4. A thermo-vacuum test (three temperatures: hot, nominal, and cold) to verify that the camera temperatures meet the thermal requirements (image sensor, FPGA, and cold finger temperatures are read by the camera itself; the temperature sensors placed on the camera cover are read by the test setup).
5. Finally, a second optical characterization gives us confidence in its behavior, before delivering the camera for assembly into the instrument.

7 Camera optics

The optical performance of the SPGCams has been checked in order to know whether they meet the various optical requirements in **Table 1** and whether dust particles have settled on the sensor surface. A homogeneous Lambertian source is approximated by means of a collimated LED beam at 530 nm, which enters an integrating sphere and is reflected several times before coming out. A nearby,

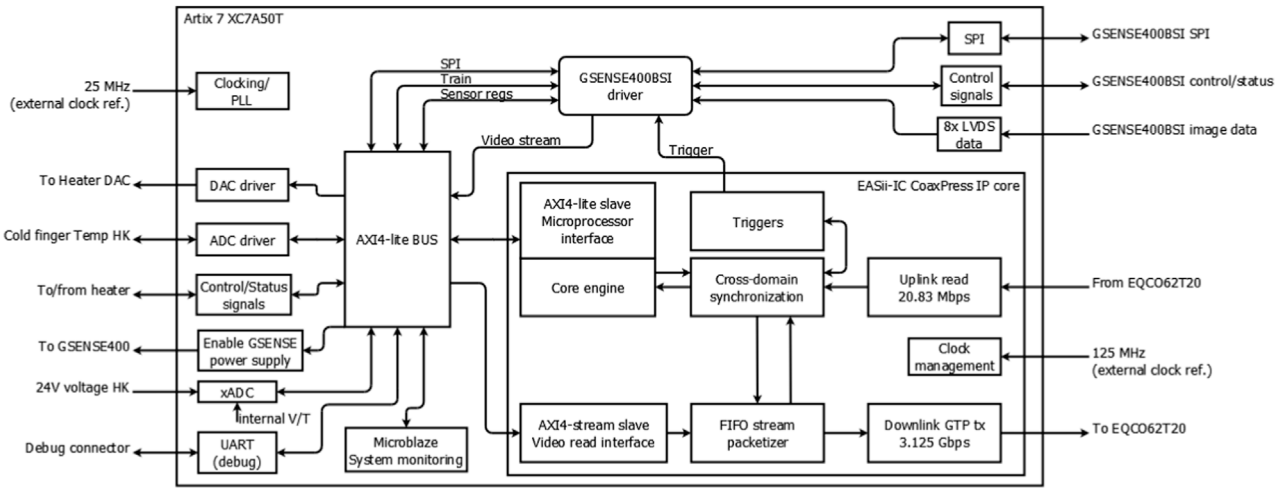


FIGURE 15
FPGA block diagram.

removable USAF pattern is illuminated by that source and used to focus the camera and verify that it is read properly. Next, there is a shutter to control the diameter of the beam and a system of two achromatic lenses to focus the USAF pattern onto the camera, which is inside a vacuum chamber. Light enters the chamber through an optical window aligned with the camera and optical setup.

A number of tests have been carried out in order to check the performance of the cameras. The tests were made in a near-vacuum environment (5 mbar), and the image sensor temperature is regulated at 20°C. These conditions are similar to the flight conditions of the Sunrise iii mission. Two types of tests can be distinguished: 1) a series of dark images were taken (LED off) with the image sensor configured with different acquisition gain codes (1, 2, 3, 4, and 5; each code involves 5,000 images) and exposure times. To reduce the amount of data to process, a region of interest of 128 px × 1024 px (rows [860, 988] and columns [512, 1536]) was used. 2) Linearity and gain tests to determine L and g. These tests were also performed for gain codes 1 through 5. The sensor was illuminated by the 530 nm LED. For each gain, a sequence of 30 different exposure times was taken. The exposure time was increased from 0.0205 ms until the full well capacity was reached (it depends on the gain mode). Two images were taken per exposure time. Dark images were also acquired under the same conditions.

To calculate the various parameters characterizing the noise, we have followed the guidelines described in **Section 3**. A summary of the inferred noise parameters is presented in **Table 3**. A few dust particles are identified in those regions exhibiting anomalous low signal levels. The number of pixels covered by dust particles remains below 0.001% of the whole image for all the cameras. **Figure 14** shows a USAF target, a dark current, and a map of the offset fix pattern noise images acquired with one of the prototype cameras during characterization tests. The latter is derived from a sequence of dark current images taken at different exposure times.

8 Camera firmware

The FPGA firmware is in charge of managing all the tasks of the camera. Several blocks can be distinguished, as shown in **Figure 15**. The brain of the system is an embedded MicroBlaze processor that interprets commands from the frame grabber and acts accordingly, making the camera to perform what it is ordered. The MicroBlaze achieves it by interfacing through the standard AXI4 protocol with the rest of the blocks/IP cores that perform any singular task. There is a block dedicated to controlling the image sensor. It is based on the code provided by GPIXEL on its GSENSE400 evaluation board but adapted with other features and migrated to another Xilinx FPGA (Artix 7 vs. Spartan 6) and development suite (Vivado). This is the biggest advantage of choosing Artix 7 because Vivado is not supported by Spartan 6 (which uses the older and superseded Xilinx ISE suite). It is not futile because Vivado implies a huge improvement in the efficiency of firmware development. Another advantage of the Artix 7 is that it is a newer device (2017 vs. 2009), with a next-generation manufacturing process (higher performance, higher circuit density, and lower power) and will likely be discontinued later.

9 Conclusion

This paper reports on the design, fabrication, and verification of the SPGCams, initially thought suitable for the TuMag and SCIP instruments aboard the Sunrise iii mission but with capabilities which can be exploited by other solar instruments. They have been specifically designed to comply with the requirements of those two spectropolarimeters. In particular, of special importance is the available signal-to-noise ratio provided by a system which is illuminated such that shot noise (proportional to the square root of the signal) is the most important source of measurement contamination, hence avoiding other sources proportional to the

signal itself. Circumventing these other signals is achieved by filling the detector well up to less than half its capacity.

Author contributions

DO was the scientist who led the development, designing the experiments, and checking all the results. DA was the main electronic engineer of the cameras and the main writer of the technical parts of the paper. AL was the system engineer, hence in charge of the interfaces with the rest of the instruments. MB was the project manager of the whole TuMag and SCIP development and responsible for all the technical documentation. DH was mainly responsible for the firmware. JM, BA, AS, and EB contributed to firmware development. PL was the mechanic designer of the SPGCams. FB was responsible for the optics design and verification. IB was responsible for the mechanical fabrication of the cameras. AT was responsible for the assembly of the cameras. AM and JR contributed to the electronic verification. IP and JP were the thermal engineers of the project. YKat and MK were responsible for defining specifications of the SCIP cameras. YKaw and TO carried out camera performance tests. MR and EM contributed to the firmware and embedded system software. JD was the principal investigator of the project and was overall responsible for writing the paper.

Funding

This work was funded by the Spanish MCIN/AEI, under projects RTI 2018-096886-B-C5, PID 2021-125325OB-C5, and PCI 2022-135009-2, and co-funded by European FEDER funds, "A way of making Europe," under grants CEX 2021-001131-S and 10.13039/501100011033.

Acknowledgments

The authors thank Shinnosuke Ishikawa and Noriyuki Narukage from the National Astronomical Observatory of Japan who originally suggested using GPIXEL back-illuminated CMOS sensors for Sunrise III in the early phase of the project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Del Toro Iniesta, J. C. (2003). *Introduction to spectropolarimetry*. Cambridge, UK: Cambridge University Press.

Del Toro Iniesta, J. C., Orozco Suárez, D., Álvarez-Herrero, A., Sanchis Kilders, E., Pérez Grande, I., Ruiz Cobo, B., et al. (2023). TuMag: The tunable magnetograph for the Sunrise III mission. *Sol. Phys.* (in preparation).

Katsukawa, Y., Del Toro Iniesta, J. C., and Solanki, S. K. (2023). The Sunrise chromospheric infrared spectroPolarimeter. *Sol. Phys.* (in preparation).

Lites, B. W. (1987). Rotating waveplates as polarization modulators for Stokes polarimetry of the sun: Evaluation of seeing-induced crosstalk errors. *Appl. Opt.* 26, 3838–3845. doi:10.1364/AO.26.003838

Martínez Pillet, V., Del Toro Iniesta, J. C., Alvarez-Herrero, A., Domingo, V., Bonet, J. A., Gonzalez Fernandez, L., et al. (2011). The Imaging Magnetograph eXperiment (IMaX) for the Sunrise balloon-borne solar observatory. *Sol. Phys.* 268, 57–102. doi:10.1007/s11207-010-9644-y

Rust, D. M. (1994). The flare Genesis project. *Adv. Space Res.* 14, 89–93. doi:10.1016/0273-1177(94)90072-8

Solanki, S. K., Gandorfer, A., Chares, B., Curdt, W., Deutsch, W., Feller, A., et al. (2023). The Sunrise III mission. *Sol. Phys.* (in preparation).

Zhang, Y. H., Zhang, H. F., Qu, W. Q., Wang, Z. Y., Zhu, Z. Y., Zhu, J., et al. (2020). "Design of a large-format high-rate scientific CMOS camera," in *Society of photo-optical instrumentation engineers (SPIE) conference series* (SPIE), 1145407. doi:10.1117/12.2559909

Zhu, J., Zhang, H. F., Zhang, Y., Tang, Y., Zhu, Z., Zhang, J., et al. (2022). "Design of a scientific CMOS camera for astronomical observation," in *X-ray, optical, and infrared detectors for astronomy X*. Editors A. D. Holland, and J. Beletic (Montréal, Québec, Canada: SPIE), 1219126. doi:10.1117/12.2629345

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Mini Review
PUBLISHED 10 May 2023
DOI 10.3389/fspas.2023.1179000

Check for updates

OPEN ACCESS

EDITED BY
Steve B. Howell,
National Aeronautics and Space Administration, United States

REVIEWED BY
Daniel Thorngren,
Johns Hopkins University, United States
Jun Yang,
Peking University, China

*CORRESPONDENCE
Simon Müller,
simonandres.mueller@uzh.ch

RECEIVED 03 March 2023
ACCEPTED 21 April 2023
PUBLISHED 10 May 2023

CITATION
Müller S and Helled R (2023), Warm giant exoplanet characterisation: current state, challenges and outlook.
Front. Astron. Space Sci. 10:1179000.
doi: 10.3389/fspas.2023.1179000

COPYRIGHT
© 2023 Müller and Helled. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Warm giant exoplanet characterisation: current state, challenges and outlook

Simon Müller* and Ravit Helled

Center for Theoretical Astrophysics and Cosmology, Institute for Computational Science, University of Zurich, Zurich, Switzerland

The characterisation of giant exoplanets is crucial to constrain giant planet formation and evolution theory and for putting the solar-system's giant planets in perspective. Typically, mass-radius (M-R) measurements of moderately irradiated warm Jupiters are used to estimate the planetary bulk composition, which is an essential quantity for constraining giant planet formation, evolution and structure models. The successful launch of the James Webb Space Telescope (JWST) and the upcoming ARIEL mission open a new era in giant exoplanet characterisation as atmospheric measurements provide key information on the composition and internal structure of giant exoplanets. In this review, we discuss how giant planet evolution models are used to infer the planetary bulk composition, and the connection between the compositions of the interior and atmosphere. We identify the important theoretical uncertainties in evolution models including the equations of state, atmospheric models, chemical composition, interior structure and main energy transport processes. Nevertheless, we show that atmospheric measurements by JWST and ARIEL and the accurate determination of stellar ages by PLATO can significantly reduce the degeneracy in the inferred bulk composition. Furthermore, we discuss the importance of evolution models for the characterisation of direct-imaged planets. We conclude that giant planet theory has a critical role in the interpretation of observation and emphasise the importance of advancing giant planet theory.

KEYWORDS

planets and satellites: gaseous planets, formation, evolution, interiors, composition, characterisation

1 Introduction

The study of giant exoplanets gives a unique peek into the formation of planets, because their composition is linked to their form history (e.g., Mousis et al., 2009; Helled et al., 2014; Johansen and Lambrechts, 2017; Ginzburg and Chiang, 2020). Since the first discovery of the hot Jupiter 51 Peg b (Mayor and Queloz, 1995), there have been over a thousand detections of giant exoplanets with diverse masses, sizes, and equilibrium temperatures (see **Figure 1**). Most observed giant exoplanets are hot Jupiters, but a fraction are so-called warm giants with equilibrium temperatures below ~1000 K. Warm giants are particularly interesting objects, since, compared to hot Jupiters, they are superior for characterisation. This is because hot Jupiters are inflated by a poorly understood mechanism (e.g., Fortney and Nettelmann, 2010; Weiss et al., 2013; Baraffe et al., 2014; Fortney et al., 2021; Sarkis et al., 2021), and therefore their interiors are more difficult to characterise. The large observed radii of giant exoplanets imply the presence of massive hydrogen-helium

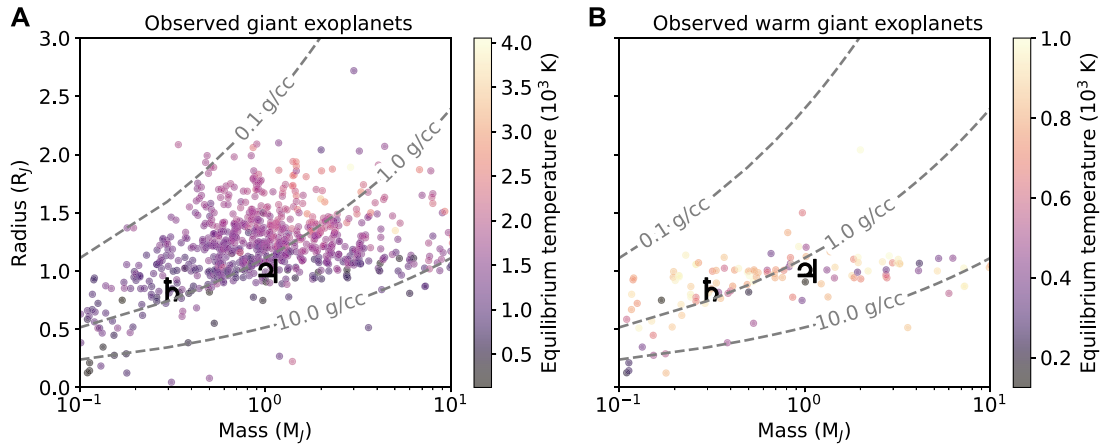


FIGURE 1
Masses and radii of giant exoplanets with masses between $0.1 \leq M_J \leq 10$ **(A)** and equilibrium temperatures below 1000 K **(B)**. The dashed lines show lines of constant density. Jupiter (♃) and Saturn (♄) are shown with their respective symbols. Error bars are omitted for readability. The data was taken from the NASA Exoplanet Archive.

envelopes, which can progressively contract and cool as the planets evolve in time (e.g., Hubbard, 1977; Burrows et al., 2001). This implies that any effort to characterise these planets must rely on theoretical (numerical) models that simulate the planetary evolution.

The combination of evolution models with measurements of planetary radius, mass, and stellar age enables the estimation of the planetary bulk composition, which is, in the case of giant planets, represented by the heavy-element mass (e.g., Guillot et al., 2006; Fortney et al., 2007; Guillot, 2008; Miller and Fortney, 2011; Mordasini, 2014; Thorngren et al., 2016). This is done by comparing the observed radius to the one predicted by evolution models. In order to include the measurement uncertainties, often the bulk metallicity is inferred as a posterior distribution. Since giant planets cool and contract, their evolution path directly affects their characterisation: At a given observed planetary age, a different planetary radius would be predicted, yielding a different estimate of the composition. Furthermore, a determination of the bulk composition of giant planets relies on accurate measurements of the planetary radius, mass and stellar ages. Otherwise, the inferred heavy-element mass is highly uncertain. While planetary masses and radii are often somewhat well constrained, stellar ages are not well-determined and have uncertainties of a few Gyr. The upcoming PLATO mission (Rauer et al., 2014) is expected to accurately determine stellar ages, which is crucial for constraining the planetary composition, in particular for young planets that cool and shrink rapidly.

Furthermore, JWST (Gardner et al., 2006) and the upcoming ARIEL mission (Tinetti et al., 2018) will push exoplanet characterisation to the next level by determining the atmospheric composition of many giant exoplanets. To-date only several hot Jupiter atmospheres have been studied (e.g., Madhusudhan et al., 2014; Welbanks et al., 2019; Ben-Yami et al., 2020; Line et al., 2021), with recent detection of both water (Rustamkulov et al., 2022) and carbon dioxide (The JWST Transiting Exoplanet Community Early Release Science Team et al., 2022). Future JWST and ARIEL observations will provide a more complete picture of the atmospheric compositions of hot Jupiters. They will also observe warm Jupiters, whose interior models have much less theoretical uncertainties compared to hot Jupiters and are therefore superior targets for interior characterisation. Therefore, the characterisation of warm giant exoplanets is essential to better understand the origins of giant planets (Teske et al., 2019; Turrini et al., 2021; Knierim et al., 2022).

Additionally, masses and radii of exoplanets are now more accurately determined by observations. This is in part due to more precise radial velocity and transit measurements, but also thanks to improved information of the host stars from the GAIA mission (Gaia Collaboration et al., 2016). Therefore, exoplanetary science is now in an era at which theoretical uncertainties are no longer negligible and can affect the data interpretation. In order to interpret this current and upcoming wealth of data, a robust theoretical foundation and knowledge of how giant planets evolve are required.

In this review, we summarise the current state-of-the art in warm giant planet characterisation, and discuss the challenges due to theoretical uncertainties of evolution models. We discuss the role of evolution models in the interpretation of upcoming PLATO, JWST and ARIEL measurements, and show that these missions will significantly improve our understanding of giant exoplanets.

2 Giant planet evolution models

Giant planet evolution models are constructed assuming that the planets are spherically symmetric and in hydrostatic equilibrium. The equations of planetary evolution are then written as a coupled set of partial differential equations describing the mass, momentum and energy conservation, and the transport of chemical elements and energy (e.g., Kippenhahn et al., 2012). In most models, energy transport is assumed to be by radiation (and conduction) or convection, however there is also the possibility of double-diffusive energy transport (Wood et al., 2013; Radko et al., 2014). The Ledoux criterion (Ledoux, 1947) is used locally to determine the dominant energy mechanism. Convection is usually modelled with the

mixing-length theory (Böhm-Vitense, 1958). A good approximation is that rigorous convection leads to (nearly) adiabatic planetary interiors. The missing pieces are an appropriate equation of state, a description of the opacity, and atmospheric boundary conditions.

Since there is no general analytical solution to the evolution equations they must be solved numerically (e.g., Henyey et al., 1965; Guillot and Morel, 1995; Burrows et al., 1997; Vazan et al., 2015; Thorngren et al., 2016). A popular open-source code is Modules for Experiments in Stellar Astrophysics (MESA; Paxton et al. (2011); Paxton et al., 2013; Paxton et al., 2015; Paxton et al., 2018; Paxton et al., 2019); Jermyn et al. (2022)), which is also suitable to calculate the evolution of planets (e.g., Mankovich et al., 2016; Berardo and Cumming, 2017; Cumming et al., 2018; Müller et al., 2020b; Malsky and Rogers, 2020). Recently, Müller and Helled (2021) used MESA to calculate an extensive grid of giant planet evolution models, and developed a python module to generate cooling tracks by interpolation.

2.1 Theoretical uncertainties in giant planet evolution models

Giant planet evolution models have to make a range of assumptions, such as the equations of state (hereafter EoS) for the different elements and their distribution in the interior, the opacity and the atmospheric model, as well as what metal is used to represent the heavy elements. These choices influence the cooling, and therefore the predictions from the evolution models (such as radius or luminosity) to various degrees (Guillot, 1999). In the following sections, we discuss the key assumptions and parameters that affect giant planet evolution models, and show how they can influence the interpretation of exoplanetary data.

2.1.1 Equations of state and atmospheric models

Typically, giant planet evolution models involve three different equations of state that are combined with the ideal-mixing law: hydrogen (H), helium (He) and a heavy element (Z). There are several hydrogen-helium equations of state that are often used in the modelling of giant planets (Saumon et al., 1995; Militzer and Hubbard, 2013; Becker et al., 2014; Chabrier et al., 2019). Since giant planets are hydrogen-dominated in composition, the uncertainties in the H EoS have important implications for the interior modelling of giant planets (Helled et al., 2020; Howard et al., 2023). For example, switching between the one from Saumon et al. (1995) to Chabrier et al. (2019) leads to predicted radii that are smaller by up to ~10% (Müller et al., 2020a). In addition, non-ideal H-He interactions that are often neglected are non-negligible and can change the predicted radius by up to 8%, especially at younger ages (Howard and Guillot, 2023).

A common simplification of evolution models is that they are limited to representing all the heavy elements with one component. This is often water, a type of rock such as olivine or SiO_2, or a water-rock mixture (see, e.g., More et al., 1988; Thompson, 1990; Mazevet et al., 2019). However, in reality, giant planets are not expected to contain a single heavy element or rock-type, but rather a combination. The uncertainty from the heavy-element composition also affects the predicted radius, likely by a few percent (e.g., Baraffe et al., 2008; Vazan et al., 2015). This is usually most significant for lower-mass giant planets, with masses below ~0.3M_J.

Since giant planets cool down by radiating energy from their atmospheres, the opacity and the atmospheric model are crucial for the evolution. Unfortunately, these parameters are not well constrained. For the opacity, different contributions must be considered: molecular, grains and clouds. A commonly used molecular opacity is from Freedman et al. (2014) which also includes the effect of heavy elements. This is particularly important for irradiated planets with enriched atmospheres. However, this also means that if the atmospheric composition is unknown, the molecular opacity is not well determined. It is also not known whether clouds or grains are present and how efficient they are at blocking infrared radiation as clouds could trap heat in the interior and slow down cooling (Vazan et al., 2013; Mordasini, 2014; Poser et al., 2019). An additional complication is that most giant exoplanets observed today are highly irradiated and have an unknown albedo. A variety of models exist to account for the instellation, including semi-gray and full atmospheric models (Fortney et al., 2007; Guillot, 2010; Parmentier and Guillot, 2014). The uncertainties linked to the opacity are particularly large for young planets due to their fast contraction. In that case, differences in opacity can significantly affect the estimates of the planetary parameters. Overall, the uncertainties in opacities and atmospheric models probably lead to ~10% uncertainty in the predicted planetary radius (Valsecchi et al., 2015; Müller et al., 2020a), but are potentially even larger for young planets.

2.1.2 Primordial state, energy transport, and distribution of elements

For young giant planets (~1–10 Myr), an additional complication is that the primordial internal structure and its thermal state are unknown and likely depends on the formation history (e.g., Baraffe et al., 2003; Marley et al., 2007; Spiegel and Burrows, 2012). If giant planets are fully convective shortly after their formation, this issue is resolved after a few 10 Myr, since different initial configurations converge to the same cooling track (e.g., Marley et al., 2007; Berardo and Cumming, 2017; Berardo et al., 2017; Cumming et al., 2018). However, if young giant planets are not fully convective, the primordial state will significantly affect the planetary contraction. Recent formation models predict that young giant planets are expected to have composition gradients and therefore may not be fully convective, which complicates the situation (e.g., Helled and Stevenson, 2017; Lozovsky et al., 2017; Valletta and Helled, 2020; Stevenson et al., 2022). Indeed, studies of Jupiter and Saturn suggest that parts of their interiors are not fully convective today (e.g., Debras and Chabrier, 2019; Mankovich and Fuller, 2021).

If young giant planets are not mostly convective, the initial conditions are not lost as quickly and the cooling is slower. Therefore, the luminosity at young ages is lower than predicted from fully convective models. After a few Gyr, the predicted luminosity would be higher, since the energy transport is less efficient (Leconte and Chabrier, 2012; Wood et al., 2013; Radko et al., 2014) and there is more primordial heat trapped inside the planet. This introduces an additional uncertainty to evolution models. For example, Kurokawa and Inutsuka (2015) showed that old giant planets cooling by double-diffusive convection are inflated by ~10%, which would mean an increase in their luminosity by ~20%. In addition to the heat transport mechanism, the radius of a giant planet is also influenced

by the distribution of chemical elements in the interior. The two extremes are that all heavy elements are in the core (core-envelope structure) or homogeneously distributed. The latter generally results in a smaller planet by a few percent (e.g., Baraffe et al., 2008; Vazan et al., 2013; Müller et al., 2020a).

Lastly, it is also possible that extraneous events, such as giant impacts, contribute to the inflation of giant planets. These are usually not included in evolution models due to their stochastic nature. However, unless collisions are frequent and violent, which is unlikely after a few Gyr, the energy deposited during the impacts is quickly re-radiated, and the planets are only inflated for a short time, up to a few 10^4 years (Müller and Helled, 2023). Nevertheless, collisions could be important for the interpretation of individual giant planets.

3 Warm giant exoplanet characterisation

3.1 Mass-metallicity trends of warm giant planets

Thorngren et al. (2016) used evolution models to infer the metallicity of 47 warm giant planets ($20M_\oplus < M < 20M_J$) with moderate instellation fluxes ($F_\star < 2 \times 10^8$ erg s^{-1} cm^{-2}). The main results from the study were that there is 1) a correlation between the heavy-element mass of a planet and its total mass, and 2) a strong relation between the metal-enrichment of a planet ($Z/Z_\star$, where $Z_\star$ is the metallicity of the host star) and its mass. They found that the mass-scaling was approximately $M_z \propto \sqrt{M}$.

Teske et al. (2019) performed a slightly different study by focusing on whether different host-star heavy elements are correlated with the bulk heavy-element mass of warm giants. Their findings suggest that the stellar metallicity is not correlated with the planetary residual metallicity, i.e., the residual metallicity that is not explained by the trend with planetary mass. Using different evolution models, Müller et al. (2020a) later independently inferred a qualitative mass-metallicity correlation in agreement with the results from Thorngren et al. (2016). However, they also showed that different model parameters, such as EoS and opacity, can have a large influence on the inferred bulk metallicity of giant planets. A similar mass-metallicity trend was also found by Müller and Helled (2023) who analysed warm giants in the Ariel mission reference sample (Edwards and Tinetti, 2022). Their results suggest a lower heavy-element mass for a given planetary mass compared to Thorngren et al. (2016), and also a less statistically significant correlation. It is fair to say that currently it is unclear to what extent, if any, trend exists between the planetary heavy-element mass and planetary mass. This is in particular due to the many theoretical uncertainties associated with the models used for the data interpretation. Therefore, there is a risk of over-interpreting the comparisons between predictions from formation models with the inferred mass-metallicity trends from observations.

The inferred mass-metallicity correlation serves as an important constraint for planet formation models, and can be used to test planet formation theory. The observation that the metal enrichment decreases with planetary mass is in qualitative agreement with the core-accretion model (e.g., Pollack et al., 1996), in which a heavy-element core accretes large amounts of metal-poor gas as the planet grows. Planetary population synthesis models (Mordasini et al., 2014) also yield a similar power-law, however with different exponents. It has also been suggested that the observed metal-enrichment can be explained if it traces the final assembly of giant planets, where the heavy elements are predominantly accreted from a planetesimal disks with large gaps (Hasegawa et al., 2018). One of the major unresolved problems is that the large heavy-element masses of some giant exoplanets cannot be explained by standard formation models. Formation models suggest that the core masses of giant planets are limited to a few $10M_\oplus$ (e.g., Pollack et al., 1996; Helled et al., 2014; Bitsch et al., 2018). Since the accreted gas likely has a stellar metallicity, the additional metal enrichment has to come from different sources, such as planetesimal accretion (Mousis et al., 2009; Shibata and Ikoma, 2019; Shibata et al., 2020), pebble accretion (Johansen and Lambrechts, 2017), and collisions between planetary embryos (Ginzburg and Chiang, 2020).

4 Connection to future observations

In this section, we use the evolution models from Müller and Helled (2021) to demonstrate how upcoming observations can improve the characterisation of giant exoplanets (**Section 4.1** and **Section 4.2**) and the importance of evolution models in determining the mass of direct-imaged planets (**Section 4.3**).

4.1 PLATO: the importance of accurate stellar age measurements

Stellar ages are currently often only determined within a few Gyr, which makes the inferred composition of giant exoplanets degenerate. This is demonstrated in **Figure 2A** for three exoplanets that cover the typical mass-range of warm giants (Kepler 16b, Kepler 167e and K2 144b). The coloured lines show the calculated evolution of the three planets for different bulk metallicities. By comparing the observed radii to the one from the evolution models, it is clear that the large uncertainty of the stellar age causes a degeneracy in the inferred composition. Kepler 16b, for example, has a very precisely determined radius, but its composition still cannot be clearly determined due to its large age uncertainty. This is where the PLATO mission will clearly improve giant exoplanet characterisation: It will determine stellar ages to within an accuracy of ~ 10%, breaking the degeneracy in the inferred heavy-element mass and reduce its uncertainty by about a factor of two (Müller and Helled, 2023).

4.2 JWST and ariel: atmospheric measurements and the connection to the bulk composition

Measuring the chemical composition of giant planets' atmospheres promises a new dawn in giant exoplanets characterisation: Knowledge of the atmospheric composition can reveal information on the planetary interior and origin (Burrows, 2014; Teske et al., 2019; Turrini et al., 2021; Edwards et al., 2022; Helled et al., 2022). Although the connection between

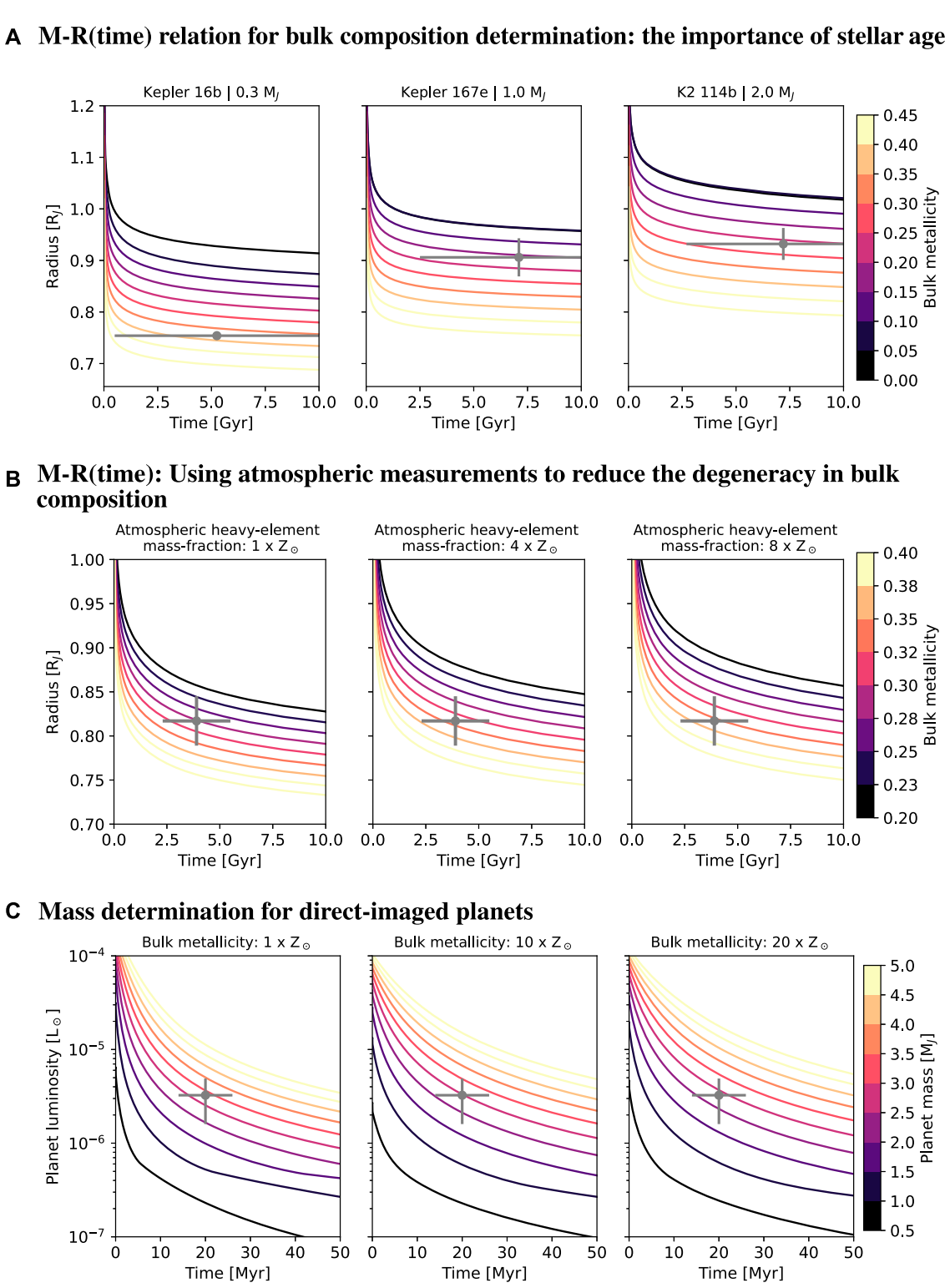


FIGURE 2
(A) Radius evolution for different bulk metallicites shown together with three exoplanets (Kepler 16b, Kepler 167e, K2 114b). The large age uncertainties cause a degeneracy in the inferred composition. **(B)** Radius evolution for different bulk (coloured lines) and atmospheric (columns) heavy-element mass-fraction shown together with the exoplanet NGTS 11b. Measuring the atmospheric metallicity breaks the degeneracy in the inferred composition. **(C)** Luminosity evolution for different masses (coloured lines) and bulk metallicites (columns) shown together with the directly imaged exoplanet 51 Eri b. The inferred mass of the planet depends on the assumed composition.

the atmospheric composition and that of the planetary bulk is challenging and is yet to be determined (e.g., Helled et al., 2022), in the case of warm Jupiters the planets can be better characterised. In particular, it was clearly shown that atmospheric measurements can break the degeneracy in determining the planetary bulk composition and it is expected that measurements of warm Jupiter atmospheres can reduce the uncertainty of the bulk composition by at least a factor of four (Müller and Helled, 2023). We demonstrate this in **Figure 2B**, where we simulate the evolution of NGTS-11 b for various bulk and atmospheric heavy-element mass-fractions (atmospheric metallicity). Depending on the assumed atmospheric metallicity, the observed age and radius of NGTS-11 b matches cooling curves with a different heavy-element fraction. Therefore, the determination of a planet's atmospheric composition clearly improves the planet's characterisation. JWST will further improve the planetary characterisation since white-light curves can constrain the planetary transit depths to very high precision. This will significantly reduce the uncertainties on the planetary radius and therefore on the bulk composition.

However, a key question remains: *What is the link between the atmospheric and the bulk composition?* We know that the solar system gas giants likely have different interior and atmospheric compositions (e.g., Helled, 2018). Characterising the atmospheres of many giant exoplanets will further constrain the internal structure and will reveal how well-mixed they are (Thorngren and Fortney, 2019). It should be noted, however, that the measurements only trace the very upper atmospheres and it is unclear whether and how this composition changes with depth. Finally, since the mass of the outermost atmosphere is very small and the outermost part of the atmosphere is radiative, the measured atmospheric composition could be a result of a recent "pollution", for example, by accreting a small object which vaporised in the atmosphere. It is therefore crucial to better understand under what conditions the atmospheric composition can be linked to the bulk composition.

4.3 Characterisation of direct-imaged planets

Another important application for evolution models is the mass-determination of exoplanets that are detected by direct imaging (e.g., Mordasini et al., 2017). This is achieved by using the planet's thermal emissions and comparing it to evolution models. Direct-imaged exoplanets are commonly younger than 1 Gyr and orbit at large radial distances. This allows probing a completely different regime than the usual mass-radius measurements, and therefore provides additional constraints on planet formation theory. Earlier mass-characterisations used tables of planetary isochrones (e.g., Baraffe et al., 2003). However, in addition to the mass there are many other parameters that influence the planet's luminosity, for example, its composition. It is therefore important to use evolution models that go beyond only accounting for the planetary mass to estimate the mass of direct-imaged exoplanets. In **Figure 2C** we show the luminosity as a function of time for evolution models between $0.5 - 5M_J$ and different bulk metallicites. The observed luminosity and the planet's age are shown in the same figure. It is evident that the inferred mass depends on the assumed composition.

Most evolution models assume a hot-start formation scenario (Baraffe et al., 2003; Marley et al., 2007). While currently hot-starts are the expected formation pathway in the core accretion framework (Berardo and Cumming, 2017; Berardo et al., 2017; Cumming et al., 2018), it is important to note that cold- (Marley et al., 2007; Fortney et al., 2008) and warm-start scenarios would (Spiegel and Burrows, 2012; Marleau and Cumming, 2014) yield significantly different mass estimates.

5 Conclusion

Advanced evolution models are crucial for the characterisation of giant exoplanets. Evolution models provide information of the planetary bulk composition using measurements of mass and radius (with planetary age), connect the atmospheric composition with that of the bulk, and determine the planetary mass of direct-imaged planets. There are theoretical uncertainties associated with the models that can affect the data interpretation, and taking full advantage of future observations requires the determination of the most realistic parameters for evolution models, which is work in progress. Advances in giant exoplanet characterisation are expected with the accurate determination of stellar ages from PLATO and atmospheric measurements from JWST and ARIEL as well as ground-based observations. In particular, future measurements will constrain the bulk heavy-element mass of giant exoplanets and reveal information on the link between the atmospheres and the interiors of giant exoplanets.

Author contributions

SM was the main author of the article and prepared the figures. RH made significant contributions and edits to the text. Both authors approved the article for publication.

Funding

We acknowledge support from Swiss National Science Foundation (SNSF) grant 200020_188460.

Acknowledgments

We thank Daniel Thorngren and an anonymous referee for valuable comments. We acknowledge additional support from the National Centre for Competence in Research "PlanetS" supported by SNSF. This research used data from the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. Extensive use was made of the python packages numpy (Harris et al., 2020), matplotlib (Hunter, 2007) and jupyter (Kluyver et al., 2016).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baraffe, I., Chabrier, G., Barman, T. S., Allard, F., and Hauschildt, P. H. (2003). Evolutionary models for cool Brown dwarfs and extrasolar giant planets. The case of HD 209458. *A&A* 402, 701–712. doi:10.1051/0004-6361:20030252

Baraffe, I., Chabrier, G., and Barman, T. (2008). Structure and evolution of super-Earth to super-Jupiter exoplanets. I. Heavy element enrichment in the interior. *A&A* 482, 315–332. doi:10.1051/0004-6361:20079321

Baraffe, I., Chabrier, G., Fortney, J., and Sotin, C. (2014). "Planetary internal structures," in *Protostars and planets VI*. Editors H. Beuther, R. S. Klessen, C. P. Dullemond, and T. Henning (University of Arizona Press), 763. doi:10.2458/azu_uapress_9780816531240-ch033

Becker, A., Lorenzen, W., Fortney, J. J., Nettelmann, N., Schöttler, M., and Redmer, R. (2014). *Ab initio* equations of state for hydrogen (H-REOS.3) and helium (He-REOS.3) and their implications for the interior of Brown dwarfs. *Astrophysical J. Suppl. Ser.* 215, 21. doi:10.1088/0067-0049/215/2/21

Ben-Yami, M., Madhusudhan, N., Cabot, S. H. C., Constantinou, S., Piette, A., Gandhi, S., et al. (2020). Neutral Cr and V in the atmosphere of ultra-hot jupiter WASP-121 b. *Astrophysical J.* 897, L5. doi:10.3847/2041-8213/ab94aa

Berardo, D., and Cumming, A. (2017). Hot-start giant planets form with radiative interiors. *Astrophysical J.* 846, L17. doi:10.3847/2041-8213/aa81c0

Berardo, D., Cumming, A., and Marleau, G.-D. (2017). The evolution of gas giant entropy during formation by runaway accretion. *Astrophysical J.* 834, 149. doi:10.3847/1538-4357/834/2/149

Bitsch, B., Morbidelli, A., Johansen, A., Lega, E., Lambrechts, M., and Crida, A. (2018). Pebble-isolation mass: Scaling law and implications for the formation of super-Earths and gas giants. *A&A* 612, A30. doi:10.1051/0004-6361/201731931

Böhm-Vitense, E. (1958). Über die Wasserstoffkonvektionszone in Sternen verschiedener Effektivtemperaturen und Leuchtkräfte. Mit 5 Textabbildungen. *Z. für Astrophysi* 46, 108.

Burrows, A. S. (2014). Spectra as windows into exoplanet atmospheres. *Proc. Natl. Acad. Sci.* 111 (35), 12601–12609.

Burrows, A., Hubbard, W. B., Lunine, J. I., and Liebert, J. (2001). The theory of brown dwarfs and extrasolar giant planets. *Rev. Mod. Phys.* 73 (3), 719–765.

Burrows, A., Marley, M., Hubbard, W. B., Lunine, J. I., Guillot, T., and Saumon, D. (1997). A nongray theory of extrasolar giant planets and brown dwarfs. *ApJ* 491 (2), 856–875.

Chabrier, G., Mazevet, S., and Soubiran, F. (2019). A new equation of state for dense hydrogen-helium mixtures. *Astrophysical J.* 872, 51. doi:10.3847/1538-4357/aaf99f

Cumming, A., Helled, R., and Venturini, J. (2018). The primordial entropy of Jupiter. *Mon. Notices R. Astronomical Soc.* 477, 4817–4823. doi:10.1093/mnras/sty1000

Debras, F., and Chabrier, G. (2019). New models of jupiter in the context of juno and galileo. *Astrophysical J.* 872, 100. doi:10.3847/1538-4357/aaff65

Edwards, B., Changeat, Q., Tsiaras, A., Hou Yip, K., Al-Refaie, A. F., Anisman, L., et al. (2022). *Exploring the ability of HST WFC3 G141 to uncover trends in populations of exoplanet atmospheres through a homogeneous transmission survey of 70 gaseous planets. arXiv e-prints* , arXiv:2211.00649:10.48550/arXiv.2211.00649.

Edwards, B., and Tinetti, G. (2022). *The Ariel target list: The impact of TESS and the potential for characterising multiple planets within a system. arXiv e-prints* , arXiv:2205.05073.

Fortney, J. J., Dawson, R. I., and Komacek, T. D. (2021). Hot Jupiters: Origins, structure, atmospheres. *J. Geophys. Res. (Planets)* 126, e06629. doi:10.1029/2020JE006629

Fortney, J. J., Marley, M. S., and Barnes, J. W. (2007). Planetary radii across five orders of magnitude in mass and stellar insolation: Application to transits. *Astrophysical J.* 659, 1661–1672. doi:10.1086/512120

Fortney, J. J., Marley, M. S., Saumon, D., and Lodders, K. (2008). Synthetic spectra and colors of young giant planet atmospheres: Effects of initial conditions and atmospheric metallicity. *Astrophysical J.* 683, 1104–1116. doi:10.1086/589942

Fortney, J. J., and Nettelmann, N. (2010). The interior structure, composition, and evolution of giant planets. *Space Sci. Rev.* 152, 423–447. doi:10.1007/s11214-009-9582-x

Freedman, R. S., Lustig-Yaeger, J., Fortney, J. J., Lupu, R. E., Marley, M. S., and Lodders, K. (2014). Gaseous mean opacities for giant planet and ultracool dwarf atmospheres over a range of metallicities and temperatures. *Astrophysical J. Suppl. Ser.* 214, 25. doi:10.1088/0067-0049/214/2/25

Gardner, J. P., Mather, J. C., Clampin, M., Doyon, R., Greenhouse, M. A., Hammel, H. B., et al. (2006). The James Webb Space telescope. *Space Sci. Rev.* 123, 485–606. doi:10.1007/s11214-006-8315-7

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., Vallenari, A., Babusiaux, C., Bailer-Jones, C. A. L., et al. (2016). The Gaia mission. *A&A* 595, A1. doi:10.1051/0004-6361/201629272

Ginzburg, S., and Chiang, E. (2020). Heavy-metal Jupiters by major mergers: Metallicity versus mass for giant planets. *MNRAS* 498, 680–688. doi:10.1093/mnras/staa2500

Guillot, T. (1999). Interiors of giant planets inside and outside the solar system. *Science* 286, 72–77. doi:10.1126/science.286.5437.72

Guillot, T., and Morel, P. (1995). Cepam: A code for modeling the interiors of giant planets. *A&AS* 109, 109–123.

Guillot, T. (2010). On the radiative equilibrium of irradiated planetary atmospheres. *A&A* 520, A27. doi:10.1051/0004-6361/200913396

Guillot, T., Santos, N. C., Pont, F., Iro, N., Melo, C., and Ribas, I. (2006). A correlation between the heavy element content of transiting extrasolar planets and the metallicity of their parent stars. *A&A* 453, L21–L24. doi:10.1051/0004-6361:20065476

Guillot, T. (2008). The composition of transiting giant extrasolar planets. *Phys. Scr.* 130, 014023. doi:10.1088/0031-8949/2008/T130/014023

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi:10.1038/s41586-020-2649-2

Hasegawa, Y., Bryden, G., Ikoma, M., Vasisht, G., and Swain, M. (2018). The origin of the heavy-element content trend in giant planets via core accretion. *Astrophysical J.* 865, 32. doi:10.3847/1538-4357/aad912

Helled, R., Bodenheimer, P., Podolak, M., Boley, A., Meru, F., Nayakshin, S., et al. (2014). "Giant planet formation, evolution, and internal structure," in *Protostars and planets VI*. Editors H. Beuther, R. S. Klessen, C. P. Dullemond, and T. Henning (University of Arizona Press), 643. doi:10.2458/azu_uapress_9780816531240-ch028

Helled, R., Mazzola, G., and Redmer, R. (2020). Understanding dense hydrogen at planetary conditions. *Nat. Rev. Phys.* 2, 562–574. doi:10.1038/s42254-020-0223-3

Helled, R., and Stevenson, D. (2017). The fuzziness of giant planets' cores. *Astrophysical J.* 840, L4. doi:10.3847/2041-8213/aa6d08

Helled, R. (2018). The interiors of jupiter and Saturn. 175. doi:10.1093/acrefore/9780190647926.013.175

Helled, R., Werner, S., Dorn, C., Guillot, T., Ikoma, M., Ito, Y., et al. (2022). Ariel planetary interiors white paper. *Exp. Astron.* 53, 323–356. doi:10.1007/s10686-021-09739-3

Henyey, L., Vardya, M. S., and Bodenheimer, P. (1965). Studies in stellar evolution. III. The calculation of model envelopes. *Astrophysical J.* 142, 841. doi:10.1086/148357

Howard, S., and Guillot, T. (2023). *Accounting for non-ideal mixing effects in the hydrogen-helium equation of state. arXiv e-prints* , arXiv:2302.07902. doi:10.48550/arXiv.2302.07902

Howard, S., Guillot, T., Bazot, M., Miguel, Y., Stevenson, D. J., Galanti, E., et al. (2023). *Jupiter's interior from Juno: Equation-of-state uncertainties and dilute core extent. arXiv e-prints* , arXiv:2302.09082. doi:10.48550/arXiv.2302.09082

Hubbard, W. B. (1977). The jovian surface condition and cooling rate. *Icarus* 30, 305–310. doi:10.1016/0019-1035(77)90164-6

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* 9, 90–95. doi:10.1109/MCSE.2007.55

Jermyn, A. S., Bauer, E. B., Schwab, J., Farmer, R., Ball, W. H., Bellinger, E. P., et al. (2022). *Modules for Experiments in stellar Astrophysics (MESA): Time-dependent convection, energy conservation, automatic differentiation, and infrastructure. arXiv e-prints* , arXiv:2208.03651.

Johansen, A., and Lambrechts, M. (2017). Forming planets via pebble accretion. *Annu. Rev. Earth Planet. Sci.* 45, 359–387. doi:10.1146/annurev-earth-063016-020226

Kippenhahn, R., Weigert, A., and Weiss, A. (2012). *Stellar structure and evolution*. Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-30304-3

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., et al. (2016). "Jupyter notebooks - a publishing format for reproducible computational workflows," in *Positioning and power in academic publishing: Players, agents and agendas*. Editors F. Loizides, and B. Scmidt (Netherlands: IOS Press), 87–90.

Knierim, H., Shibata, S., and Helled, R. (2022). Constraining the origin of giant exoplanets via elemental abundance measurements. *A&A* 665, L5. doi:10.1051/0004-6361/202244516

Kurokawa, H., and Inutsuka, S.-i. (2015). On the radius anomaly of hot Jupiters: Reexamination of the possibility and impact of layered convection. *Astrophysical J.* 815, 78. doi:10.1088/0004-637X/815/1/78

Leconte, J., and Chabrier, G. (2012). A new vision of giant planet interiors: Impact of double diffusive convection. *A&A* 540, A20. doi:10.1051/0004-6361/201117595

Ledoux, P. (1947). Stellar models with convection and with discontinuity of the mean molecular. *Astrophysical J.* 105, 305. doi:10.1086/144905

Line, M. R., Brogi, M., Bean, J. L., Gandhi, S., Zalesky, J., Parmentier, V., et al. (2021). A solar C/O and sub-solar metallicity in a hot Jupiter atmosphere. *Nature* 598, 580–584. doi:10.1038/s41586-021-03912-6

Lozovsky, M., Helled, R., Rosenberg, E. D., and Bodenheimer, P. (2017). Jupiter's formation and its primordial internal structure. *Astrophysical J.* 836, 227. doi:10.3847/1538-4357/836/2/227

Madhusudhan, N., Crouzet, N., McCullough, P. R., Deming, D., and Hedges, C. (2014). H_2O abundances in the atmospheres of three hot Jupiters. *Astrophysical J.* 791, L9. doi:10.1088/2041-8205/791/1/L9

Malsky, I., and Rogers, L. A. (2020). *Coupled thermal and compositional evolution of photo evaporating planet envelopes*. *arXiv e-prints* , arXiv:2002.06466.

Mankovich, C., Fortney, J. J., and Moore, K. L. (2016). Bayesian evolution models for jupiter with helium rain and double-diffusive convection. *Astrophysical J.* 832, 113. doi:10.3847/0004-637X/832/2/113

Mankovich, C., and Fuller, J. (2021). *A diffuse core in Saturn revealed by ring seismology*. *arXiv e-prints* , arXiv:2104.13385.

Marleau, G. D., and Cumming, A. (2014). Constraining the initial entropy of directly detected exoplanets. *MNRAS* 437, 1378–1399. doi:10.1093/mnras/stt1967

Marley, M. S., Fortney, J. J., Hubickyj, O., Bodenheimer, P., and Lissauer, J. J. (2007). On the luminosity of young Jupiters. *Astrophysical J.* 655, 541–549. doi:10.1086/509759

Mayor, M., and Queloz, D. (1995). A Jupiter-mass companion to a solar-type star. *Nature* 378, 355–359. doi:10.1038/378355a0

Mazevet, S., Licari, A., Chabrier, G., and Potekhin, A. Y. (2019). *Ab initio* based equation of state of dense water for planetary and exoplanetary modeling. *A&A* 621, A128. doi:10.1051/0004-6361/201833963

Militzer, B., and Hubbard, W. B. (2013). *Ab initio* equation of state for hydrogen-helium mixtures with recalibration of the giant-planet mass-radius relation. *Astrophysical J.* 774, 148. doi:10.1088/0004-637X/774/2/148

Miller, N., and Fortney, J. J. (2011). The heavy-element masses of extrasolar giant planets, revealed. *Astrophysical J.* 736, L29. doi:10.1088/2041-8205/736/2/L29

Mordasini, C. (2014). Grain opacity and the bulk composition of extrasolar planets: II. An analytical model for grain opacity in protoplanetary atmospheres⋆. *A&A* 572, A118. doi:10.1051/0004-6361/201423702

Mordasini, C., Klahr, H., Alibert, Y., Miller, N., and Henning, T. (2014). Grain opacity and the bulk composition of extrasolar planets: I. Results from scaling the ISM opacity⋆. *A&A* 566, A141. doi:10.1051/0004-6361/201321479

Mordasini, C., Marleau, G. D., and Mollière, P. (2017). Characterization of exoplanets from their formation. III. The statistics of planetary luminosities. *A&A* 608, A72. doi:10.1051/0004-6361/201630077

More, R. M., Warren, K. H., Young, D. A., and Zimmerman, G. B. (1988). A new quotidian equation of state (QEOS) for hot dense matter. *Phys. Fluids* 31, 3059–3078. doi:10.1063/1.866963

Mousis, O., Marboeuf, U., Lunine, J. I., Alibert, Y., Fletcher, L. N., Orton, G. S., et al. (2009). Determination of the minimum masses of heavy elements in the envelopes of jupiter and Saturn. *Astrophysical J.* 696, 1348–1354. doi:10.1088/0004-637X/696/2/1348

Müller, S., Ben-Yami, M., and Helled, R. (2020a). Theoretical versus observational uncertainties: Composition of giant exoplanets. *Astrophysical J.* 903, 147. doi:10.3847/1538-4357/abba19

Müller, S., Helled, R., and Cumming, A. (2020b). The challenge of forming a fuzzy core in Jupiter. *A&A* 638, A121. doi:10.1051/0004-6361/201937376

Müller, S., and Helled, R. (2021). Synthetic evolution tracks of giant planets. *Mon. Notices R. Astronomical Soc.* 507, 2094–2102. doi:10.1093/mnras/stab2250

Müller, S., and Helled, R. (2023). Towards a new era in giant exoplanet characterisation. *A&A* 669, A24. doi:10.1051/0004-6361/202244827

Parmentier, V., and Guillot, T. (2014). A non-grey analytical model for irradiated atmospheres. I. Derivation. *A&A* 562, A133. doi:10.1051/0004-6361/201322342

Paxton, B., Bildsten, L., Dotter, A., Herwig, F., Lesaffre, P., and Timmes, F. (2011). Modules for Experiments in stellar Astrophysics (MESA). *Astrophysical J. Suppl. Ser.* 192, 3. doi:10.1088/0067-0049/192/1/3

Paxton, B., Cantiello, M., Arras, P., Bildsten, L., Brown, E. F., Dotter, A., et al. (2013). Modules for Experiments in stellar Astrophysics (MESA): Planets, oscillations, rotation, and massive stars. *Astrophysical J. Suppl. Ser.* 208, 4. doi:10.1088/0067-0049/208/1/4

Paxton, B., Marchant, P., Schwab, J., Bauer, E. B., Bildsten, L., Cantiello, M., et al. (2015). Modules for Experiments in stellar Astrophysics (MESA): Binaries, pulsations, and explosions. *Astrophysical J. Suppl. Ser.* 220, 15. doi:10.1088/0067-0049/220/1/15

Paxton, B., Schwab, J., Bauer, E. B., Bildsten, L., Blinnikov, S., Duffell, P., et al. (2018). Modules for Experiments in stellar Astrophysics (MESA): Convective boundaries, element diffusion, and massive star explosions. *Astrophysical J. Suppl. Ser.* 234, 34. doi:10.3847/1538-4365/aaa5a8

Paxton, B., Smolec, R., Schwab, J., Gautschy, A., Bildsten, L., Cantiello, M., et al. (2019). Modules for Experiments in stellar Astrophysics (MESA): Pulsating variable stars, rotation, convective boundaries, and energy conservation. *Astrophysical J. Suppl. Ser.* 243, 10. doi:10.3847/1538-4365/ab2241

Pollack, J. B., Hubickyj, O., Bodenheimer, P., Lissauer, J. J., Podolak, M., and Greenzweig, Y. (1996). formation of the giant planets by concurrent accretion of solids and gas. *Icarus* 124, 62–85. doi:10.1006/icar.1996.0190

Poser, A. J., Nettelmann, N., and Redmer, R. (2019). The effect of clouds as an additional opacity source on the inferred metallicity of giant exoplanets. *Atmosphere* 10, 664. doi:10.3390/atmos10110664

Radko, T., Bulters, A., Flanagan, J. D., and Campin, J. M. (2014). Double-diffusive recipes. Part I: Large-scale dynamics of thermohaline staircases. *J. Phys. Oceanogr.* 44, 1269–1284. doi:10.1175/JPO-D-13-0155.1

Rauer, H., Catala, C., Aerts, C., Appourchaux, T., Benz, W., Brandeker, A., et al. (2014). The PLATO 2.0 mission. *Exp. Astron.* 38, 249–330. doi:10.1007/s10686-014-9383-4

Rustamkulov, Z., Sing, D. K., Mukherjee, S., May, E. M., Kirk, J., Schlawin, E., et al. (2022). *Early Release science of the exoplanet WASP-39b with JWST NIRSpec PRISM*. *arXiv e-prints* , arXiv:2211.10487:10.48550/arXiv.2211.10487.

Sarkis, P., Mordasini, C., Henning, T., Marleau, G. D., and Mollière, P. (2021). Evidence of three mechanisms explaining the radius anomaly of hot Jupiters. *A&A* 645, A79. doi:10.1051/0004-6361/202038361

Saumon, D., Chabrier, G., and van Horn, H. M. (1995). An equation of state for low-mass stars and giant planets. *Astrophysical J.* 99, 713. doi:10.1086/192204

Shibata, S., Helled, R., and Ikoma, M. (2020). The origin of the high metallicity of close-in giant exoplanets. Combined effects of resonant and aerodynamic shepherding. *A&A* 633, A33. doi:10.1051/0004-6361/201936700

Shibata, S., and Ikoma, M. (2019). Capture of solids by growing proto-gas giants: Effects of gap formation and supply limited growth. *Mon. Notices R. Astronomical Soc.* 487, 4510–4524. doi:10.1093/mnras/stz1629

Spiegel, D. S., and Burrows, A. (2012). Spectral and photometric diagnostics of giant planet formation scenarios. *Astrophysical J.* 745, 174. doi:10.1088/0004-637X/745/2/174

Stevenson, D. J., Bodenheimer, P., Lissauer, J. J., and D'Angelo, G. (2022). Mixing of condensable constituents with H-He during the formation and evolution of jupiter. *PSJ* 3, 74. doi:10.3847/PSJ/ac5c44

The JWST Transiting Exoplanet Community Early Release Science Team, Ahrer, E.-M., Alderson, L., Batalha, N. M., Batalha, N. E., Bean, J. L., et al. (2022). *Identification of carbon dioxide in an exoplanet atmosphere*. *arXiv e-prints*, arXiv:2208.11692. doi:10.48550/arXiv.2208.11692

Teske, J. K., Thorngren, D., Fortney, J. J., Hinkel, N., and Brewer, J. M. (2019). Do metal-rich stars make metal-rich planets? New insights on giant planet formation from host star abundances. *AJ* 158, 239. doi:10.3847/1538-3881/ab4f79

Thompson, S. L. (1990). *ANEOS analytic equations of state for shock physics codes input manual*. doi:10.2172/6939284

Thorngren, D., and Fortney, J. J. (2019). Connecting giant planet atmosphere and interior modeling: Constraints on atmospheric metal enrichment. *Astrophysical J.* 874, L31. doi:10.3847/2041-8213/ab1137

Thorngren, D. P., Fortney, J. J., Murray-Clay, R. A., and Lopez, E. D. (2016). The mass-metallicity relation for giant planets. *Astrophysical J.* 831, 64. doi:10.3847/0004-637X/831/1/64

Tinetti, G., Drossart, P., Eccleston, P., Hartogh, P., Heske, A., Leconte, J., et al. (2018). A chemical survey of exoplanets with ARIEL. *Exp. Astron.* 46, 135–209. doi:10.1007/s10686-018-9598-x

Turrini, D., Schisano, E., Fonte, S., Molinari, S., Politi, R., Fedele, D., et al. (2021). Tracing the formation history of giant planets in protoplanetary disks with carbon, oxygen, nitrogen, and sulfur. *Astrophysical J.* 909, 40. doi:10.3847/1538-4357/abd6e5

Valletta, C., and Helled, R. (2020). Giant planet formation models with a self-consistent treatment of the heavy elements. *Astrophysical J.* 900, 133. doi:10.3847/1538-4357/aba904

Valsecchi, F., Rappaport, S., Rasio, F. A., Marchant, P., and Rogers, L. A. (2015). Tidally-driven roche-lobe overflow of hot Jupiters with MESA. *Astrophysical J.* 813, 101. doi:10.1088/0004-637X/813/2/101

Vazan, A., Helled, R., Kovetz, A., and Podolak, M. (2015). Convection and mixing in giant planet evolution. *Astrophysical J.* 803, 32. doi:10.1088/0004-637X/803/1/32

Vazan, A., Kovetz, A., Podolak, M., and Helled, R. (2013). The effect of composition on the evolution of giant and intermediate-mass planets. *Mon. Notices R. Astronomical Soc.* 434, 3283–3292. doi:10.1093/mnras/stt1248

Weiss, L. M., Marcy, G. W., Rowe, J. F., Howard, A. W., Isaacson, H., Fortney, J. J., et al. (2013). The mass of KOI-94d and a relation for planet radius, mass, and incident flux. *Astrophysical J.* 768, 14. doi:10.1088/0004-637X/768/1/14

Welbanks, L., Madhusudhan, N., Allard, N. F., Hubeny, I., Spiegelman, F., and Leininger, T. (2019). Mass-metallicity trends in transiting exoplanets from atmospheric abundances of H_2O, Na, and K. *Astrophysical J.* 887, L20. doi:10.3847/2041-8213/ab5a89

Wood, T. S., Garaud, P., and Stellmach, S. (2013). A new model for mixing by double-diffusive convection (Semi-convection). II. The transport of heat and composition through layers. *Astrophysical J.* 768, 157. doi:10.1088/0004-637X/768/2/157

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 15 June 2023
DOI 10.3389/fspas.2023.1143909

Check for updates

OPEN ACCESS

EDITED BY
Joseph E. Borovsky,
Space Science Institute, United States

REVIEWED BY
Jingnan Guo,
University of Science and Technology of China, China
Jeffrey Linsky,
University of Colorado Boulder, United States

*CORRESPONDENCE
Merav Opher,
mopher@bu.edu

RECEIVED 13 January 2023
ACCEPTED 10 May 2023
PUBLISHED 15 June 2023

CITATION
Opher M, Richardson J, Zank G, Florinski V, Giacalone J, Sokół JM, Toth G, Buxner S, Kornbleuth M, Gkioulidou M, Nikoukar R, Van der Holst B, Turner D, Gross N, Drake J, Swisdak M, Dialynas K, Dayeh M, Chen Y, Zieger B, Powell E, Onubogu C, Ma X, Bair E, Elliott H, Galli A, Zhao L, Adhikari L, Nakanotani M, Hill ME, Mostafavi P, Du S, Guo F, Reisenfeld D, Fuselier S, Izmodenov V, Baliukin I, Cummings A, Miller J, Wang B, Ghanbari K, Kota J, Loeb A, Burgess J, Hokanson SC, Morrow C, Hong A and Boldon A (2023), Solar wind with Hydrogen Ion charge Exchange and Large-Scale Dynamics (SHIELD) DRIVE Science Center. *Front. Astron. Space Sci.* 10:1143909. doi: 10.3389/fspas.2023.1143909

COPYRIGHT
© 2023 Opher, Richardson, Zank, Florinski, Giacalone, Sokół, Toth, Buxner, Kornbleuth, Gkioulidou, Nikoukar, Van der Holst, Turner, Gross, Drake, Swisdak, Dialynas, Dayeh, Chen, Zieger, Powell, Onubogu, Ma, Bair, Elliott, Galli, Zhao, Adhikari, Nakanotani, Hill, Mostafavi, Du, Guo, Reisenfeld, Fuselier, Izmodenov, Baliukin, Cummings, Miller, Wang, Ghanbari, Kota, Loeb, Burgess, Hokanson, Morrow, Hong and Boldon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Solar wind with Hydrogen Ion charge Exchange and Large-Scale Dynamics (SHIELD) DRIVE Science Center

Merav Opher[1,2]*, John Richardson[3], Gary Zank[4], Vladimir Florinski[4], Joe Giacalone[5], Justyna M. Sokół[6], Gabor Toth[7], Sanlyn Buxner[8], Marc Kornbleuth[1], Matina Gkioulidou[9], Romina Nikoukar[9], Bart Van der Holst[7], Drew Turner[9], Nicholas Gross[1], James Drake[10], Marc Swisdak[10], Kostas Dialynas[11], Maher Dayeh[6], Yuxi Chen[1], Bertalan Zieger[1], Erick Powell[1], Chika Onubogu[1], Xiaohan Ma[1], Ethan Bair[1], Heather Elliott[6], Andre Galli[12], Lingling Zhao[4], Laxman Adhikari[4], Masaru Nakanotani[4], Matthew E. Hill[9], Parisa Mostafavi[9], Senbei Du[13], Fan Guo[13], Daniel Reisenfeld[13], Stephen Fuselier[6], Vladislav Izmodenov[14], Igor Baliukin[15], Alan Cummings[16], Jesse Miller[1], Bingbing Wang[4], Keyvan Ghanbari[4], Jozsef Kota[5], Abraham Loeb[17], Juditra Burgess[1], Sarah Chobot Hokanson[18], Cherilyn Morrow[19], Adam Hong[20] and Andrea Boldon[21]

[1]Astronomy Department, Boston University, Boston, MA, United States, [2]Radcliffe Institute for Advanced Studies at Harvard University, Cambridge, MA, United States, [3]Kavli Center for Astrophysics and Space Science, Massachusetts Institute of Technology, Cambridge, MA, United States, [4]Department of Space Science and Center for Space Plasma and Aeronomic Research (CSPAR), University of Alabama in Huntsville, Huntsville, AL, United States, [5]Lunar and Planetary Laboratory, University of Arizona, Tucson, AZ, United States, [6]Southwest Research Institute, San Antonio, TX, United States, [7]Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MA, United States, [8]Planetary Science Institute, Tucson, AZ, United States, [9]Applied Physics Lab, John Hopkins University, Laurel, MD, United States, [10]Department of Physics, University of Maryland, College Park, MD, United States, [11]Office of Space Research and Technology, Academy of Athens, Athens, Greece, [12]Space Research and Planetary Sciences Division, University of Bern, Bern, Switzerland, [13]Los Alamos National Laboratory, Los Alamos, NM, United States, [14]Lomonosov Moscow State University, Moscow Center for Fundamental and Applied Mathematics, GSP-1, Leninskie Gory, Moscow, Russia, [15]Space Research Institute (IKI) of Russian Academy of Sciences, Moscow, Russia, [16]California Institute of Technology, Pasadena, CA, United States, [17]Institute of Theory and Computation, Harvard University, Cambridge, MA, United States, [18]Office of Provost, Boston University, Boston, MA, United States, [19]Consultant SWRI, Boulder, CO, United States, [20]Adam Hong Graphic Design, New York, NY, United States, [21]The Implementation Group (TIG), Washington, DC, United States

Most stars generate winds and move through the interstellar medium that surrounds them. This movement creates a cocoon formed by the deflection

Abbreviations: ACR, Anomalous cosmic ray; BI, Broadening Impacts; BP, Broadening Participation; BSD, Bench Strength Development; CD, Center Director; CE, Charge exchange; CR, Cosmic ray; EAB, External Advisory Board; EC, Executive Committee; GCR, Galactic cosmic rays; HP, Heliopause; HS, Heliosheath; IHA, The Institute for Habitable Astrospheres; IMAP, Interstellar Mapping and Acceleration Probe; LISM, Local interstellar medium; LECP, Low energy charge particle; NH, New Horizons; OH, Outer Heliosphere; PM, Project Manager; PI, Principal Investigator; PSAB, Postdoc and Student Advisory Board; PLS, Plasma science experiment; RT, Research Thrusts; TS, Termination Shock; URM, Under-represented Minority; V1, Voyager 1; V2, Voyager 2.

of these winds that envelops and protects the stars. We call these "cocoons" astrospheres. The Sun has its own cocoon, the heliosphere. The heliosphere is an immense shield that protects the Solar System from harsh, galactic radiation. The radiation that enters the heliosphere affects life on Earth as well as human space exploration. Galactic cosmic rays are the dominant source of radiation and principal hazard affecting space missions within our Solar System. Current global heliosphere models do not successfully predict the radiation environment at all locations or under different solar conditions. To understand the heliosphere's shielding properties, we need to understand its structure and large-scale dynamics. A fortunate confluence of missions has provided the scientific community with a treasury of heliospheric data. However, fundamental features remain unknown. The vision of the Solar wind with Hydrogen Ion charge Exchange and Large-Scale Dynamics (SHIELD) DRIVE Science Center is to understand the nature and structure of the heliosphere. Through four integrated research thrusts leading to the global model, SHIELD will: 1) determine the global nature of the heliosphere; 2) determine how pickup ions evolve from "cradle to grave" and affect heliospheric processes; 3) establish how the heliosphere interacts with and influences the Local Interstellar Medium (LISM); and 4) establish how cosmic rays are filtered by and transported through the heliosphere. The key deliverable is a comprehensive, self-consistent, global model of the heliosphere that explains data from all relevant *in situ* and remote observations and predicts the radiation environment. SHIELD will develop a "digital twin" of the heliosphere capable of: (a) predicting how changing solar and LISM conditions affect life on Earth, (b) understanding the radiation environment to support long-duration space travel, and (c) contributing toward finding life elsewhere in the Galaxy. SHIELD also will train the next-generation of heliophysicists, a diverse community fluent in team science and skilled working in highly transdisciplinary collaborative environments.

KEYWORDS

heliosphere, space physics, solar wind, interstellar medium, magnetic field, galactic cosmic ray (GCR)

1 Introduction

The vision of the Solar wind with Hydrogen Ion charge Exchange and Large-Scale Dynamics (SHIELD) DRIVE Science Center is to understand the nature and structure of the heliosphere. To realize our vision and create a new model for the heliosphere, SHIELD will enlist experts on observations, kinetic physics and MHD physics, and high energy particle transport and acceleration physics, and will facilitate their interaction in a Center environment. SHIELD will reach to the broader community for participation. Through four integrated research thrusts leading to the global model, SHIELD will: 1) determine the global nature of the heliosphere; 2) determine how pickup ions evolve from "cradle to grave" and affect heliospheric processes; 3) establish how the heliosphere interacts with and influences the LISM; and 4) establish how cosmic rays are filtered by and transported through the heliosphere. (Instead of LISM VLISM has been also coined to describe the region beyond the HP where interstellar material and fields are modified by the Sun).

The key deliverable is a comprehensive, self-consistent, global model of the heliosphere that explains data from all relevant *in situ* and remote observations and predicts the radiation environment. SHIELD will develop a "digital twin" of the heliosphere capable of: (a) predicting how changing solar and LISM conditions affect life on Earth, (b) understanding the radiation environment to support long-duration space travel, and (c) contributing toward finding life elsewhere in the Galaxy. SHIELD also will train the next-generation of heliophysicists, a diverse community fluent in team science and skilled working in highly transdisciplinary collaborative environments.

SHIELD is one of three DRIVE centers chosen for Phase II in 2022 after a two-year competition between 9 DRIVE centers selected for Phase I. The DRIVE centers came from the recommendation of the Decadal Survey 2013–2023 Solar and Space Physics "A Science for a Technological Society", NRC 2012. As part of the (DRIVE—Diversify, Realize, Integrate, Venture, Educate) a critical element was the funding of Heliophysics DRIVE Science Centers.

Heliophysics Phase II DRIVE (Diversify, Realize, Integrate, Venture, Educate) Science Centers (DSCs) are intended to support science that cannot be effectively done by individual investigators or small teams, but requires the synergistic, coordinated efforts of a research center. The DRIVE Centers "must address grand challenge science goals that are both ambitious and focused enough to be achievable within the lifetime of the center [5 years-(2022–2027)]–in other words, science problems poised and ready for major advances."

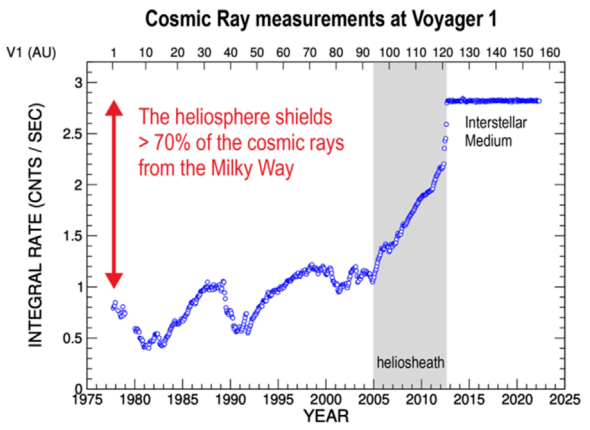


FIGURE 1
The heliosphere shields 70% of the harsh galactic radiation. Global models are far from being able to predict the radiation environment in all directions and conditions. SHIELD will provide for the community improved radiation prediction capabilities. The blue curve shows the integral counting rate of galactic cosmic rays from Voyager 1 (for more information see Cummings et al., 2016). The gray shading indicates the region between the solar termination shock and the heliopause.

1.1 Knowledge gaps

Stars are not inanimate luminous structures in the night sky. Most stars generate winds and move through the interstellar medium that surrounds them. This movement creates a cocoon formed by the deflection of these winds that envelops and protects the stars. We call these "cocoons" astrospheres.

The Sun has its own cocoon, the heliosphere: The heliosphere is an immense shield that protects the Solar System from harsh, galactic radiation. When Voyager 1 (V1) crossed the heliopause (HP), it discovered that the heliosphere shields 70% of galactic cosmic rays (GCRs) from Earth (**Figure 1**). The radiation that enters the heliosphere affects life on Earth as well as human space exploration. GCRs are the dominant source of radiation and the principal hazard (Elgart, 2021) affecting long-term space missions within our Solar System. Current global heliosphere models do not successfully predict the radiation environment at all locations or under different solar conditions. To understand the heliosphere's shielding properties, we need to understand its structure and large-scale dynamics.

The heliosphere is formed by: 1) the solar wind; 2) the ionized local interstellar medium (LISM) that surrounds and confines the heliosphere; 3) LISM neutrals that enter the heliosphere, become ionized and join the solar wind as hot pick-up ions (PUIs); and 4) GCRs that stream from the Galaxy and penetrate the heliosphere. A fortunate confluence of missions has provided the scientific community with a treasury of heliospheric data. The *in situ* measurements by the Voyager and New Horizons (NH) spacecraft, combined with the all-sky Energetic Neutral Atoms (ENA) images of the heliospheric boundary region by the Interstellar Boundary Explorer (IBEX) and Cassini missions, have revealed the size of the heliosphere in the nose region (V1 and V2), the key role PUIs play in the solar wind and heliosheath (V2, NH), and the existence of the IBEX ribbon and Cassini belt.

However, fundamental features remain unknown. Knowledge gaps include the basic shape of the heliosphere (**Figure 2**), the extent of its tail, the flow pattern and the width of the heliosheath, and the extent of the heliosphere's influence on the LISM. Current global models fail to explain major observations (Box 1), a key driver of the SHIELD Drive Science Center.

Global models (e.g., Izmodenov & Alexashov, 2015; Opher et al., 2015; Pogorelov et al., 2015) have grown in sophistication and accelerated our understanding of the heliosphere. But they fail to explain many observations. Models do not integrate the influences of suprathermal particles, reconnection, turbulence phenomena, and other microprocesses. There are models of interplanetary shocks, of the termination shock, of cosmic ray (CR) transport, of turbulence, and of the heliosheath and heliosphere. However, no global model includes all the necessary physics. Microprocesses have spatial scales vastly separate from the global scales but are critical pieces of a heliospheric model and must be parameterized. The transport of CRs and the evolution of PUIs require kinetic transport models that are coupled self-consistently to a global MHD model. The next quantum leap in understanding the heliosphere will be realized by synthesizing knowledge in theory, modeling, and observations to create a comprehensive, self-consistent global model that connects all the physical processes that effect the heliosphere as a one system.

1.2 Vision

The vision of the Solar wind with Hydrogen Ion charge Exchange and Large-Scale Dynamics (SHIELD) DRIVE Science Center (DSC) is to understand the nature and structure of the heliosphere. SHIELD is built around four coupled research thrusts (RTs), each addressing a major science question required to create a novel, global heliosphere model. They are: 1) What is the global structure of the heliosphere?; 2) How do pickup ions evolve from cradle to grave, and how do they affect heliospheric processes?; 3) How does the heliosphere interact with and influence the interstellar medium?; and 4) How are cosmic rays filtered by and transported through the heliosphere? The key deliverable is a comprehensive, self-consistent, global model of the heliosphere that explains data from all relevant *in situ* and remote observations and predicts the radiation environment. SHIELD's broadening impact goal is to attract, train and mentor the next-generation of team-based heliophysicists while accelerating knowledge integration, transfer, and communication across traditional boundaries.

SHIELD's model will include all components of the heliosphere—thermal and suprathermal components of the solar wind, PUIs, ACRs, GCRs, and interstellar neutrals (**Figure 4**). In addition, the model will define, quantify, and simulate all relevant processes that affect these components, from microphysical processes such as reconnection and turbulence to PUI and cosmic ray transport and acceleration, with physics guided by theory and models that are constrained by observations. SHIELD will develop a "digital twin" of the heliosphere capable of: (a) predicting how changing solar and LISM conditions affect life on Earth, (b) understanding the radiation environment to support long-duration space travel, and (c) contributing toward finding life elsewhere in the Galaxy.

BOX 1 Critical Observations to be Explained by SHIELD

- The acceleration mechanism and location for anomalous cosmic rays (ACRs) are not known. The Voyager spacecraft did not find evidence for acceleration at the termination shock (TS) but detected an increase in ACR intensities as they moved across the heliosheath (HS). *SHIELD* will determine where and how ACRs are accelerated.
- The HS is 30%–50% thinner than current models predict. SHIELD will determine physical mechanisms responsible for extracting energy from the HS.
- The flow patterns (Figure 3) determined by plasma and energetic particles instruments are drastically different at V1 and V2. *SHIELD* will reconcile the flow observations from different instruments with those derived from models.
- The magnetic field direction did not change at the heliopause (HP) crossings, a major surprise, and the LISM field direction has changed little since the HP. *SHIELD* will determine the extent of the influence of the heliosphere on the *LISM*.
- The significant increase in GCRs just prior to the HP at V1 and V2 and the unexpected GCR anisotropies observed in the LISM are not understood. *SHIELD* will determine the porosity of the HP and the shielding of *GCRs* by the *HP* in all directions. *SHIELD* also will determine the source of *GCR* anisotropies in the *LISM*.
- Proposed heliosphere shapes range from comet-like to "spherical" to "croissant-like" with a truncated heliotail. *SHIELD* will determine the physical drivers for the shape of the heliosphere and the length and structure of the heliotail.
- Models predict ENA intensities only half as large as those observed by IBEX. *SHIELD* will determine the source and distribution of ions in the heliosphere to explain *IBEX* data and prepare for *ENA* maps that will be produced by *IMAP*.
- IBEX detected a global feature, the ribbon, organized by the interstellar magnetic field (B_{ISM}). Cassini measured a similar, broader "belt" feature. Models proposed to explain these features rely on assumptions such as the draping of the B_{ISM} and the level of turbulence in the LISM. *SHIELD* will determine the source of the turbulence in the LISM and the direction of the interstellar magnetic field, which will help establish the origin of the ribbon and belt.

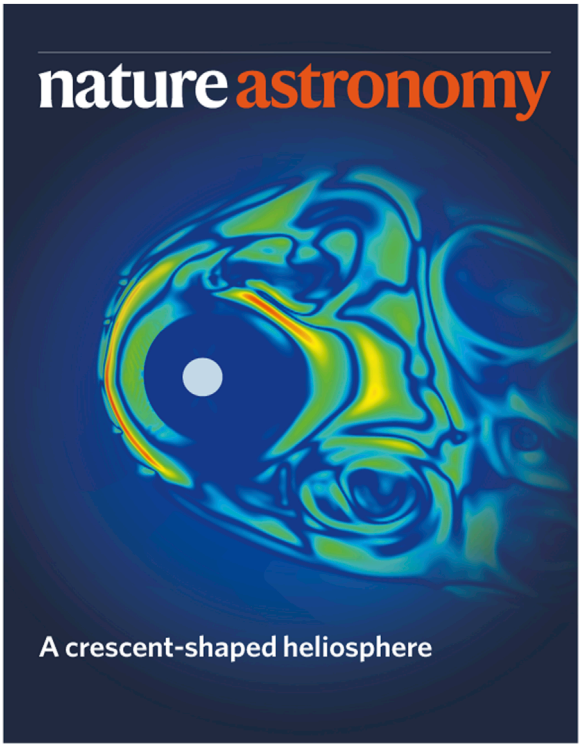


FIGURE 2
Fundamental features of the heliosphere are not well understood (Opher et al., 2020). SHIELD will determine the main physical drivers contributing to its shape.

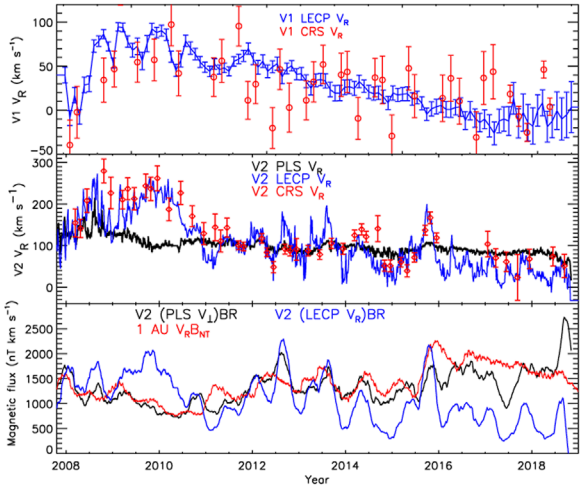
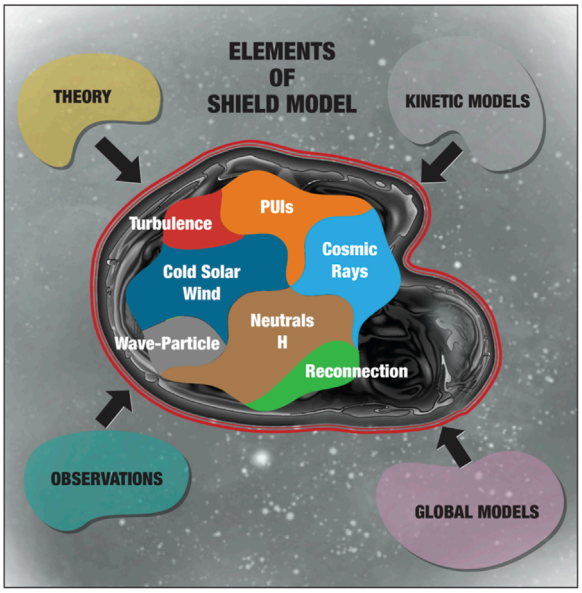


FIGURE 3
The drastically different HS flows at V1 and V2 are a source of current debate. SHIELD will address why the flow speeds differ each other and from model predictions (Richardson et al., 2021).

1.3 Potential for Transformative Impact

SHIELD research will inform the shape, size, and structure of the heliosphere, closing long-standing knowledge gaps. In addition to predicting how changing solar and LISM conditions affect the heliospheric radiation environment in support of long-duration space travel, SHIELD's impact is expected to be felt in neighboring scientific communities. The heliosphere is a template for other astrospheres and is the only astrosphere where we can study *in situ* the processes that determine its structure and radiation environment. Knowledge gained from SHIELD can be extended to other astrospheres and to other times in the Solar System's history. For example, when life started on Earth, the solar wind had different characteristics and the Sun was embedded in a different interstellar cloud. SHIELD will help predict the radiation environment at exoplanets in



SHIELD combines theory, modeling (global MHD & localized kinetic), observations to build a comprehensive model of the heliosphere. These elements are included in all research thrusts of SHIELD.

FIGURE 4
The SHIELD Model.

other astrospheres and throughout our Solar System's history. Through the Institute for Habitable Astrospheres (IHA, https://shielddrivecenter.com/iha/), our sustainability initiative, SHIELD will collaborate cross-disciplinarily with astronomers, biologists, and chemists to determine questions and challenges concerning the origin of life that can be solved by combining expertise in radiation, biology, astrobiology, chemistry, and heliophysics.

There is a time urgency: IBEX and Cassini provide ENA data, and, in 2025, NASA will launch Interstellar Mapping and Acceleration Probe (IMAP) (McComas et al., 2018) to make high-resolution ENA measurements. Since ENA observations are line-of-sight images, a global model like SHIELD is required to pin down the sources of the ENAs and interpret the data. The synergy of a global and comprehensive model such as SHIELD with ENA data will transform our view of the heliosphere. Voyager has made the only *in situ* observations of the TS, HS, HP, and the LISM thus far. Observers from Voyager, several of whom participate in SHIELD, will benchmark our global model and pass their knowledge of the data to SHIELD's young investigators, allowing Voyager's legacy to continue.

SHIELD's work responds directly to the Solar and Space Physics Decadal Survey's (Solar and Space Physics, NRC, 2013) Science Goal 3, determine the interaction of the Sun with the Solar System and the interstellar medium. We investigate the fundamental processes of turbulence, reconnection, and wave-particle interactions that occur both within the heliosphere and throughout the Universe (Decadal Survey Science Goal 4). The coupling of microprocesses and global codes will explain key observations (Box 1) and thus transform the field of heliophysics. **Figure 5** shows that the goals and research thrusts (RTs) of SHIELD are tightly interwoven, justifying the need for a Center. (See **Section 5**, *Justification for DSC Mode of Support*.)

1.4 Synergy within the SHIELD science team

SHIELD has gathered more than 45 experts with diverse viewpoints on current heliospheric puzzles (See **Table 1** in **Section 7.3**). Our team includes experts specializing in all key physical processes that play a role in the distant solar wind, HS and LISM, including turbulence, reconnection, wave-particle interactions, PUI and GCR acceleration and transport, global MHD codes, ENA modeling, and localized kinetic models. To validate the SHIELD model, the team includes experts on *in situ* and remote observations from all the relevant instruments on Voyager, IBEX, Cassini, New Horizons, and IMAP. An established NASA science educator with expertise in literary, public outreach, communication, and evaluation leads the Broader Impacts activities. Knowledge integration is assured via SHIELD's management approach, emphasizing shared decision making, team science, collaboration, robust communication as well as leadership and accountability of all team members (**Section 8**, Management.)

Solving SHIELD's science questions requires an integrated approach. An example of how our experts in theory, modeling, and data analysis worked together in Phase I (when 30 original DSC were downsized in a period of 2 years to the final 3 awarded for a full DSC) is demonstrative of our synergy. The PUI team (RT2) used multiple models to derive how PUI spectra evolve at the TS. We first modeled the effect of turbulence on the supersonic solar wind (Nakanotani et al., 2021) (Nakanotani; throughout the text we added the names of people from the center that will be involved in each task-see **Table 1**) constrained by NH observations (Elliott). This solar wind solution was input into kinetic models [hybrid (Giacalone), PIC (Drake, Swisdak), and multi-ion (Zieger)] to find how reconnection and turbulence modify the ion spectra downstream of the TS. We validated these model results with Voyager observations (Richardson, Hill). A consensus PUI spectra was derived and used to model ENAs at IBEX (Shrestha et al., 2021) and Cassini energies (Gkioulidou et al., 2022) (Kornbleuth, Baliukin) using global MHD models (Opher). These model results were compared with ENA observations (Dialynas, Dayeh, Fuselier, Gkioulidou). This work, which shows the integration of the team, quantified the additional PUI acceleration that must occur to match ENA data and sets the stage for our Phase II work (see **Section 2.2.2**, RT2).

2 SHIELD research plan

2.1 Research progress in Phase I

SHIELD made substantial progress on its science questions in Phase I, including several publications mentioned below. **Table 2** profiles the key research accomplishments for each thrust and implications for Phase II.

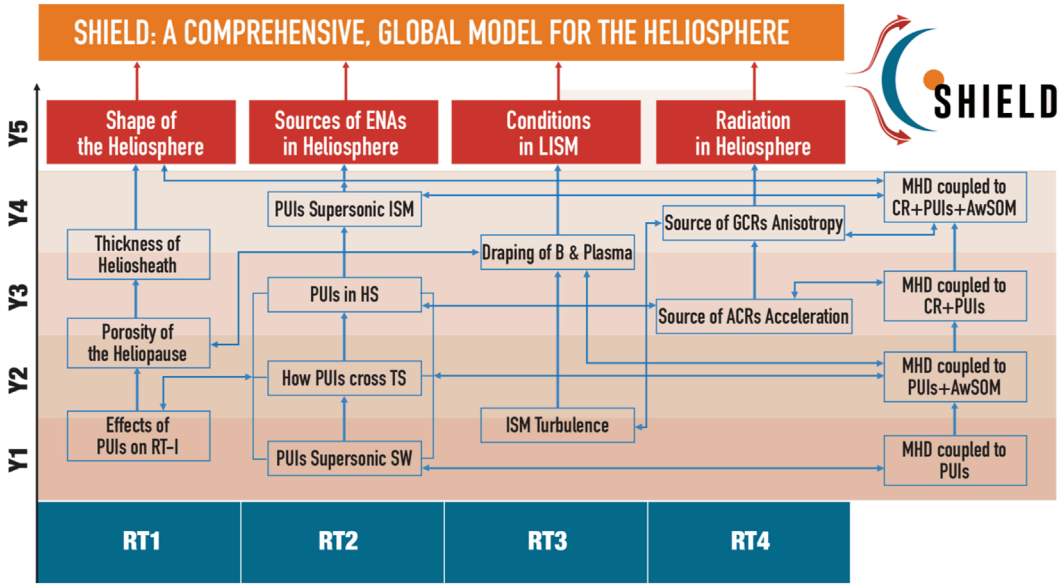


FIGURE 5
Each science topic leads to Center-wide outcomes (red boxes) and toward the building of SHIELD's comprehensive, self-consistent global model of the heliosphere. Years 1–5 corresponds to 5 years of Phase II of SHIELD.

2.2 Research proposed in Phase II

SHIELD will unravel the main physical processes that drive the heliosphere, assembling them into a novel, global model of the heliosphere. It will inform how the heliosphere shields Earth from galactic cosmic rays and support future work on the habitability of other worlds. To fulfill the promise of this vision, we face a set of knowledge integration challenges: 1) modeling huge expanses of space while including small scale processes, 2) modeling a system that is time dependent on large and small-time scales, and 3) validating the models. (See **Section 7.3**, Risk Mitigation.) To this end, SHIELD is organized around four complementary and integrated research thrusts: 1) the Global Structure of the Heliosphere; 2) Evolution of PUIs; 3) the Influence of the Heliosphere on the LISM; and 4) Cosmic Ray Transport/Filtration (**Figure 6**). Each thrust is led by a Director (senior investigator) and Deputy Director (junior investigator), and integrates the expertise areas necessary for the research: observations, theory, numerical simulations, data analysis, localized kinetic physics, and MHD models.

2.2.1 RT1: Global Structure of the Heliosphere

RT1 has as director Justyna Sokol and as Deputy Director Marc Kornbleuth.

2.2.1.1 RT1.1 Shape of the Heliosphere

State of Knowledge: The standard picture of a comet-like heliosphere (Baranov & Malama, 1993) has been the dominant paradigm since the 1960s (Davis, 1955; Parker, 1961; Axford et al., 1963; Dessler, 1967; Baranov & Malama, 1993). Recent modeling and observations challenge this paradigm (Drake et al., 2015; Opher et al., 2015) and suggest that the solar magnetic field tension organizes the HS solar wind into two jet-like structures at the poles. The BU model (Opher et al., 2020) (**Figure 7**: top middle panel) suggests that the heliosphere has a croissant-like shape with a truncated tail. Other models find an elongated tail [**Figure 7**: top left (Izmodenov & Alexashov, 2015) and top right (Pogorelov et al., 2015)]. Cassini (Dialynas et al., 2017) (**Figure 7**: bottom left) and IBEX ENA (McComas et al., 2013) (**Figure 7**: bottom right) data give contradictory HS shapes. Knowledge Gap: The size and shape of the heliosphere are not known. Current models have problems. First, most assume Maxwellian plasma distributions, but suprathermal particles contain most of the pressure in the outer Solar System. Second, the effects of CRs are generally not included (Scherer and Ferreira, 2005), and third, microprocesses are excluded. PUIs and CRs must be included to correctly model heliosphere energetics and derive plasma distributions in the HS, a critical step for making modeled ENA maps to compare with data. Approach: In Phase I, we compared the BU and Moscow models and found that in both models the heliotail plasma was confined by the solar magnetic field (Kornbleuth et al., 2021a) (AC2) (**Figure 8**). Plasma solutions in the nose of the heliosphere are similar, but reconnection in the BU model leads to a shorter heliotail compared to the Moscow model. Higher-energy (> 10 keV) ENA maps differ between these models, since they can see farther down the tail, so ENA data could determine which model is correct. A key advance in Phase I is the realization that a Rayleigh-Taylor-like instability (RT-I) may occur along the solar magnetic axis in the HS. This instability destroys the coherence of the heliospheric jets and generates turbulence which leads to magnetic reconnection, allowing LISM material to enter the heliotail (AC1) (**Figure 9**). The turbulence driven by the instability is macroscopic and may accelerate particles. In Phase II, we will explore the time evolution of the MHD solution and the mixing of solar and LISM plasma (Opher, Drake, Izmodenov, and Zank).

10.3389/fspas.2023.1143909

TABLE 1 SHIELD team and participation, by thrust area.

Surname	Name	Affiliation	Role	RT1	RT2	RT3	RT4	Code devel	RI
Opher	Merav	BU	PI/Center Director	■	■	■	■	■	■
Richardson	John	MIT	Project Manager	■	■	■			■
Buxner	Sanlyn	UA/PSI	Director, BI	■	■	■	■	■	■
Florinski	Vladimir	UAH	Director, RT 3	■	■	■	■	■	■
Sokol	Justyna	SwRI	Director, RT 1	■	■				■
Giacalone	Joe	UA	Director, RT 4	■	■		■		■
Toth	Gabor	UM	Director, Code Coupling	■		■	■	■	■
Zank	Gary	UAH	Director, RT 2	■	■	■	■	■	■
Gkioulidou	Matina	JHU/APL	Deputy Director, RT 2	■	■				■
Kornbleuth	Marc	BU	Deputy Director, RT 1	■		■		■	■
Nikoukar	Romina	JHU/APL	Deputy Director, RT 4				■		■
Turner	Drew	JHU/APL	Deputy Director, RT 3	■		■			■
van der Hoist	Bart	UM	Dep. Dir, Code Coupling	■		■		■	■
Baliukin	Igor	IRI Ras	Co-I	■		■			■
Cummings	Alan	CalTech	Co-I	■		■	■		■
Dayeh	Maher	SwRI	Co-I	■	■				■
Dialynas	Kostas	AA	Co-I	■	■				■
Drake	James	UMD	Co-I	■	■		■		■
Elliot	Heather	SwRI	Co-I		■				■
Galli	Andre	Universitat Bern	Co-I	■					■
Gross	Nicholas	BU	Co-I	■	■	■	■	■	■
Hill	Matt	JHU/APL	Co-I		■		■		■
Izmodenov	Vladislav	Moscow State	Co-I	■		■			■
Mostafavi	Parisa	JHU/APL	Co-I		■	■			■
Reisenfeld	Dan	LANL	Co-I	■					■
Swisdak	Marc	UMD	Co-I	■	■		■		■
Zieger	Bertalan	BU	Co-I	■	■	■			■
Zhao	Ling-Ling	UAH	Co-I	■	■				■
Adhikari	Laxman	UAH	Research Scientist		■				■
Elliot	Heather	SwRI	Co-I		■				■
Du	Senbei	LANL	Research Scientist		■				■
Ghanbari	Keyvan	UAH	Postdoc			■			■
Miller	Jesse	BU	Postdoc	■					■
Nakanotani	Masaru	UAH	Research Scientist		■				■
Wang	Bingbing	UAH	Postdoc		■				■
Ma	Xiaohan	BU	PhD Student	■		■			■
Onubogu	Chika	BU	PhD Student	■					■
Powell	Erick	BU	PhD Student	■		■			■
Fuselier	Stephen	SwRI	Collaborator, RT1	■	■				■

(Continued on the following page)

TABLE 1 (*Continued*) **SHIELD team and participation, by thrust area.**

Surname	Name	Affiliation	Role	RT1	RT2	RT3	RT4	Code devel	RI
Guo	Fan	LANL	Collaborator, RT2		■				■
Kota	Jozsef	UA	Collaborator, RT4				■		■
Loeb	Abraham	Harvard	Collaborator, RT1						■
Michael	Adam	APL	Collaborator, RT1					■	■
Burgess	Juditra	BU	Collaborator, BI						■
Hokanson	Sarah	BU	Consultant, BI	■	■	■	■	■	■
Morrow	Cherilynn	PUNCH Mission	Consultant, BI	■	■	■	■	■	■

TABLE 2 **Phase I research accomplishments (AC) and next steps.**

SHIELD phase I research accomplishment (AC)	Implications for research in phase II
AC1: Discovered a Rayleigh-Taylor-like instability in the HS, driven by the neutral H atoms, that may cause the heliospheric tail to open (Opher et al., 2021) [RT1]	Explore how this instability effects the heliotail structure and if it leads to turbulent acceleration of ions in the HS [RT1]
AC2: Compared two MHD solutions (BU and Moscow); both found that the plasma in the heliotail is confined by the solar magnetic field (Kornbleuth et a. 2021a) [RT1]	Identify how turbulence and reconnection drive heliotail model differences, incorporate into SHIELD. Compare heliotail data and model results [RT 1]
AC3: Found that less than 10% of > 0.1 keV ENAs are produced in the inner HS (Fuselier et al., 2021) [RT1]	Identify origin of low-energy ENAs with implications for IBEX and IMAP [RT1]
AC4: Found that Voyager HS flows derived from the Compton-Getting (CG) effect using MeV and keV ions agree, but differ from plasma instrument (PLS) values (Cummings et al., 2021). Showed that magnetic flux conservation implies CG flows are incorrect (Richardson et al., 2021) [RT1, RT4]	Understand why CG flows differ from PLS observations and determine the true HS flow pattern [RT1, RT4]
AC5: Produced an ion spectrum downstream of the TS based on a hybrid model, PIC model, and theory and compared this spectrum with observations (Shrestha et al., 2021; Gkioulidou et al., 2022; Giacalone et al., 2021; Kornbleuth et al., 2021b). ENA spectra derived from these results differ from observations [RT2]	Understand the discrepancy between modeled and observed ENA spectra [RT2]
AC6: Discovered that the TS Mach number in the tail is larger than at the nose providing more heating of the PUIs in the tail (Giacalone et al., 2021) [RT2]	Understand the global physics of the heliospheric TS and the impact on PUIs and ENA maps as a function of energy [RT2]
AC7: Derived the ion energy spectra in the HS in the V2 direction by combining data from multiple instruments (Dialynas et al., 2020) [RT2]	Explore implications of these spectra on pressure balance at the HP and ENA creation and extend this study to V1 [RT2]
AC8: Determined the turbulence spectrum generated by dispersive waves in the HS (Zieger et al., 2020) [RT2]	Determine how HS turbulence energizes PUIs and other ions in the HS.
AC9: Derived the abundance of PUIs ahead of the TS in all directions (Nakanotani et al., 2021) [RT2]	Incorporate the PUI abundances in global MHD modeling [RT2]
AC10: Showed that BU and Moscow model solutions are similar at large distances from the HP although there are differences at the HP (Kornbleuth et al., 2021a) [RT3]	Constrain the draping of the magnetic field and B_{ISM} direction based on observations [RT3]
AC11: Found that GCRs are trapped where the magnetic field in the LISM is parallel to the HP surface (Florinski et al., 2021) [RT4]	Determine the consequences for the entry and exit of CRs in the heliosphere [RT4]
AC12: Discovered that GCR anisotropies in the LISM are energy dependent (Nikoukar et al., 2022) [RT4]	Find how the energy dependent LISM GCR anisotropies are generated [RT4]
AC13: Discovered that the intensity of ~10–50 keV PUIs protons, accelerated at the TS, are relatively uniform over multiple azimuthal locations (Giacalone et al., 2021) [RT2, RT4]	Provide low-energy ACR injection energy as a function of location along the TS [RT2, RT4]

Differences between the BU and Moscow models will be investigated (Kornbleuth and Baliukin), including why the BU model has a short heliotail. Simulations will show how the RT instability (AC1) develops under different solar cycle conditions (Powell). Next, we will explore the effects of solar-driven temporal changes (CMEs, solar cycle changes) on the global HS structure (Onubogu, Sokół). The effects of CRs and PUIs on the RT instability will be studied as well as the development of turbulence, and the consequences of this turbulence on the acceleration of HS PUIs (Ma and Opher) (RT2.3). The acceleration of PUIs and other ions will be modeled by hybrid codes (Giacalone and Du), PIC codes (Drake, Swisdak), and kinetic models (Zhao and Zank), using the MHD-derived RT-I turbulence. Model and observed high-energy ENA maps (> 10 keV), which probe the tail to larger distances, will be compared and we

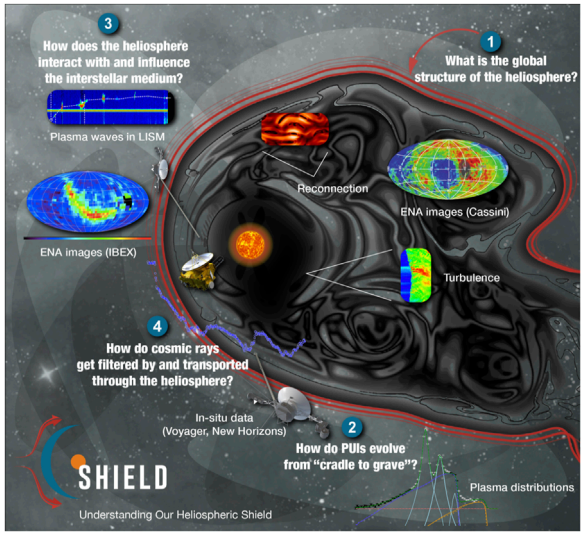


FIGURE 6
SHIELD unites and leverages the expertise of more than 45 world-class researchers from 18 partner institutions to understand our heliospheric shield.

will make predictions for IMAP (Gkioulidou, Dialynas, Fuselier, Sokół, and Dayeh). The heliotail structure and evolution, measured by IBEX (> 0.5 eV) and IBEX polar maps, will be used to validate SHIELD (Fuselier, Dayeh, Sokół, Reisenfeld). All these tasks depend on the global structure of the heliosphere, which depends on the solar wind, PUIs (RT2), CRs (RT4) and LISM conditions (RT3). The PUI and CR acceleration (RT1) will be added to the MHD model. SHIELD will use the complete plasma distribution to create model ENA maps to compare with IBEX, Cassini and, in the future, IMAP data and constrain the shape of the heliosphere.

The next task will be to solve the discrepancy between observed and modeled HS flows (AC4). V2 HS flows derived from particle data using the CG effect and from direct plasma measurements do not agree. In Phase I, SHIELD used Voyager CRS observations to measure particle anisotropies and found that LECP (keV particles) and CRS (> 0.5 MeV) CG speeds agree with each other, but not with plasma data (AC4). These observations suggest that the V1 HS stagnation region may not be real (**Figure 3**), which would affect models of ENA production. In Phase I, we assembled an expert team to study these results. We found that if the speed profile at V1 were similar to that found at V2 by PLS, inconsistencies with magnetic flux conservation and the Rankine-Hugoniot conditions at the shock would be reconciled (Richardson et al., 2021) (AC4).

In Phase II, we will determine why the CG calculations give different speeds and if the plasma and particle speeds are decoupled (Richardson, Cummings, and Zank), which would affect particle transport. The high-energy ions may set up HS pressure gradients that invalidate the CG calculations (Roelof, 2021). SHIELD will investigate this possibility using the HD code coupled to CRs and PUIs. We will identify the origin of low-energy ENAs with implications for IBEX and IMAP (Fuselier with Galli) (AC3). Outcome: SHIELD will determine the flow speeds in the HS, the structure of the heliotail, the shape of the heliosphere, and its evolution in time during the solar cycle.

2.2.1.2 RT1.2 Thickness of the Heliosheath

State of Knowledge: The thickness of the HS measured by Voyagers is thinner than models predict by almost a factor of 2. Mechanisms suggested to resolve this problem are solar cycle variations (Izmodenov et al., 2005; 2008), removal of hot HS ions by charge exchange (Malama et al., 2006; Opher et al., 2020), inclusion of thermal conductivity (Izmodenov et al., 2014), and escape of ACRs across the HP (Guo et al., 2010). All produce insufficient HS thinning. Knowledge Gap: The HS is much thinner than predicted. Approach: SHIELD will explore three mechanisms that may thin the HS: 1) charge exchange of neutral H with the full plasma distribution; 2) the inclusion of ACRs and their transport, and 3) the inclusion of thermal conduction. The SHIELD model will calculate accurate plasma distribution functions to test mechanism 1) (Powell, Ma, and Kornbleuth). Mechanism 2) will be tested using models of ACR escape (Guo et al., 2018) with transport parameters from the MHD code coupled to the CR transport code (Florinski and a postdoc). To explore mechanism 3), thermal conduction, already in SWMF, will be activated in OH (van der Holst, Toth) (see B3, Development of the SHIELD Model). These tasks will determine which mechanism, or combination of mechanisms, account for the HS thicknesses observed at V1 and V2. Outcome: SHIELD will determine why the HS is thinner than expected and illuminate the physics needed in the global model.

2.2.1.3 RT1.3 Structure of the Heliopause

State of Knowledge: The structure of the HP and the degree of porosity determine the transport rates of energetic particles into and out of the heliosphere. Possible processes at the HP are magnetic reconnection (Swisdak et al., 2013), interchange instabilities (Florinski, 2015), flux transfer events (Schwadron & McComas, 2013), and turbulent reconnection (Grygorczuk et al., 2014). Knowledge Gap: The HP crossings at V1 and V2 are significantly different (e.g., the particle dropouts/enhancements were on the opposite sides of the HP). The variations of the HP structure and properties with time and location are not known. Approach: SHIELD will run localized PIC (Drake, Swisdak), hybrid (Giacalone) and multi-ion simulations (Zieger) at key HP locations using input from MHD simulations (Kornbleuth, Ma, Powell, Opher). Localized MHD simulations (Florinski) and theoretical studies (Florinski, Zank) will explore MHD instabilities at the HP. SHIELD results in the V1 and V2 directions will be compared with observations of the magnetic field at the HP (Szabo) to test model accuracy and validate SHIELD's global HP results. Solar cycle variation effects on HP properties will be explored. Outcome: SHIELD will determine the HP porosity that affects cosmic ray entry into the HS and LISM CR intensities.

2.2.2 RT2: Evolution of PUIs

RT2 has director Gary Zank and as Deputy Director Matina Gkioulidou.

2.2.2.1 RT2.1–2.4 Evolution of PUIs in the Supersonic SW, HS and LISM

State of Knowledge: PUIs dominate the thermal energy of the solar wind from ~20 AU to the HP (Richardson et al., 2008). New Horizons measures PUIs but Voyager cannot; HS suprathermal particle distributions below 28 keV are unknown. Turbulence plays a

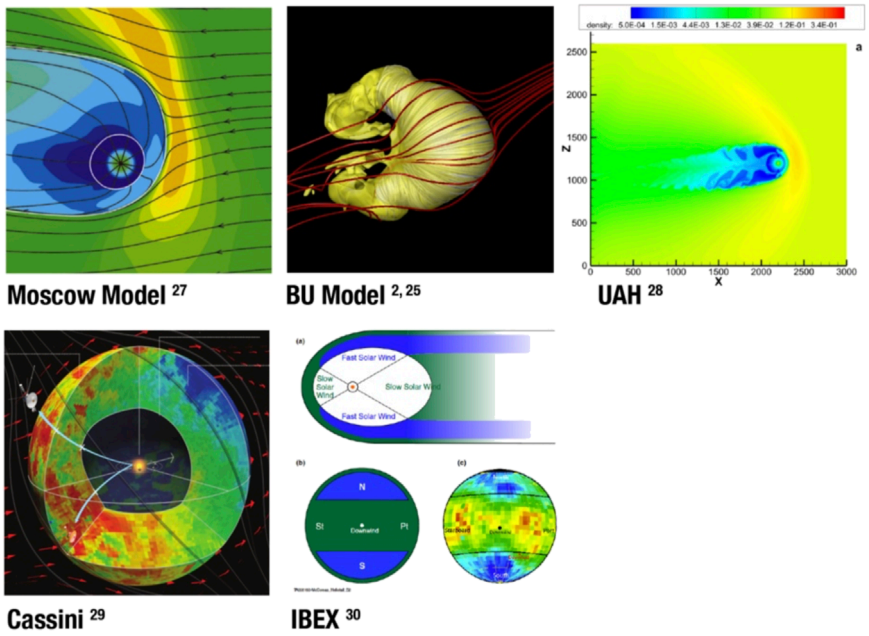


FIGURE 7
A major controversy: What is the shape of the heliosphere? Solving it is one of the goals of SHIELD.

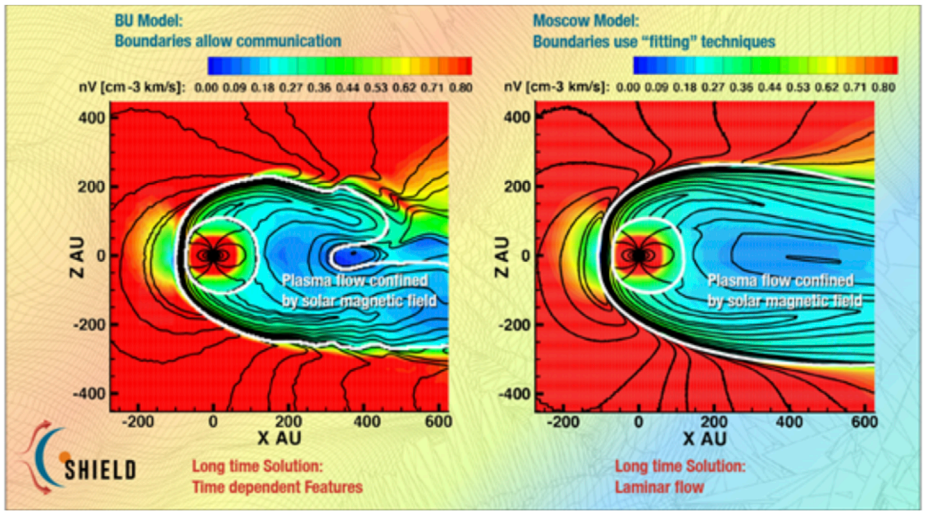


FIGURE 8
Different (BU and Moscow) models provide different plasma solutions in the tail. Differences need to be understood that will lead to the Center Milestone M1.

major role in heating the solar wind from the corona to the TS. PUIs provides a source of waves and turbulence that, through dissipation, heat the solar wind and mediate the transport of energetic particles (Lee et al., 1987; Williams & Zank, 1994; Zank, 1999; Smith et al., 2006; Smith et al., 2017). The solar wind turbulence transmitted across the TS is dominated by compressible modes. Heliospheric processes that effect the LISM PUIs are the ionization of solar wind and HS neutral H, and transmission of HS compressible turbulence (Burlaga et al., 2015; Zank et al., 2017a) across the HP and its conversion to incompressible modes (Burlaga et al., 2018; Zank et al., 2019; Zhao et al., 2020). Secondary charge exchange (CE) produces hot PUIs (Zank et al., 1996) that form the IBEX ribbon (Heerikhuisen et al., 2010). ENAs produced by CE from these PUIs serve as probes of the LISM and B_{ISM}. The CE of heliospheric neutrals can drive instabilities that generate weak interstellar turbulence (Florinski et al., 2010), heating the LISM, and weakly scattering the PUIs. Knowledge Gap: We do not understand how PUIs are accelerated at the TS, how they are transported

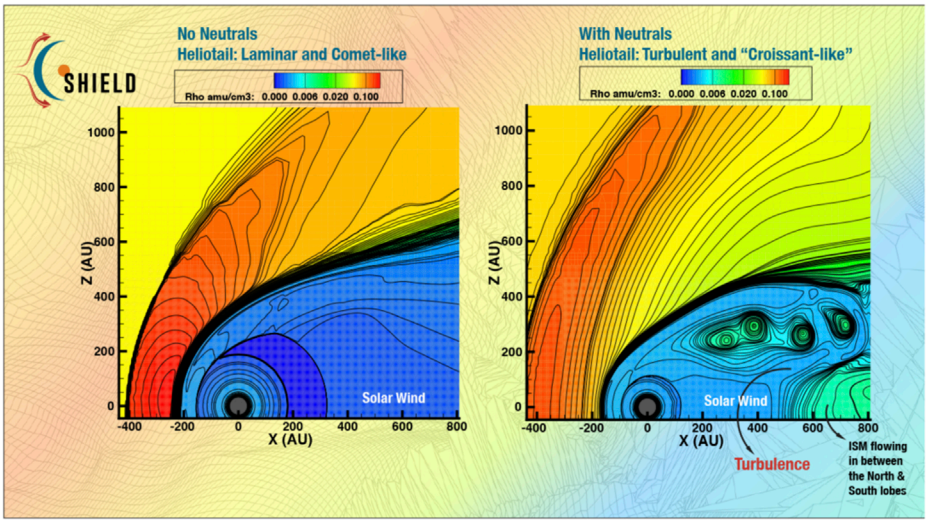


FIGURE 9
SHIELD discovered an instability in the heliosheath, SHIELD will investigate its consequences for particle acceleration and structure of heliotail.

and heated in the HS and LISM, or the distribution of ENAs produced when the PUIs are lost via charge exchange. The effect of compressible HS turbulence on the PUIs is an outstanding question. The origin of the compressible turbulence in the LISM is not known (Burlaga et al., 2018; 2020; Zank et al., 2019; Zhao et al., 2020). Approach: Localized simulations (hybrid, PIC, and multi-fluid models) and theory (Zank et al., 2017a; 2017b; 2021) will be used to investigate how PUIs are affected by turbulence, reconnection, and non-linear waves. Outcome: The simulations and theory will determine the PUI properties and how best to incorporate the PUIs into the SHIELD model (B3, Development of the SHIELD Model).

2.2.2.2 RT2.1 PUI Evolution in the Supersonic Solar Wind

NH observations and theoretical models (Zank et al., 1996a; Zank et al., 2010) show that shocks preferentially heat PUIs (Zirnstein et al., 2018). Zank, Adhikari, and Zhao will include turbulence generated by PUIs in localized MHD simulations to study which processes heat the PUIs (AC9). Interplanetary shocks observed by NH will be simulated using PIC simulations (Swisdak, Drake) (Drake et al., 2010) and multi-ion fluid simulations (Zieger, Keebler) (Zieger et al., 2015; 2020). Turbulence parameters will be constrained by Voyager (Fraternale et al., 2019) (Richardson, Szabo) and NH (Keebler et al., 2022) data (Elliott, Hill). Outcome: SHIELD will predict the PUI spectra upstream of the TS for input into global MHD models and production of ENA maps.

2.2.2.3 RT2.2 PUI Heating and Acceleration at the TS

SHIELD made progress in Phase I toward understanding the transmission and acceleration of PUIs at the TS. Plasma and magnetic field parameters upstream of the TS in the nose, the flank, and tail-ward directions were determined by a solar-wind/PUI/turbulence model (Nakanotani et al., 2021) (AC9). Hybrid simulations, which included large-scale turbulence upstream of the TS, found that the 10–50 keV PUI intensity was similar at all three TS locations. However, we found significant differences in the distribution of 0.5–1 keV ENAs coming from the tail region of the heliosphere compared to those from the nose and flank directions. PIC simulations that included electron physics and provided better estimates of the reflected (accelerated) PUI population were also performed (Drake, Swisdak). The transmitted > 4 keV ion intensities are comparable to those from the hybrid model and predicted by theory (Zank et al., 2010; Wang et al., 2023) verifying the results. We used the ion spectra produced by the hybrid simulations (Giacalone et al., 2021) as input to model ENA spectra (Gkioulidou et al., 2022). The model ENA intensities were lower than those measured (**Figure 10**). We will identify additional physical process for the energization of PUIs (Nakanotani, Zank, Zhao, Wang, Du, Guo, Giacalone, Drake, Opher) and compare with ENA and ion data (Dialynas, Gkioulidou, Dayeh, Fuselier, Decker, Kornblueth, Wang). Outcome: SHIELD will predict the PUI spectra in the downstream the TS for input into global MHD models and production of ENA maps.

2.2.2.4 RT2.3 Determine How PUIs Evolve in the HS

The difference between predicted and observed ENA distributions (**Figure 10**) could be a signature of PUI acceleration in the HS. Turbulence in the HS is highly compressible, different from the supersonic solar wind. The transport theory for PUIs in compressible turbulence will be developed (Zank, Zhao) using a stochastic differential equation (SDE) code. Hybrid simulations (Giacalone, Du, Guo) will model the evolution and acceleration of HS particles by compressible turbulence. PIC simulations will test if reconnection energizes HS PUIs (Drake et al., 2010; Dahlin et al., 2016; 2017) (Swisdak, Drake). Multi-ion simulations (Zieger) will determine if compressible turbulence is generated by dispersive magnetosonic waves (Zieger et al., 2020). Outcome: SHIELD will predict the PUI spectra in the HS needed for global MHD models and production of ENA maps.

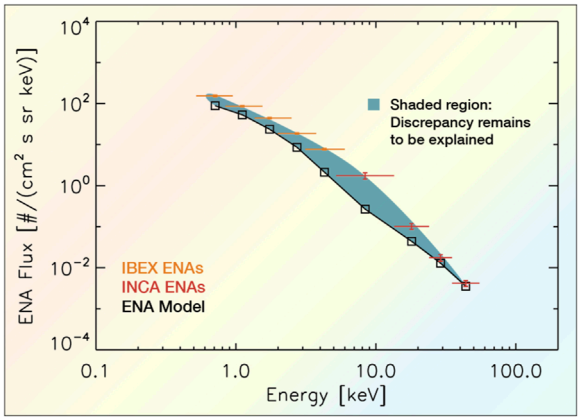


FIGURE 10
Discrepancy between modeled ENAs and observations needs to be explained (Gkioulidou et al., 2021).

2.2.2.5 RT2.4 PUI Evolution in the LISM

SHIELD will develop a transport model for heliospheric PUIs in the LISM. A theoretical transport equation (Zank, 2014) will be derived that will be solved via a SDE method (Zhao et al., 2019) (Zank, Florinski, Zhao). Complementary hybrid simulations (Giacalone, Florinski) will model PUI creation, wave/turbulence generation, and transport in the LISM to ensure that the theoretical transport formalism is valid. Outcome: SHIELD will predict the PUI spectra in the LISM needed for global MHD models and production of ENA maps.

2.2.3 RT3: Influence of the Heliosphere on the LISM

RT3 have as a director Vladimir Florinski and as Deputy Director Drew Turner.

2.2.3.1 RT3.1 Heliospheric and LISM Plasma Interactions

State of Knowledge: The LISM is strongly influenced by solar transients (Gurnett et al., 2015; 2021) (**Figure 11**). The magnetic field was predicted to rotate at the HP, but a surprise was that B did not change direction at either the V1 or V2 crossing. B_{ISM} remains near the solar direction 33 AU past the HP. Constraining the direction of the pristine B_{ISM} is important for models of the IBEX ribbon (McComas et al., 2009) and the Cassini belt. Krimigis et al. (2009) Another V1 discovery was that the plasma density increased by a factor of ~30 at the HP and then continued to slowly increase for an additional 7 AU (Gurnett et al., 2013; Gurnett and Kurth, 2019), indicative of a plasma depletion layer (PDL) in the LISM (Zwan & Wolf, 1976; Wang & Richardson, 2004; Cairns & Fuselier, 2017). Knowledge Gap: We do not know how far the heliospheric influence extends into the LISM. Central questions include: 1) What controls the draping of B_{ISM} at the HP? 2) What are the intensity and direction of B_{ISM} and the plasma density in the pristine LISM? 3) How does the density change with distance? What are the properties and geometry of the PDL? Approach: SHIELD will compare B_{ISM} from MHD models with observations (Kornbleuth, Baliukin, Izmodenov, Opher, Ma, Powell, Richardson) to see if reconnection (Opher et al., 2017) and/or instabilities (Florisnki, 2015) (Florinski) twist the B_{ISM} to the solar direction and how far from the HP this twist occurs (AC10). We will derive PUI energy spectra and combine them with *in situ* inner/outer heliosheath ion energy spectra to obtain energetic particle spectra on both sides of the HP (Cummings, Decker). Existing theories and orbit/scattering models of keV to GeV charged particles will be used to determine the strength of coupling between particle species exhibiting diffusive behavior and the background plasma to establish the effect on pressure balance (Florinski). Global MHD models will evaluate the properties and geometry of the PDL using enhanced resolution near the HP (Florinski, Toth, Opher). Zank and Florinski will revisit the theory of tangential discontinuities in the presence of energetic particles. The SWMF anisotropic pressure (CGL) model will be used to perform local-scale CGL-MHD simulations within 10 AU of the HP (Florinski, Toth) to determine the PDL pressure anisotropy and relate it to the observed plasma depletion and magnetic field enhancement. Outcome: SHIELD will reconstruct the magnetic field pattern and the distribution of plasma and energetic particle pressures, their anisotropies, and current flows within the entire HP boundary layer. Draping of B and plasma in LISM will inform conditions in LISM.

2.2.3.2 RT3.2 the Nature of Transient Disturbances in the LISM

State of Knowledge: Large solar events cross the TS and collide with the HP, injecting compressive waves and structures into the LISM (Burlaga et al., 2013; Gurnett et al., 2015; Washimi et al., 2015; 2017; Burlaga and Ness, 2016; Kim et al., 2017). Interstellar shocks and pressure fronts are typically weak, having a compression ratio of less than 1.5. Knowledge Gap: LISM transients are not resolved by current models (Kim et al., 2017), making it difficult to relate LISM structures to their precursor solar events. The measured width of the shocks or pressure fronts was about 0.1 AU, which is four orders of magnitude larger than the ion inertial length in the LISM (a scale that defines the shock thickness). Mostafavi and Zank (2018) argued for collisional broadening of the shocks using an analytic model; however kinetic models have not reproduced these results. Approach: High resolution MHD simulations, using an adaptively refined mesh, will model the propagation of transients (Zieger, Toth, Mostafavi). More sophisticated and accurate theoretical models such as Braginskii type fluid models are also being utilized to study the nature of shocks in the VLISM/LISM (Ghanbari, Zank). The global heliospheric model provides TS and HP locations (RT1). Shocks will be modeled on kinetic scales using hybrid particle-mesh models, resolving ion dynamics only, with a very long simulation box (Florinski, Giacalone). Outcome: SHIELD will explain the timing and magnetic field signatures of the shock waves and pressure fronts driven by events inside the heliosphere and observed in the LISM by the Voyagers.

2.2.3.3 RT.3.3 Character of Turbulence and How Far the Heliospheric Effects Extend into the LISM

State of Knowledge: V1 observed both transverse and longitudinal magnetic fluctuations in the LISM (Burlaga et al., 2015; 2018; Fraternale et al., 2020) with a Kolmogorov power spectrum. Fluctuations were primarily compressive near the HP, but a transverse component appeared 2 years after the HP crossing while the compressive component showed little change. Magnetic fluctuation intensity was a factor of 100 times higher

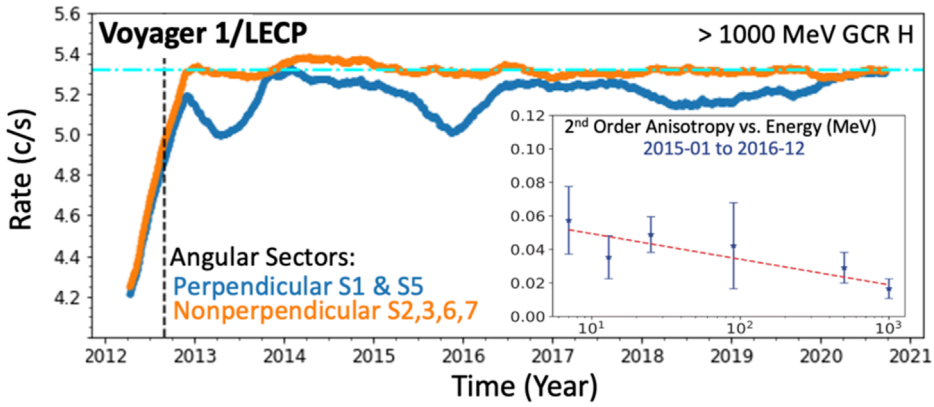


FIGURE 11
Effect of the heliosphere is seen throughout the LISM (Nikoukar et al., 2022). SHIELD addresses how far this influence extends.

than the spectrum of interstellar fluctuations from radio astronomy observations (Haverkorn et al., 2008). The compressive component may result from refraction of solar wind waves near the HP (Zank et al., 2017b). Knowledge Gap: We do not understand the source of the LISM turbulence. A global model of turbulent transport in the LISM has not been developed and the distribution of turbulent fluctuations throughout the outer heliosheath remains a complete mystery. The transverse fluctuations may result from nonlinear interactions between magnetosonic waves and incompressible 2D structures (Zank et al., 2019), but this hypothesis does not explain why the total power in the magnetic fluctuations increases with distance from the HP. Compressive turbulence could have a very different effect on particle pitch-angle scattering in the LISM than slab and 2D turbulence. A closure on the question of the IBEX ribbon origin (Heerikhuisen et al., 2010; Florinski et al., 2016) and GCR penetration of the HP transition layer is impossible without a detailed knowledge of scattering and cross field diffusion in the LISM. Approach: SHIELD will use measured LISM turbulence spectra (Florinski), particle transport simulations (Florinski) and theory (Zank) to address the nature of the LISM turbulence. We will investigate how the magnetic shear at the HP effects the cosmic rays. We will also develop theoretical underpinnings for energetic charged particle transport in compressive and anisotropic turbulence (Florinski). Direct numerical simulations of the orbits of a very large number of test particles in pre-specified, data-driven magnetic turbulence models will supplement and verify the theory (Giacalone & Jokipii, 1999; Giacalone, 2005). The inclusion of anisotropic magnetic turbulence is especially important for GCR transport in the LISM. We will model the transport of multi-component turbulence in the LISM using an updated turbulence model including both incompressible and weakly compressive fluctuations (Zank, van der Holst). Results will be used to model GCR transport in the outer HS and their filtration by the HP to assess how Earth is protected from GCRs in RT4 (Florinski, Toth, Opher). These simulations will be based on integration of a large number of guiding center averaged trajectories backward in time from the observation point (Florinski et al., 2010) with pitch angle scattering and perpendicular diffusion included as stochastic processes that will feed into the overall SHIELD model (through SPECTRUM). Outcome: SHIELD will establish physically realistic transport coefficients as a function of particle energy appropriate for particle transport in the LISM, which will feed into predicting the radiation environment in heliosphere.

2.2.3.4 RT3.4 Spatial Extent of the LISM Region Disturbed by the Heliosphere

State of Knowledge: Astrospheres are obstacles to the interstellar gas and sources of waves and particles in the LISM. The heliosphere is a source of ENAs produced by charge exchange with interstellar material; the leading IBEX ribbon explanation requires that ENAs fill the space around the HP to a distance of several hundred AU (Möbius et al., 2013). Knowledge Gap: A global picture of the interstellar environment near the heliosphere is far from complete. The turbulent layer that exists in front of the HP could extend to the bow wave, whose location is unknown. Do turbulent fluctuations continue to grow away from the HP or begin to decrease past a certain distance? What is the primary location of the gyrating PUIs producing the IBEX ribbon? How do PUI induced instabilities contribute to the outer HS turbulence? Approach: Global simulations of ENA transport will determine the ENA distributions in the LISM (RT1). Our goal is to explain the secondary distributed flux and make predictions for the IMAP mission (Opher). We will model secondary PUI populations in the presence of turbulent fluctuations to determine the width and strength of the ribbon. The team will revisit existing theories of wave transmission across a magnetic discontinuity and nonlinear wave conversion in a low-beta plasma (Zank). To determine the turbulence extinction distance, weak Coulomb collisions will be added in a BGK or Fokker-Planck approximation (Zieger, Zank). Semi-analytic and empirical models will be developed to track evolution of the power spectrum of magnetic fluctuations with heliocentric distance (Florinski, Zank). Outcome: SHIELD will describe the extent and properties of the ribbon producing region and explain the evolution of turbulent magnetic fluctuations with distance from the HP.

2.2.4 RT4: Cosmic Ray Transport and Filtration

RT4 has as director Joe Giacalone and as Deputy Director Romina Nikoukar.

2.2.4.1 RT4.1 Cosmic-Ray Intensities and Spectra Throughout the HS

State of Knowledge: GCRs are observed at multiple locations in the heliosphere and vary with the solar cycle, radial distance, and heliolatitude. V1 and V2 observed rapid CGR increases inside the HP. In the LISM, GCR anisotropies are observed, likely associated with transient disturbances. ACRs can cross the HP and are observed in the LISM. Knowledge Gap: The transport of GCRs and ACRs near to and across the HP are not understood; neither is the magnetic field configuration leading to the GCR anisotropies. Approach: Transport of GCR will be studied with the coupled SHIELD code (Florinski, Toth) using the pitch angle transport option to study the GCR anisotropy. The decay of ACRs outside the HP at V1 and V2 provide a constraints on ACR transport in the LISM. Giacalone and Florinski will use SHIELD to numerically integrate the trajectories of a large number of 10–100's of MeV ACRs near the HP (including anisotropic turbulence), model their escape into the LISM, and compare their output with Voyager observations (Cummings, Hill, Stone). A new feature SHIELD found is a magnetic trap (AC11) on the northern flank of the HP. At the center of the trap, the magnetic field is especially weak, offering a preferred location for CRs to cross the HP. We are modeling the HP crossing points of CRs to elucidate how the magnetic trap modifies the effectiveness of the heliospheric shield (Florinski et al., 2021). SHIELD will develop code for community use on our website to calculate GCR and ACR intensities and energies at any point inside the heliosphere. A similar web interface, Helmod (https://helmod.org), is in operation, but a new run must be performed every time a new position is entered by the user. We will perform many runs using the method of backward-tracing stochastic integration of the Parker equation (similar to GALPROP) at multiple locations and create a lookup table of values which can be accessed quickly by the user. Similar look up tables will be made as more sophisticated methods become available from the coupled SHIELD model. SHIELD will also consider solar cycle variation of GCR modulation, including the effects of the varying HCS position and diffusion coefficients. Outcome: SHIELD will determine GCR transport and the radiation environment throughout the heliosphere.

2.2.4.2 RT4.2 GCR Anisotropies in the LISM

State of Knowledge: Voyager observed multiple regions with anisotropic GCR distributions in the LISM due to dropouts of >211 MeV proton intensities perpendicular to the magnetic field. Cosmic ray electrons remain isotropic. Knowledge Gap: The cause of these anisotropies or why electrons behave differently than protons is not known. Approach: We will analyze data for these events and compare these data with existing theories (Kota, Giacolone). The energy dependence of these events will be determined (Nikoukar, Hill) by using LECP pulse height analysis data in a wide energy range (from MeV to GeV). A combination of magnetic trapping and cooling downstream of solar-induced transient disturbances and GCRs escaping from a single compression in the draping region has been proposed to explain observed anisotropies. Outcome: SHIELD will elucidate the physical processes that suppress GCR intensities near 90°C pitch angle and be used to improve GCR models of modulation and transport.

2.2.4.3 RT4.3 Acceleration of Anomalous Cosmic Rays (ACRs)

State of Knowledge: A surprise of the TS crossings was that neither Voyager spacecraft found evidence for ACR acceleration at the TS (Giacalone, 2012). Before Voyager, the paradigm was that diffusive shock acceleration of PUIs created ACRs at the TS (Fisk et al., 1974; Pesses et al., 1981; Giacalone, 2012). ACR intensities did not change at the TS, but slowly increased as the spacecraft moved deeper into the HS (Decker et al., 2010). Hypotheses for the ACR acceleration mechanism and location are: 1) in the flanks of the TS (Cummings et al., 2008); 2) in "hot spots" along a turbulent TS (McComas & Schwadron, 2006; Kota, 2010); 3) by reconnection (Guo et al., 2010; Opher et al., 2011; Keebler et al., 2022); 4) by turbulence processes (Zank et al., 2015); 5) by second-order Fermi acceleration; and 5) by turbulence generated by multi-ion magnetosonic waves (Dialynas et al., 2020). Knowledge Gap: We need to understand where and how ACRs are accelerated. Approach: SHIELD will explore acceleration of ACRs at the TS and in the HS using the coupled ACR transport and MHD models. Opher will explore consequences of the RT instability for ACR acceleration. SHIELD output will be compared with V1 and V2 ACR data from V1 and V2 (Florinski, Cummings); model runs will determine the contribution of reconnection, turbulence and TS conditions to ACRs. ACR anisotropies when V2 was outside and inside the sector zone (defined by the heliolatitudinal extent of the heliospheric current sheet will be compared (Hill); preliminary work suggests an energy dependence in the anisotropies that will constrain ACR acceleration models. Outcomes: SHIELD will determine where and how ACRs are accelerated and predict the ACR spectra and intensities throughout the heliosphere.

3 Development of the SHIELD model

At the end of 5 years of Phase II (2022–2027), SHIELD will deliver a self-consistent, comprehensive global model of the heliosphere as a system that incorporates all critical physical processes (**Figure 5**; where year 1–5 corresponds to 5 years of Phase II of SHIELD). Toth with van der Holst leads the development of the SHIELD code. The global MHD model is the OH (Opher et al., 2003; 2009; 2020) component of SWMF (Toth et al., 2005; 2012), which is time dependent, can treat the thermal solar wind and PUIs (Opher et al., 2020) separately, and includes a kinetic treatment of the neutral H atoms (Michael et al., 2021). The suprathermal PUI and CR transport is done by the SPECTRUM (Florinski, 2015; Guo et al., 2018) particle code that will be inserted into the SWMF framework and coupled to OH (Kornbleuth, Opher). Solar wind turbulence is simulated with the Alfven Wave Solar Model (AWSoM) (van der Holst et al., 2014) model that is already part of the SWMF framework; SHIELD will extend AWSoM to handle compressible fluctuations (van der Holst, Zank). Turbulence and MHD parameters will be fed into SPECTRUM from OH. Both SPECTRUM and AwSOM parameters will be constrained by localized models, theory and observations. OH will be coupled to SPECTRUM, and AwSOM will drive the solar wind solution. This

task builds on the prior experience of Florinski (Florinski et al., 2003; 2004; 2008) who has already developed a Parker-SDE solver (Florinski & Pogorelov, 2009; Florinski et al., 2013). To simulate PUI transport, a pitch angle transport module will be added (Chalov & Fahr, 2000; Malama et al., 2006; Guo & Florinski, 2014) to SPECTRUM. In the pitch angle option, the gyrotropic distribution is evolved by the focused transport equation (FTE) that depends on coefficients related to the mean and the fluctuating components of the magnetic field computed by AwSOM. The pitch-angle diffusion coefficients are obtained from quasi-linear theory; the perpendicular diffusion coefficient is from nonlinear guiding center theory (Zank et al., 1996b; Shalchi et al., 2004; le Roux et al., 2010) and the velocity diffusion coefficient is estimated from the intensity of the large-scale velocity fluctuations (Chalov et al., 1997; Florinski et al., 2003). In Phase I, SHIELD implemented a more efficient algorithm to evaluate the source terms for kinetic neutrals (Powell et al., 2021). SHIELD has already coupled OH to AMPS, a code that describes neutrals kinetically (Michael et al., 2022). We will perform the coupling between OH to FLEKS being developed now by Yuxi Chen. We will use the OH multi-fluid version to set the initial solution, then couple OH to SPECTRUM and update the solution. The description of kinetic neutrals will be extended (Chen, Kornbleuth, Michaels) to the new version of the code. When completed and validated, the model will be open source (DMP).

4 Team contribution

The PI provides overall leadership; the PM provides strategy and technical support and ensures integration. The RT Directors, Deputy Directors, Code Coupling Director, and BI Director execute thrusts. The roles are detailed in **Sections 1.4**, **2.2**, **7**, and **Table 1**.

5 Justification for DSC mode of support

The breadth of physics, of data analysis required for model verification, and of associated coding are well beyond the expertise of any single or even pair of PIs or institution. A global model inclusive of all relevant physical processes at all spatial resolutions (microphysics and macrophysics) can only be done in a Center environment with a tight focus on knowledge integration. We capture the kinetic transport of PUIs and CRs with the kinetic code SPECTRUM coupled self-consistently to a fluid code (SWMF/OH). The input is provided by localized simulations that treat from first principles kinetic effects like turbulence (Giacalone), reconnection (Drake, Swisdak) and acceleration of particles (Zank). SHIELD will couple SPECTRUM (Florinski) to the MHD global model (Opher, Toth) with input from localized codes. To benchmark the global code as well as the localized model and theory output, we use data from Voyager (Richardson, Cummings, Nikoukar, Gkioulidou, Hill, Dialynas), NH (Elliot, Zirnstein), IBEX (Fuselier, Dayeh, Reisenfeld) and Cassini (Dialynas).

6 Building on Phase I

SHIELD began with a large team, recognizing a broad collaboration is needed to realize Center goals and build team-based processes. This strategic decision positions us well for Phase II; it resulted in tight integration and early critical outputs. The team grew with Nikoukar studying GCR anisotropies, Gkioulidou with IMAP expertise, and van der Holst with expertise in turbulence and the AwSOM code. SHIELD also appointed a new BI director.

7 Management plan

7.1 Leadership

Figure 12 shows the management plan of SHIELD. The Center Director (CD) is Merav Opher, Professor of Astronomy at BU and William Bentinck-Smith Fellow at Harvard Radcliffe Institute, 2021–2022. Opher is responsible for the Center's direction, progress and evolution. She ensures that SHIELD's research, workforce development, diversity, outreach, and sustainability goals are met, and that the Center's schedule, reports, and budget remain on track. A leader in the global modelling of the heliosphere and the comparisons of these models with data, Opher's work has challenged the decades old paradigm of a comet-like heliosphere. She will ensure the Center works effectively within NASA policies and guidelines. The Project Manager (PM) is John Richardson (MIT). Richardson is responsible for day-to-day management, provides strategic, technical and logistical guidance, and ensures integration of Center activities. Additionally, he identifies and brings new members into SHIELD and assumes responsibility for SHIELD programming in the CD's absence. Richardson is PI of the Voyager plasma experiment and a leader in observational studies of the solar wind, including making the first plasma observations of the TS, HS, HP, and LISM. Opher and Richardson work with the Executive Committee (EC), Advisory Boards, and NASA to ensure SHIELD has appropriate guidance, institutional resources and administrative support.

7.2 Roles of internal leadership

The EC serves as the core administrative team. Led by the CD and PM, the EC is comprised of the Research Thrust (RT) Directors, Code Coupling Director, and Broader Impacts Director. RT Directors (Sokol, Zank, Florinski, Giacalone) provide management, oversight and scientific direction for SHIELD. They are responsible for driving and integrating the research, executing on short- and long-term goals and milestones, and mentoring SHIELD's young investigators (YIs). RT Directors meet with their teams bi-weekly, and with the CD and PM monthly. Deputy Directors (DDs, usually postdocs and early career researchers) work with the RT Directors and are part of the EC. DDs implement and augment the research by working across multiple RTs, run the weekly RT meetings and prepare annual reports. Two of the six DDs [Kornbleuth, Gkioulidou*, Nikoukar*, van der Holst, Turner, Gross] are female, advancing SHIELD's diversity goals. The Code Coupling Director (Tóth) leads the coupling and BI Deputy (Gross)

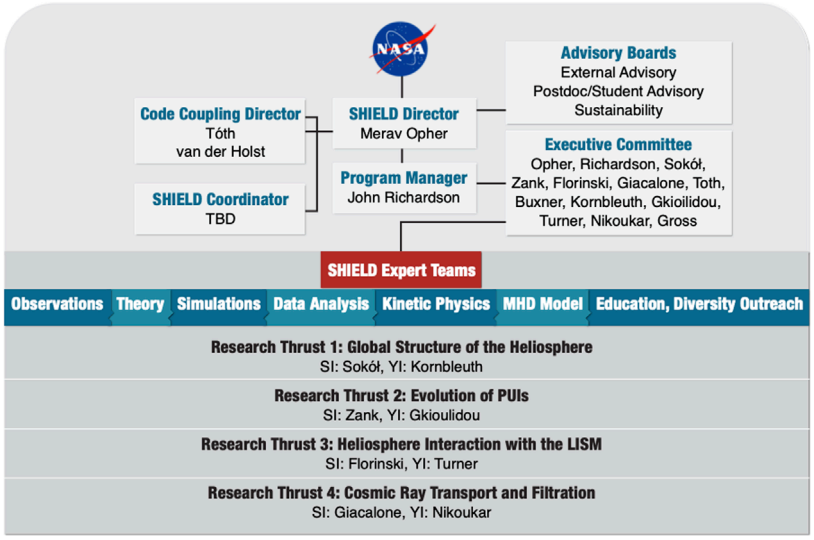


FIGURE 12
SHIELD organizational chart.

of the codes and his team meets weekly. The BI Director (Buxner) directs training, outreach and diversity programs, and ensures tight integration between the research and BI activities; her team also meets weekly. SHIELD will hire a full-time Project Administrator to provide administrative, programmatic, and financial support services. Collaborators (non-funded) support SHIELD research and BI activities. **Table 1** profiles the team.

7.3 Decision making and operational procedures

SHIELD is structured to ensure transparency in decision-making, coordination across the research and BI thrusts, and the flexibility to make timely and appropriate changes in programmatic direction. The ability to react to new discoveries is particularly important for the outer heliosphere where new data and models are changing current views. Agility and teaming are prioritized; protocols support effective resource allocation, continual strategic planning and robust communication. The success of Phase I is a direct result of rigorous operational protocols; key elements are retained in Phase II.

Decision Making: Decision making is the responsibility of the CD, following consultation with the PM, EC and Advisory Boards. The CD, PM, and EC evaluate projects annually to assess alignment and progress of thrusts, using performance indicators. Gaps are identified and mitigated. Evaluation leads to resource recommendations with final decisions made by the Director in close consultation with the PM. The CD communicates decisions to and receives concurrence from the EC, EABs and NASA. The CD shares decisions with SHIELD team and updates the strategic plan.

Oversight: Active oversight has and will continue to ensure effective operations and management. The External Advisory Board (EAB), a Postdoc and Student Advisory Board (PSAB), and a Sustainability Board (SB) ensure effective oversight, operation and management. EAB: The EAB provides guidance and oversight with regard to vision, scientific direction and strategy, project execution and progress, conflict resolution, risk mitigation, and sustainability. It meets yearly to review SHIELD; findings are documented in an annual report. Board members also meet with the CD and PM as needed. Members are: *Ralph McNutt* (Chair), Chief Scientist, Space Science, Space Exploration Sector, JHU/APL; *Jeffrey Hughes*, Professor of Astronomy, BU; *Mae Jemison*, PI, 100 Year Starship; *Janet Luhmann*, Senior Fellow, Space Sciences Laboratory PI, IMPACT, SSL/UC-Berkeley; *Sara Seager*, Class of 1941 Professor of Physics and Planetary Science, MIT; *Alan Stern*, Associate VP, Space Science and Engineering Division, SwRI; PI, New Horizons; and *Janet Vertesi*, Associate Professor of Sociology, Princeton U. PSAB: The PSAB will be formed from SHIELD postdocs and students. SHIELD will support 5 postdoctoral fellows (1 at BU 1 at UAH, 1 at SwRI, 1 at UA, and a SHIELD postdoc, funded by CERIF, to be competed for by the SHIELD institutions). The co-Is and Directors will nominate 1-2 PhD students to serve on the PSAB, which will meet quarterly with the EC to provide input and suggestions on strategy, mentoring, professional development, and Center communications, supporting SHIELD's reverse mentoring (Jordan & Sorell, 2019) goals (**Section 8.4**, Mentoring). The PSAB will elect its own chairperson; each will serve a one-year term to allow multiple postdocs and students to develop leadership skills. Assignments will be for 2 years, assuring continuity. The PSAB will present to the EAB during its annual meeting. SB: The Sustainability Board will advise SHIELD in formalizing and launching the Center's sustainability initiative, the Institute for Habitable Astrospheres (IHA). The IHA, using SHIELD's global model, will support convergent research focused on the origin of life that can be solved by combining expertise in radiation, biology, chemistry, and heliophysics. A one-day workshop will kickstart the IHA. The Board is composed of *Avi Loeb* (Harvard), *Sara Seager* (MIT) (co-directors of IHA), *Alan Stern*, and *Ralph McNutt*. The SB will meet with the CD, PM, and DDs annually.

TABLE 3 Risks related to Research.

	Mitigation
Codes from multiple researchers and institutions need to be coupled to achieve research goals	A designated Code Coupling Director leads weekly team meeting to ensure progress
Multiple theoretical and numerical models will address the same physical processes; must decide which is best for use in SHIELD.	Coordination of research among the institutions to use common inputs and bi-weekly team meetings to ensure collaboration and progress
Critical sections of SHIELD code lag in implementation due to coding or personnel issues	Code coupling team deputy to take over as lead; reserve resources added if needed to maintain timeline
Data cannot distinguish between models	Science questions will be refocused
Self -consistent codes to take too much time to run	Science questions will be refocused
IMAP does not fly/give good data to validate model	IBEX and Cassini data will be used for model verification

Risk Management and Conflict Resolution: Clear procedures are in place to mitigate risks and manage and resolve conflict. The CD and PM have successfully resolved conflicts, e.g., differences about publications, through consensus building and shared decision making. As conflicts arise and initial recommendations are made, the CD and PM solicits comments from EC. In the unlikely event differences cannot be resolved through consensus building, the EAB is engaged to reach an outcome that protects NASA's investment. RTs that fail to deliver against milestones will be terminated or redirected, although SHIELD management will be mindful of individual researchers' career development. The CD, in consultation with the EC and the EAB, may reallocate resources, re-direct funding between thrusts, and add, remove, or re-assign Directors. Advice and concurrence from NASA will be sought when required. SHIELD's full risk mitigation plan is articulated in its Strategic Plan (C7); select operational risks are noted in **Table 3**.

As established, SHIELD's research is tightly integrated. This coupling is an inherent risk. The dependency of the various thrusts and tasks, and in some cases, their linear nature, could potentially delay or stop work in other areas if key tasks are not completed on time. The CD has funding in reserve (CERIF) and will make available the necessary resources (human, financial, technical) to advance critical tasks. The monthly EC meetings will enable early identification and mitigation of challenges and risks.

7.4 Integration

SHIELD is designed to assimilate current knowledge of the outer heliosphere into a single model while identifying areas where our understanding of the physics is lacking and filling these gaps. This research requires close communications within each thrust and a close integration of efforts between the thrusts to achieve the Center's vision (**Figure 5**). Thrust integration is advanced by active participation of the EC in Center activities. The CD and PM participate in all thrusts, the development of the SHIELD model and BI. EAB member Vertisi, who has published extensively on what institutions and flows of people are required (Vertesi, 2020) to craft scientific knowledge, advised SHIELD on our broad approach. We exploit team science, support multi-directional information flow and leverage established infrastructure and managerial experience to realize objectives. The PM plays a major role in establishing and ensuring that communication lines remain open and efficient, that data is constantly collected and fully up to date, and that research and BI thrusts are on schedule and achieving specific milestones. Additional points of integration are SHIELD's students and young investigators and supported BI activities. Training opportunities are provided for young investigators, post-doctoral researchers, graduate students and undergraduates, who are not only mentored in the practice of team science but have opportunities to pursue collaborative research through the Central Education, Recruitment, and Impact Fund (CERIF), discussed below. To promote integration, SHIELD allocated funds to send postdocs and students to partner institutions to support cross-fertilization and new skill acquisition.

7.5 Collaboration plan

SHIELD will rely on the extensive research from organizational science and the science of team science (Guimera et al., 2005; Jones et al., 2008; Börner et al., 2010; NRC, 2015), to advance our team-based approach and address key challenges. Our management model supports collaboration at every level; top down, bottom up, and horizontally. SHIELD has already established a culture of collaboration as evidenced by its Phase I accomplishments and publications. As discussed in BI, SHIELD will provide training for young investigators focused on team science, reflexivity training, team coordination, and communication. Effective communication and collaboration are central to integrating SHIELD's research, building and supporting teams, and developing strategic directions. **Table 4** shows our meeting schedule, which was used successfully in Phase 1 and will be maintained in Phase II. Frequent interactions keep collaborations active and ensure progress. Quarterly Center-wide virtual seminars are another vehicle for collaboration, knowledge sharing and engagement with the wider heliospheric community. State-of-the-art technology enables productive virtual meetings, facilitates long-distance collaboration and provides platforms for file storage and sharing.

7.6 Coordination, engagement with broader community and new members

SHIELD recognizes the urgent need to build a broad coalition to realize its ambitious vision and longer-term sustainability goals (IHA). We welcome and invite participation of outside researchers to help realize strategic objectives. In Phase I, for instance, SHIELD established a workgroup to investigate Compton-Getting (CG) observations, which included members from two RTs and three members from outside SHIELD (identified by PM Richardson). The CG group is responsible for two SHIELD publications (Cummings et al., 2021; Richardson et al., 2021). Engagement with Broader Community: SHIELD seeks to attract talented researchers whose interests match or expand SHIELD and contribute to a Center culture that is diverse with respect to gender, age, racial/ethnic background, and sexual orientation. SHIELD will: 1) maintain an

TABLE 4 Communications meeting schedule.

Party	Frequency
Director and PM	Biweekly
Director, PM and RT Touchbase	Monthly
RT Team Communications	Biweekly
Coupling Team	Weekly
BI Team Communications	Weekly
Executive Committee	Monthly
Strategic Planning	Triannual
Virtual Seminar	Quarterly
Postdoc Advisory Committee	Quarterly
External Advisory Board	Annually
Evaluation	Annually
"All Hands Meeting"	Annually

active public website to share information and progress as well as enable researchers to join as "affiliate" members; 2) engage the broader community by making 1 day of its annual meetings a workshop open to everyone; 3) use the IHA to engage the broader community; and 4) invite experts and role models to participate in BI activities (**Section 8.4**).

7.7 Evaluation and continuous improvement

Evaluation: All scientific and BI activities (education, training, diversity, and outreach) will be assessed internally and externally. The EC and the EAB will evaluate the research; an external evaluator will assess BI activities and determine if SHIELD is fulfilling its goals. Each thrust will submit an annual report that summarizes progress in its scientific and BI activities and proposes plans for the following year. These documents will form the basis of SHIELD's annual report. BI programs will be evaluated at the start of the program for baseline assessments; annual surveys and interviews with SHIELD participants will provide summative data related to the overall program success. Performance indicators used by the EC to evaluate the status of the RTs include: 1) progress towards SHIELD's milestones, 2) scientific and technological advances (as measured by publications, conference presentations, invited presentations and seminars, impact factors); 3) level of interaction both within the RT and with other RTs; 4) evidence of mentorship and training of young scientists. RT Directors will provide quarterly reports listing notable advances and difficulties as well as publications (See Decision Making and Risk Management). Criteria used to evaluate the status of the BI activities include: 1) the number and diversity of participating faculty, scientists, young investigators, and students at all levels; 2) inclusion of the participation of women, URMs, persons with disabilities, and LGBTQIA + individuals in BI activities; 3) integration into Center research; and 4) evidence of effective inter-institutional collaboration. The EAB will complete an independent assessment and report findings and recommendations. SHIELD will use all findings to enable continuous improvement. Continuous Improvement: The EC will solicit seed proposals each year to bring new ideas, expertise, and personnel, including students, to the Center through CERIF. CERIF: This new initiative will support 5 graduate students (1 allocated to IHA) to establish direct linkages with collaborators to deepen appreciation for SHIELD's team-based science. The EC will evaluate proposals with advice from the EAB; final selection of the awarded graduate students will be made by the CD, with a preference for diverse researchers. Applicants will be encouraged and funded to visit and interact with SHIELD partners. After 1–2 years, the results of these seed projects will be reviewed, after which the activity will either be incorporated into the RT or BI or phased out.

7.8 Living strategic plan

The strategic plan (SP) is a living document. It is updated annually. The CD and PM lead strategic planning and assessment exercises and convene the EAB, PSAB and SB annually. SHIELD's SP provides a detailed research plan, timeline and risk mitigation strategies.

7.9 Lessons learned

An important lesson learned from Phase I is the centrality of team-based science. SHIELD established team processes, prioritized team composition and supported virtual collaboration. The team realized relatively early the need to have flexible funding to support team science, which resulted in *CERIF* in Phase II. CERIF will support students, seed projects, research exchanges, and will be used to manage risks. A second lesson learned is the need to elevate the Center's knowledge transfer and professional development activities. SHIELD lacked a formal mechanism to develop young and diverse scientists. The new Bench-Strength-Development program provides multi-level and reverse mentoring, training, exchanges, and professional development, and prepares each young scientist to become an independent investigator able to execute team-based science within SHIELD and any other future endeavors. Another lesson is the need to focus dually on both the Center's 5-year goal and longer-term vision. In Phase I, the EAB urged a spin-off Institute to ensure broad impact and the continuation of SHIELD well beyond the 5 years of the DSC. In Phase II, SHIELD will advance the IHA as discussed in **Section 1.3** and **Section 7.3**, Oversight. The CD, PM and DDs will work in concert with the Sustainability Board to develop a workshop series broadly focused on habitability. SHIELD's model will be open source and archived with CCMC to remain a living tool for the community.

8 Broadening impacts

8.1 Goal

Figure 13 shows the overall structure of the Broadening Impacts activites of SHIELD. In Phase I, SHIELD launched signature education and outreach programming, spanning K-12 to young investigators, in pursuit of its vision to change the face of

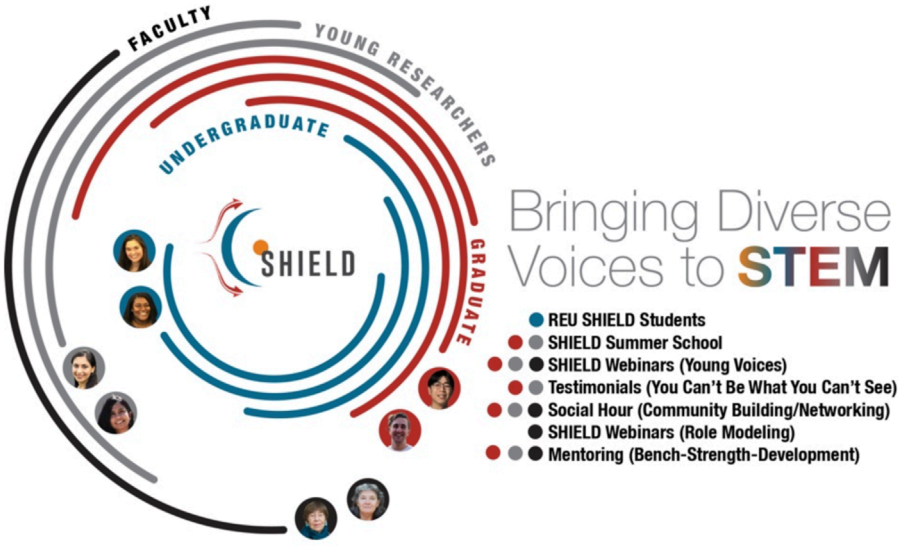


FIGURE 13
SHIELD BI initiatives.

heliophysics. As part of a campaign to bring diverse voices to STEM, SHIELD initiated You Can't Be What You Can't See testimonials and Creating Science Quartets (Box 2) webinars. This innovative programming, the brainchild of CD Opher and Phase I BI Director Wong, collects stories and testimonials from graduate students, young researchers and their mentors on finding community and overcoming challenges. Testimonials have touched on topics such as the impostor syndrome, gender fluidity, the immigrant experience, and the impact learning disabilities can have on trajectories in STEM fields. SHIELD's webinars profile eminent scientists and trailblazers, emerging scholars, and NASA managers who talk on career development, space physics, and scientific discovery. Notable speakers include: Fran Bagenal (UC-Boulder), Suzy Dodd (NASA), Nicky Fox (NASA), Margaret Kivelson (UCLA), Andrea Razzaghi (NASA), and Tom Krimigis (APL). The reception from these heart-warming stories and un-shielded approach to outreach has been overwhelming. Phase II builds upon this rich and successful tapestry with new programming focused on broadening participation, professional development and mentoring, and knowledge sharing. Specific aims are to: 1) increase the recruitment, inclusion, and retention of traditionally underrepresented (URM) groups (with a predominate focus on racial/ethnic minorities, women, LGBTQIA+, and first-generation college students) pursuing STEM careers and entering post-secondary education; and 2) train, mentor, and build leadership skills for emerging and early career scientists, teaching team-based science. CD Opher is one of the very few women in space physics. A non-US born citizen who identifies as LGTBQ, Opher knows first-hand the challenges of working and succeeding in this predominately masculine field. She and BI Director Buxner are personally vested in nurturing a Center culture that is inclusive, promotes a sense of belonging, and provides academic and personal supports to retain URM students in particular. For more details on the broader BI effort see Buxner et al., 2023.

8.2 Motivation

The heliophysics community, like most other science communities, has diversity and inclusion problems (NRC, 2013), particularly around gender and ethnicity. The physics workforce is majority white and male; it is also aging and retirement-eligible, necessitating the need to accelerate efforts to train the next-generation of space physics workers. Professional societies advise science education better to prepare, "*future scientists with the knowledge and skills required for effective knowledge integration and collaboration within a science team or larger group*" (NRC, 2015) while establishing cultures where students from marginalized communities can thrive. These challenges drive SHIELD's education and human-resource-development, broadening participation and outreach goals to train a new type of heliophysicist, one fluent in team science and capable of working in highly transdisciplinary, collaborative environments and make measurable improvements in the diversity of the heliophysics workforce pathway.

8.3 Team

Sanlyn Buxner, Senior Research Scientist at the Planetary Science Institute and Director of Graduate Studies in Teaching, Learning, and Sociocultural Studies at UA, will be the BI Director. Buxner brings over 15 years of experience recruiting and supporting diverse undergraduate and graduate students. Nicholas Gross (BU) is the Deputy Director of BI and has extensive education experience within the space physics community, including past service as an Education co-director for a NSF Science and Technology Center. He will organize SHIELD's REU program and provide critical support to the Summer School. Experts supporting the core team are Sarah Hokanson (BU-Provost Office) who will focus on mentoring and Cherilyn Morrow (PUNCH Mission), who will focus on

Box 2 Select Phase I BI Accomplishments

Broadening Impacts.

- Launched 'pathway for scientists' series, *Creating Science Quartets: Bringing Diverse Voices*, with enrollment exceeding 100. Upcoming seminars focus on the Latin and Black experience in STEM, respectively.
- Established *You Can't Be What You Can't See* initiative, a video/testimonial series showcasing diverse PhD students and junior researchers in space physics. 9 stories profiled.
- Established monthly webinar series (12 webinars as of 10/21), focused on *career development* (improving writing skills and delivering effective scientific presentations), *universal themes* such as imposter syndrome and immigration, and *Young Voices* (a platform for PhD students and postdocs to discuss research and career paths).

Communication and Outreach.

- Launched website: Viewed by 2000 users since April 2020.
- Established social media presence (Facebook, YouTube).
- Hosted virtual 'happy hours' to build community.
- Interviewed by popular news outlets, including *NPR*.

Knowledge Sharing.

- Published 12 joint articles (4 or more senior investigators).
- Appointed four young scientists, 2 females, to co-lead RTs.

communication. Judi Burgess, Director of Labor Relations (BU), serves as an advisor on diversity programming.

8.4 Approach

Aim 1: Broadening Participation: SHIELD proposes to increase the inclusion of URM groups (women, racial/ethnic minorities, LGBTQIA+, individuals with a self-reported disability, first generation) pursuing STEM careers and entering post-secondary education; 2) increase retention and promote development of female and URM young investigators through professional development, mentoring, and skills training; and 3) provide diversity, equity, inclusion and access (DEIA) training and other tools to promote inclusive learning environments. The centerpiece of this effort is the SHIELD Distributed REU Program. Undergraduate research experiences have raised awareness of the discipline; fifty (50%) of space physics graduate students were involved in an undergraduate research experience (NRC, 2013). SHIELD will support a summer REU program.

Gross and Buxner will recruit and support 5 diverse undergraduate students each year to participate in a 12-week research experience at BU, UAH, or UM. Over 5 years, ~25 undergraduates will be supported, with at least one-third of students being from a targeted URM groups (female, racial/ethnic minorities, LGBTQIA+, or persons with disability. SHIELD will partner with other successful space science training programs including L'SPACE, which has over 5,000 alumni, over half of whom attended or graduated from a MSI.

AIP identified three major factors critical to supporting and retaining groups underrepresented in STEM. These include: 1) the need for students to feel a sense of belonging, 2) the need to perceive themselves, and be perceived by others, as future physicists and astronomers, and 3) the need for effective teaching and mentoring students.

To organize this distributed summer REU program, SHIELD will leverage existing REU programs at BU, UAH, and UM. This approach will permit students to use existing housing arrangements and participate in social activities. Each REU student will have a mentoring plan and at least one mentor. SHIELD will work with these participants to build a virtual REU SHIELD community to collaborates on space science research. Early career scientists and graduate researchers engaged in SHIELD will join the mentoring team for the REU students. REU students will be expected to produce an abstract and poster and funded to present their work at a professional conference. REU participants also will receive instruction about the SHIELD research enterprise, ethics, scientific writing, presentation, and pathways to graduate school or other STEM careers. Mentorship and leadership skill development will be accomplished via SHIELD's new Bench-Strength-Development (BSD) program.

SHIELD will also establish a DEIA Action Workgroup to advise on center climate issues and recommend policies, procedures and practices that will bring about more equitable outcomes for everyone. Burgess will advise on the development of this workgroup. She has significant relationships with HBCUs and served with CD Opher on BU Diversity and Inclusion (D&I) Action Team. Finally, SHIELD will train its leaders to become DEIA advocates. Training in overall DEIA efforts will be conducted by Hokanson each quarter and at the annual team meeting.

Aim 2: Education and Human Resources Development (EdHRD): EdHRD programs are built around four initiatives: 1) training/skills development; 2) SHIELD Summer School; 3) Bench-Strength-Development; and 4) Mentoring/mentoring training. Training/Skills Development: A suite of programs will support and train new, emerging, and established SHIELD scientists, while furthering BP goals of the center. Professional Development

(PD): PD courses for SHIELD students, early career and other SHIELD members will cover topics in science, communication, DEIA, mentoring, and management. Onboarding Training: All associated students and early career scientists will attend training in team science and working in transdisciplinary collaborative environments led by Opher, Buxner and Gross. Each year, new students and scientists will develop a common set of foundational skills while building teamwork and interpersonal connections. Other training will focus on reflexivity training, team coordination, and communication. Skills Training: Organized by Buxner, ongoing online interactive workshops will include (a) publication and presentation preparation; (b) proposal writing; (c) organizing conference sessions; (d) networking; and (e) outreach to internal and external stakeholders. Students will outline and chart career paths as they work toward presenting research to peers in annual meetings such as the AGU. Communication Training will focus on communicating effectively with scientific and non-scientific audiences. In addition to the resources of the SHIELD team, Morrow (PE lead, PUNCH Mission) will provide a workshop series on communicating with broad audiences. Students and early career scientists will have multiple opportunities to present at outreach events, annual meetings and conferences. Summer School: The Heliophysics SMD Decadal Survey calls out summer school as an important opportunity to train graduate students in the concepts and tools of space physics. Current summer school opportunities are routinely oversubscribed, suggesting demand for similar programming. SHIELD will offer a Summer School at BU building on the prior experience of CD Opher [who ran a successful NSF CAREER Summer School (Opher, 2012)] and Gross (an organizer and instructor for both the LWS Heliophysics Summer Schools and the Boulder Space Weather Summer School). Recruitment will be done through minority professional societies and MSIs as well as L'SPACE. We expect that 20 participants a year will attend the Summer School. Our target is to have 50% of the participants be from URM groups. SHIELD's Summer School will adopt proven techniques from undergraduate education research including flipped classroom (Abeysekera & Dawson 2005, peer instruction (Crouch and MazurPeer, 2001), tutorials, and project based learning (https://www.pblworks.org/what-is-pbl). Bench Strength Development (BSD): Creating a strong and diverse bench of talent is vitally important for the success and sustainability of SHIELD. BSD is a new Phase II initiative to bring young investigators (YIs) into leadership roles within SHIELD and provide training and support to prepares postdocs and early career faculty to lead team-based science efforts. As noted in **Section 2**, RTs are led by a senior and junior investigator (Deputy); two of the six deputies are non-US born females. YIs will co-mentored by a RT Director and a co-I from another institution, supporting cross-training. BSD programming will enable YIs to: 1) increase content knowledge and gain technical skill in a team-based environment; 2) receive mentorship; 3) develop leadership skills; 4) enhance oral and written communication and presentation skills; and 5) build internal and external networks. Through BSD, reverse mentoring is also supported. YIs will provide guidance, strategy, and new ways of thinking to senior investigators through co-leadership of thrusts, presentations to EC and EAB, and participation in annual meetings (**Sections 1** and **2**). The CD and RT Directors will oversee and implement BSD; the BI Director will provide input. Mentoring and Mentoring Training: Mentoring and professional development are critically important components of SHIELD's EdHRD. SHIELD will support multi-mentoring (De Janasz & Sullivan, 2004) and make available to postdocs and graduate students at least two mentors from different institutions. Mentors will be assigned by the EC. Creation of specific annual goals, meeting schedules, deliverables, and measures of progress will help to ensure effective impact and positive outcomes. Hokanson in collaboration with Gross and Buxner will run a 2-hour *Inclusive Mentoring Workshop* at the SHIELD annual meeting to improve research mentorship skills, from developing individual development plans to establishing and maintaining productive mentoring relationships with diverse mentees. Hokanson will conduct two more annual trainings for new team members and mentees. Training also will be provided for mentees with monthly check-ins facilitated by Gross and Buxner, and a yearly evaluation by the External Evaluator. All mentees, including DDs, postdocs, CERIF-supported graduate students, and REU students, will have a mentoring plan on file with SHIELD and will be included in the evaluation effort. Mentoring will be assessed using research-based instruments (Fleming et al., 2013) and Social Network Analysis (Ginexi, 2017).

Aim 3: Heliophysics Community Building and Outreach: SHIELD's community building activities include formal and informal interactions to address issues unique to intersectionality. Building on Phase I, we will continue to sponsor a broad set of activities open to the whole space physics community with specific focus on undergraduate and graduate students, and post-doctoral researchers. Webinars: The Webinar series (Box 2) will continue. It will be expanded to include speakers from all adjacent fields of heliophysics and target 10 webinars a year divided between "Young Voices" and "Creating Science Quartets." Public Outreach: SHIELD will provide emerging scientists (graduate students, postdocs) and early career scientists with opportunities to present their work and their experience in becoming a scientist. Morrow, in collaboration with Gross and Buxner, will conduct regular communications training for emerging and early career scientists. This will be part of the professional development workshops offered over the course of SHIELD. Podcasts: Facilitated by Buxner, SHIELD scientists will join an ongoing initiative of the PUNCH Mission public engagement program to participate in SciHArt, a podcast series facilitated by student (young professional) hosts who work with the Fiske Planetarium at the UC Boulder. The primary audience is STEM-interested learners at the late high-school, undergraduate, and early graduate levels. SciHArt interviews leaders in science, engineering, and science communication who are in different phases of their career journey, from undergraduate researchers to senior professionals playing leadership roles on NASA missions. SHIELD will provide up to five scientists a year for interviews, beginning with each Deputy Director and including early career scientists. Likewise, we will expand the reach of the SciHArt audience through our outreach networks. Video Testimonials: SHIELD will continue to support its "You Can't Be What You Can't See" video/testimonial series, which profile and provide support for emerging scientists from diverse backgrounds. In Phase II, SHIELD will invite researchers from the heliophysics community to video interviews regarding their experience, the challenges and barriers they faced, and how they persevered. These videos are important for inclusivity building and are featured on our website and Youtube channel.

SHIELD will target 5 videos a year. Website: The SHIELD website (https://shielddrivecenter.com) will be maintained/expanded. As the public face of the Center, the website will feature SHIELD research, scientific highlights, publications, access BI programming and information to become an affiliate SHIELD member.

9 Models

9.1 SHIELD global model

The SHIELD global model consists of the Outer Heliosphere (OH) component of SWMF code (UM and BU), the Particle Tracker (PT) component FLEKS (UM) (that is being currently developed), the particle transport code SPECTRUM (UAH), and the Alfven Wave Solar atmosphere Model (AWSoM) that simulates turbulence in the solar wind. (b) These codes are coupled to provide a comprehensive self-consistent model of the Outer Heliosphere (c) The SHIELD model will produce multi-component number densities, temperatures, flow velocities, particle distributions, and magnetic fields with both 3D, 2D and line plots. 3D file sizes are around 10 GB (SWMF/PT). File formats are both in binary and ascii. The restart files for a typical three-dimensional domain can reach 100 GB. (d) Several 3D and 2D files are generated per run and stored at NASA Ames Pleiades. Each person performing the runs will be responsible for the long-term storage (as in lou NASA Ames); (e) Documentation: SWMF is documented in Latex and XML that are compiled into PDF manuals. The source code is also documented. The full SWMF is on the UM GitLab server (http://gitlab.umich.edu/swmf_software), and a major part of it, including BATS-R-US used in this project is mirrored to GitHub (https://github.com/MSTEM-QUDA) under a non-commercial open-source license. The Opher group will produce the documentation for OH, Chen for FLEKS/PT, and Florinski for SPECTRUM. All will be available at Gitlab.; (f) SHIELD output will be delivered Years 1-5 to CCMC. As each milestone is completed; (g) The entire model will be delivered to CCMC.

9.2 Local models: SHIELD uses a hybrid model and a particle-in-cell code

[A] (a) The UA hybrid model describes particle evolution in turbulence.; (b) It derives the particle distributions needed for the global model. (c, d) Output and data volume: 1) multi-component number densities, flow velocities, magnetic-field vectors, electric-field vectors, energetic particles fluxes in several energy ranges, as color-coded images plotted over the entire 2D spatial domain at selected simulation times. Each simulation produces 10–20 50B images. 2) reduced 1D distribution functions vs. momentum and flux vs. energy for each ion species, averaged over portions of the simulation domain at several simulation times. There are 20–30 100 kB files in ascii format 3) Reduced 1D line plots of plasma parameters and energetic particle fluxes as a function of a given spatial direction in the simulation domain for several pre-chosen cuts through the domain. There are 10–20 1 Mb ascii. 4) A table of simulation parameters, and conversions for each simulation performed in excel format (<100 kB); (e) planned documentation will be produced and available at Gitlab (f) Year 5 (g) Archived at CCMC.

[B] PIC code p3d: (a) A particle-in-cell (PIC) code that gives a nearly complete physical description of a plasma and accurately simulates processes at kinetic scales. (b) SHIELD uses it to study small-scale processes like shock acceleration and reconnection. (c) The data products (binary) are snapshots of number densities, temperatures, magnetic fields, and particle distributions in time and restart files can be used to advance a simulation from a particular point in time. (d) A single three-dimensional PIC run can produce 50 TB of data. e) Planned documentation will be produced and available at Gitlab f) These will be stored as runs are made in years 1–5. (g). Due to its volume such data are usually stored at the facility where they are produced. The High Performance Storage System (HPSS) at the National Energy Research Scientific Computing Center (NERSC) has been used for archival storage since 1998.

[C] Data Outputs: We plan to use only publicly available spacecraft data, so archiving is not needed.

9.3 Data reduction, scientific analysis, or modeling software (including model output readers and interpolators)

[1] SHIELD uses SWMF/OH (fortran); PT (FLEKS): C++; SPECTRUM: (C++)which all will transition to open source; The SWMF is developed and maintained under Git, which is a distributed version control system. The full SWMF is on the U of M GitLab server (http://gitlab.umich.edu/swmf_software), and a major part of it, including BATS-R-US used in this project, is mirrored to GitHub (https://github.com/MSTEM-QUDA) which provides open-source access under a non-commercial license. The OH component will be released as open source by Opher's group. The UAH transport code SPECTRUM will track CR and PUI trajectories and couple to the SWMF/OH and PT (FLEKS code). The code is written in C++ with elements of CUDA. It will be released under the MIT open-source license and made into a GitHub project, which will ensure its free accessibility by the community. The code is already documented using Doxygen; a user manual will be written before the code is released. SWMF/FLEKS is being developed now (Chen et al., 2021) and will be openly accessible at the U of M GitLab server for download as well as in the CCMC. OH application will be released as open-source by Opher's group.

[2] Hybrid model: Fortran code that manipulates distribution functions into other forms and reduces them to produce densities will be placed in CCMC. The hybrid code will be available on github as open-source access under a non-commercial license as part of the proposal.

[3] PIC code: p3d is primarily written in Fortran and uses the C and m4 pre-processors. Parallelization is done via MPI. A user manual will be written before the code is released. Several Fortran and IDL programs that can be used to read and analyze the data will be uploaded as well. p3d will be available on as open-source access under a non-commercial license as part of the proposal.

9.4 Roles and responsibilities of team members for data management

The data management and coordination of data acquisition, analysis, and archiving of is overseed by the P.I. (Prof. Merav Opher). The Coupling Code Group director Dr. Gabor Toth will oversee the task to make the SHIELD model open source. The UAH postdoc will aid with SPECTRUM. Marc Kornblueth from BU will aid with SWMF/OH. Localized models: The UA Postdoc will lead the archiving of the hybrid code. For p3d, the team member that performs a numerical simulation will be responsible for archiving the data. That includes saving the data, as well as any codes necessary for interpreting it, in an accessible format to the relevant high-performance storage system.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work is supported by NASA grant 18-DRIVE18_2-0029 as part of the NASA/DRIVE program titled "Our Heliospheric Shield", 80NSSC22M0164, https://shielddrivecenter.com. Some of the figures were produced by AH, Graphic Designer, adamhong.com.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abeysekera, L., and Dawson, P. (2005). Motivation and cognitive load in the flipped classroom: Definition, rationale and a call for research. *High. Educ. Res. Dev.* 34 (1), 1–14. doi:10.1080/07294360.2014.934336

Axford, W. I., Dessler, A. J., Gottlieb, B., Dessler, A. J., and Gottlieb, B. (1963). Termination of solar wind and solar magnetic field. *Astrophysical J.* 137, 1268. doi:10.1086/147602

Baranov, V. B., and Malama, Y. G. (1993). Model of the solar wind interaction with the local interstellar medium: Numerical solution of self-consistent problem. *J. Geophys. Res.* 98 (A9), 15157–15163. doi:10.1029/93JA01171

Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S. M., Hall, K. L., Keyton, J., et al. (2010). A multi-level systems perspective for the science of team science. *Sci. Transl. Med.* 2 (49), 3001399. doi:10.1126/scitranslmed.3001399

Burlaga, L. F., Florinski, V., and Ness, N. F. (2015). *In situ* observations of magnetic turbulence in the local interstellar medium. *Astrophysical J.* 804, L31. doi:10.1088/2041-8205/804/2/L31

Burlaga, L. F., Florinski, V., and Ness, N. F. (2018). Turbulence in the outer heliosheath. *Astrophysical J.* 854, 20. doi:10.3847/1538-4357/aaa45a

Burlaga, L. F., Ness, N. F., Berdichevsky, D. B., Jian, L. K., Park, J., and Szabo, A. (2020). Voyager 1 and 2 observations of a change in the nature of magnetic fluctuations in the VLISM with increasing distance from the heliopause. *Astronomical J.* 160, 40. doi:10.3847/1538-3881/ab94a7

Burlaga, L. F., and Ness, N. F. (2016). Observations of the interstellar magnetic field in the outer heliosheath:voyager 1. *Astrophysical J.* 829, 134. doi:10.3847/0004-637x/829/2/134

Burlaga, L. F., Ness, N. F., and Stone, E. C. (2013). Magnetic field observations as voyager 1 entered the heliosheath depletion region. *Science* 341 (6142), 147–150. doi:10.1126/science.1235451

Buxner, S., Gross, N., Opher, M., Wong, J., and Richardson, J. (2023). SHIELD DRIVE science center: Efforts in diversification and inclusion in heliophysics. *Front. Space Sci. Astrophysics* 10, 1155843. in press. doi:10.3389/fspas.2023.1155843

Cairns, I. H., and Fuselier, S. A. (2017). The plasma depletion layer beyond the heliopause: Evidence, implications, and predictions for*voyager 2*and*new Horizons*. *Astrophysical J.* 834, 197. doi:10.3847/1538-4357/834/2/197

Chalov, S. V., Fahr, H. J., and Izmodenov, V. (1997). Spectra of energized pick-up ions upstream of the two-dimensional heliospheric termination shock. II. Acceleration by Alfvenic and by large-scale solar wind turbulences. *Astronomy Astrophysics* 320, 659.

Chalov, S. V., and Fahr, H. J. (2000). Pick-up ion acceleration at the termination shock and the post-shock pick-up ion energy distribution. *Astronomy Astrophysics* 60, 381.

Chen, Y., Toth, G., Zhou, H., and Wang, X. (2021). Fleks: A flexible particle-in-cell code for multi-scale plasma simulations. ESS Open Archive. December 09. Avaialabe at: https://essopenarchive.org/doi/full/10.1002/essoar.10508070.3.

Crouch, C., Mazur, E., and Peer, I. (2001). Peer Instruction: Ten years of experience and results. *Am. J. Phys.* 69, 970–977. doi:10.1119/1.1374249

Cummings, A. C., Stone, E. C., McDonald, F. B., Heikkila, B. C., Laf, N., and Webber, W. R. (2008). Anomalous cosmic rays in the heliosheath. *AIP Conf. Proc.* 1039, 343.

Cummings, A. C., Stone, E. C., Heikkila, B. C., Lal, N., Webber, W. R., Jóhannesson, G., et al. (2016). Galactic cosmic rays in the local interstellar medium:*voyager 1*observations and model results. *Astrophysical J.* 831, 18. doi:10.3847/0004-637x/831/1/18

Cummings, A. C., Stone, E. C., Richardson, J. D., Heikkila, B. C., Lal, N., and Kota, J. (2021). *Astrophysical J.* 906 (2), 126. doi:10.3847/1538-4357/abc5c0

Dahlin, J., Drake, J. F., and Swisdak, M. (2017). The role of three-dimensional transport in driving enhanced electron acceleration during magnetic reconnection. *Phys. Of Plasmas* 24 (9), 092110. doi:10.1063/1.4986211

Dahlin, J. T., Drake, J. F., and Swisdak, M. (2016). Parallel electric fields are inefficient drivers of energetic electrons in magnetic reconnection. *Phys. Of Plasmas* 23 (12), 120704. doi:10.1063/1.4972082

Davis, L. (1955). Interplanetary magnetic fields and cosmic rays. *Phys. Rev.* 100 (5), 1440–1444. doi:10.1103/physrev.100.1440

de Janasz, S. C., and Sullivan, S. E. (2004). Multiple mentoring in academe: Developing the professorial network. *J. Vocat. Behav.* 64, 263–283. doi:10.1016/j.jvb.2002.07.001

Decker, R. B., Krimigis, S. M., Roelof, E. C., and Hill, M. E. (2010). Variations of low-energy ion distributions measured in the heliosheath. *AIP Conf. Proc.* 1302, 51. doi:10.1063/1.3529990

Dessler, A. J. (1967). Solar wind and interplanetary magnetic field. *Rev. Geophys. Space Phys.* 5, 1. doi:10.1029/rg005i001p00001

Dialynas, K., Galli, A., Dayeh, M. A., Cummings, A. C., Decker, R. B., Fuselier, S. A., et al. (2020). Combined ~10 eV to ~344MeV particle spectra and pressures in the heliosheath along the voyager 2 trajectory. *Astrophysical J.* 905 (2), L24. doi:10.3847/2041-8213/abcaaa

Dialynas, K., Krimigis, S. M., Mitchell, D. G., Decker, R. B., and Roelof, E. C. (2017). The bubble-like shape of the heliosphere observed by Voyager and Cassini. *Nat. Astron.* 1, 0115. doi:10.1038/s41550-017-0115

Drake, J., Opher, M., Swisdak, M., and Chamoun, J. N. (2010). A magnetic reconnection mechanism for the generation of anomalous cosmic rays. *Astrophysical J.* 709, 963–974. doi:10.1088/0004-637x/709/2/963

Drake, J., Swisdak, M., and Opher, M. (2015). A model of the heliosphere with jets. *Astrophysical J. Lett.* 808 (2), L44. doi:10.1088/2041-8205/808/2/l44

Elgart, S. (2021). *Presentation by SMD director elgart, S robin (JSC-SK4)*. Houston, Texas: Space Radiation Element Scientist at NASA Johnson Space Center.

Fisk, L. A., Kozlovsky, B., and Ramaty, R. (1974). An interpretation of the observed oxygen and nitrogen enhancements in low-energy cosmic rays. *Astrophysical J.* 190, L35. doi:10.1086/181498

Fleming, M., House, S., Hanson, V. S., Yu, L., Garbutt, J., McGee, R., et al. (2013). The mentoring competency assessment: Validation of a new instrument to evaluate skills of research mentors. *Acad. Med.* 88 (7), 1002–1008. doi:10.1097/acm.0b013e318295e298

Florinski, V. , F., Ghanbari, K., Kleiman, J., Opher, M., Giacalone, J., Kota, J., et al. (2021). Magnetic trapping of galactic cosmic rays in the outer heliosheath. *Astrophysical J.* 2023. in review.

Florinski, V., Decker, R. B., and le Roux, J. A. (2008). Pitch angle distributions of energetic particles near the heliospheric termination shock. *J. Geophys. Res.* 113 (7), A07103. doi:10.1029/2007JA012859

Florinski, V., Ferreira, S. E. S., and Pogorelov, N. V. (2013). Galactic cosmic rays in the outer heliosphere: Theory and Models. *Space Sci. Rev.* 176 (1), 147–163. doi:10.1007/s11214-011-9756-1

Florinski, V., Heerikhuisen, J., Niemiec, J., and Ernst, A. (2016). Theibexribbon and the pickup ion ring stability in the outer heliosheath. I. Theory and hybrid simulations. *Astrophysical J.* 826, 197. doi:10.3847/0004-637x/826/2/197

Florinski, V. (2015). Magnetic flux tube interchange at the heliopause. *Astrophysical J.* 813, 49. doi:10.1088/0004-637x/813/1/49

Florinski, V., and Pogorelov, N. V. (2009). FOUR-DIMENSIONAL transport of galactic cosmic rays in the outer heliosphere and heliosheath. *Astrophysical J.* 701 (1), 642–651. doi:10.1088/0004-637x/701/1/642

Florinski, V., Zank, G. P., Heerikhuisen, J., Hu, Q., and Khazanov, I. (2010). Stability of a PICKUP ION ring-beam population in the outer heliosheath: Implications for theibexribbon. *Astrophysical J.* 719 (2), 1097–1103. doi:10.1088/0004-637x/719/2/1097

Florinski, V., Zank, G. P., Jokipii, J. R., Stone, E. C., and Cummings, A. C. (2004). Do anomalous cosmic rays modify the termination shock? *Astrophysical J.* 610, 1169–1181. doi:10.1086/421901

Florinski, V., Zank, G. P., and Pogorelov, N. V. (2003). Galactic cosmic ray transport in the global heliosphere. *J. Geophys. Res.* 108 (6), 1228. doi:10.1029/2002JA009695

Fraternale, F., Pogorelov, N. V., and Burlaga, L. F. (2020). Signatures of intermittency and fine-scale turbulence in the very local interstellar medium. *Astrophysical J. Lett.* 897, L28. doi:10.3847/2041-8213/ab9df5

Fraternale, F., Pogorelov, N. V., Richardson, J. D., and Tordella, D. (2019). Magnetic turbulence spectra and intermittency in the heliosheath and in the local interstellar medium. *Astrophysical J.* 872, 40. doi:10.3847/1538-4357/aafd30

Fuselier, S. A., Galli, A., Richardson, J. D., Reisenfeld, D. B., Zirnstein, E. J., Heerikhuisen, J., et al. (2021). Energetic neutral atom fluxes from the heliosheath: Constraints from *in situ* measurements and models. *Astrophysical J. Lett.* 915 (2), L26. doi:10.3847/2041-8213/ac0d5c

Giacalone, J. (2012). *Astrophysical J.* 761, 28. doi:10.1088/0004-637x/761/1/28

Giacalone, J., and Jokipii, J. R. (1999). The transport of cosmic rays across a turbulent magnetic field. *Astrophysical J.* 520 (1), 204–214. doi:10.1086/307452

Giacalone, J., Nakanotani, M., Zank, G. P., Kota, J., Opher, M., and Richardson, J. D. (2021). Hybrid simulations of interstellar pickup protons accelerated at the solar-wind termination shock at multiple locations. *Astrophysical J. Lett.* 911 (8), 27. doi:10.3847/1538-4357/abe93a

Giacalone, J. (2005). The efficient acceleration of thermal protons by perpendicular shocks. *Astrophysical J.* 628, L37–L40. doi:10.1086/432510

Ginexi, E., Huang, G., Steketee, M., Tsakraklides, S., MacAllum, K., Bromberg, J., et al. (2017). Social network analysis of a scientist–practitioner research initiative established to facilitate science dissemination and implementation within states and communities. *Res. Eval.* 26 (4), 316–325. doi:10.1093/reseval/rvx026

Gkioulidou, M., Opher, M., Kornbleuth, M., Dialynas, K., Dayeh, M. A., Zirnstein, E., et al. (2022). On the energization of pickup ions downstream of the heliospheric termination shock by comparing 0.52–55 keV observed energetic neutral atom spectra to ones inferred from proton hybrid simulations. *Astrophysical J. Lett.* 931, 21. doi:10.3847/2041-8213/ac6beb

Grygorczuk, J., Czechowski, A., and Grzedzielski, S. (2014). *Astrophysical J. Lett.* 789, L43. doi:10.1088/2041-8205/789/2/l43

Guimera, R., Uzzi, B., Spiro, J., and Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308 (5722), 697–702. doi:10.1126/science.1106340

Guo, F., Jokipii, J. E., and Kota, J. (2010). Particle acceleration by collisionless shocks containing large-scale magnetic-field variations. *Astrophysical J.* 725, 128–133. doi:10.1088/0004-637x/725/1/128

Guo, X., and Florinski, V. (2014). Galactic cosmic-ray modulation near the heliopause. *Astrophysical J.* 793, 18. doi:10.1088/0004-637x/793/1/18

Guo, X., Florinski, V., and Wang, C. (2018). Effects of anomalous cosmic rays on the structure of the outer heliosphere. *Astrophysical J.* 859, 157. doi:10.3847/1538-4357/aabf42

Gurnett, D. A., Kurth, W. S., Burlaga, L. F., and Ness, N. F. (2013). *In situ* observations of interstellar plasma with voyager 1. *Science* 41 (6153), 1489–1492. doi:10.1126/science.1241681

Gurnett, D. A., and Kurth, W. S. (2019). Plasma densities near and beyond the heliopause from the Voyager 1 and 2 plasma wave instruments. *Nat. Astron.* 3, 1024–1028. doi:10.1038/s41550-019-0918-5

Gurnett, D. A., Kurth, W. S., Stone, E. C., Cummings, A. C., Heikkila, B., Lal, N., et al. (2021). A foreshock model for interstellar shocks of solar origin: Voyager 1 and 2 observations. *Astronomical J.* 161 (1), 11. doi:10.3847/1538-3881/abc337

Gurnett, D. A., Kurth, W. S., Stone, E. C., Cummings, A. C., Krimigis, S. M., Decker, R. B., et al. (2015). Precursors to interstellar shocks of solar origin. *Astrophysical J.* 809, 121. doi:10.1088/0004-637x/809/2/121

Haverkorn, M., Brown, J. C., Gaensler, B. M., and McClure-Griffiths, N. M. (2008). The outer scale of turbulence in the magnetoionized galactic interstellar medium. *Astrophysical J.* 680, 362–370. doi:10.1086/587165

Heerikhuisen, J., Pogorelov, N. V., Zank, G. P., Crew, G. B., Frisch, P. C., Funsten, H. O., et al. (2010). PICK-UP IONS IN the outer heliosheath: A possible mechanism for the interstellar boundary EXplorer ribbon. *Astrophysical J.* 708, L126–L130. doi:10.1088/2041-8205/708/2/l126

Izmodenov, V., Malama, Y. G., and Ruderman, M. S. (2008). Modeling of the outer heliosphere with the realistic solar cycle. *J. Adv. Space Res.* 41, 318–324. doi:10.1016/j.asr.2007.06.033

Izmodenov, V., Malama, Y., and Ruderman, M. S. (2005). Solar cycle influence on the interaction of the solar wind with Local Interstellar Cloud. *Astronomy Astrophysics* 429, 1069–1080. doi:10.1051/0004-6361:20041348

Izmodenov, V. V., Alexashov, D. B., and Ruderman, M. S. (2014). Electron thermal conduction as A possible physical mechanism to make the inner heliosheath thinner. *Astrophysical J. Lett.* 795 (1), 3. article id. L7. doi:10.1088/2041-8205/795/1/L7

Izmodenov, V. V., and Alexashov, D. B. (2015). Three-dimensional kinetic-mhd model of the global heliosphere with the heliopause-surface fitting. *Astrophysical J. Suppl. Ser.* 220 (2), 14. article id 32. doi:10.1088/0067-0049/220/2/32

Jones, B. F., Wuchty, S., and Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* 322 (5905), 1259–1262. doi:10.1126/science.1158357

Jordan, J., and Sorell, M. (2019). Why reverse mentoring works and how to do it right. Harvard Business Review. Available at: https://hbr.org/2019/10/why-reverse-mentoring-works-and-how-to-do-it-right.

Keebler, T. B., Toth, G., Zieger, B., and Opher, M. (2022). MSWIM2D: Two-dimensional outer heliosphere solar wind modeling. *Astrophysical J. Suppl.* 260, 43. doi:10.3847/1538-4365/ac67eb

Kim, T. K., Pogorelov, N. V., and Burlaga, L. F. (2017). Modeling shocks detected by *voyager 1* in the local interstellar medium. *Astrophysical J. Lett.* 843, L32. article id. doi:10.3847/2041-8213/aa7b2b

Kornbleuth, M., Opher, M., Baliukin, I., Dayeh, M. A., Zirnstein, E., Gkioulidou, M., et al. (2021b). Signature of a heliotail organized by the solar magnetic field and the role of nonideal processes in modeled IBEX ENA maps: A comparison of the BU and Moscow MHD models. *Astrophysical J.* 921, 164. doi:10.3847/1538-4357/ac1e2a

Kornbleuth, M., Opher, M., Baliukin, I., Gkioulidou, M., Richardson, J., Zank, G., et al. (2021a). The development of a split-tail heliosphere and the role of non-ideal processes: A comparison of the BU and Moscow models. *Astrophysical J.* 923 (2), 13pp–179. doi:10.3847/1538-4357/ac2fa6

Kota, J. (2010). Particle acceleration at near-perpendicular shocks: The role of field-line topology. *Astrophysical J.* 23, 393.

Krimigis, S. M., Mitchell, D. G., Roelof, E. C., Hsieh, K. C., and McComas, D. J. (2009). Imaging the interaction of the heliosphere with the interstellar medium From saturn with cassini. *Sci.* 326 (5955), 971.

le Roux, J. A., Webb, G. M., Shalchi, A., and Zank, G. P. (2010). A generalized nonlinear guiding center theory for the collisionless anomalous perpendicular diffusion of cosmic rays. *Astrophysical J.* 716, 671. doi:10.1088/0004-637X/716/1/671

Lee, M. A., and Ip, W.-H. (1987). Hydromagnetic wave excitation by ionized interstellar hydrogen and helium in the solar wind. *J. Geophys. Res.* 92 (10), 11041.

Malama, Y. G., Izmodenov, V. V., and Chalov, S. V. (2006). Modeling of the heliospheric interface: Multi-component nature of the heliospheric plasma. *Astronomy Astrophysics* 445, 693–701. doi:10.1051/0004-6361:20053646

McComas, D. J., Allegrini, F., Bochsler, P., Bzowski, M., Christian, E. R., Crew, G. B., et al. (2009). Global observations of the interstellar interaction from the interstellar boundary explorer (IBEX). *Science* 326, 959–962. doi:10.1126/science.1180906

McComas, D. J., Christian, E. R., Schwadron, N. A., Fox, N., Westlake, J., Allegrini, F., et al. (2018). Interstellar mapping and acceleration probe (IMAP): A new NASA mission. *Space Sci. Rev.* 214 (8). article id 116. doi:10.1007/s11214-018-0550-1

McComas, D. J., Dayeh, M. A., Funsten, H. O., Livadiotis, G., and Schwadron, N. A. (2013). The heliotail revealed by theinterstellar *boundary explorer*. *Astrophysical J.* 771, 77. doi:10.1088/0004-637x/771/2/77

McComas, D. J., and Schwadron, N. A. (2006). An explanation of the Voyager paradox: Particle acceleration at a blunt termination shock. *Geophys. Res. Lett.* 33, 4. doi:10.1029/2005GL025437

Michael, A. T., Opher, M., Toth, G., Tenishev, V., and Drake, J. F. (2021). The impact of kinetic neutrals on the heliotail. *Astrophysical J.* 906, 27.

Michael, A. T., Opher, M., Toth, M., Tenishev, V., and Borovikov, D. (2022). The solar wind with hydrogen ion exchange and large-scale dynamics (shield) code: A self-consistent kinetic–magnetohydrodynamic model of the outer heliosphere. *Astrophysical J.* 924, 105. doi:10.3847/1538-4357/ac35eb

Möbius, E., Liu, K., Funsten, H., Gary, S. P., and Winske, D. (2013). Analytic model of theibexribbon with neutral solar wind based ION PICKUP beyond the heliopause. *Astrophysical J.* 766, 129. doi:10.1088/0004-637x/766/2/129

Mostafavi, P., and Zank, G. P. (2018). The structure of shocks in the very local interstellar medium. *Astrophysical J. Lett.* 854, L15. article id. L15. doi:10.3847/2041-8213/aaab54

Nakanotani, M., Zank, G. P., Adhikari, L., Zhao, L.-L., Giacalone, J., Opher, M., et al. (2021). Downwind solar wind and Pioneer 10 observations. *Astrophysical J.* 901 (2), L23.

Nikoukar, R., Hill, M. E., Brown, L., Krimigis, S. M., Decker, R. B., Roelof, E. C., et al. (2022). On the energy dependence of galactic cosmic ray anisotropies in the very local interstellar medium. *Astrophysical J.* 934, 41. doi:10.3847/1538-4357/ac6fe5

NRC (2015). *Enhancing the effectiveness of team science*. Washington, DC: The National Academies Press. doi:10.17226/19007

NRC (2013). *Solar and space physics: A science for a technological society*. Washington, DC: The National Academies Press. Chapter 4, page 92. doi:10.17226/13060

Opher, M., Bibi, F. A., Toth, G., Richardson, J. D., Izmodenov, V. V., and Gombosi, T. I. (2009). A strong, highly-tilted interstellar magnetic field near the Solar System. *Nature* 462, 1036–1038. doi:10.1038/nature08567

Opher, M. (2012). BU summer school on plasma processes in space physics. Available at: http://people.bu.edu/mopher/PlasmaSummerSchool/Welcome.html.

Opher, M., Drake, J. F., Swisdak, M., Zieger, B., and Toth, G. (2017). The twist of the draped interstellar magnetic field ahead of the heliopause: A magnetic reconnection driven rotational discontinuity. *Astrophysical J. Lett.* 839, L12. doi:10.3847/2041-8213/aa692f

Opher, M., Drake, J. F., Zank, G., Toth, G., Powell, E., Shelley, W., et al. (2021). A turbulent heliosheath driven by the Rayleigh–taylor instability. *Astrophysical J.* 922, 181. doi:10.3847/1538-4357/ac2d2e

Opher, M., Drake, J. F., Zieger, B., and Gombosi, T. I. (2015). Magnetized jets driven by the Sun: The structure of the heliosphere revisited. *Astrophysical J. Lett.* 800, L28. doi:10.1088/2041-8205/800/2/l28

Opher, M., Drake, J., Swisdak, M., Schoeffler, K. M., Richardson, J. D., Decker, R. B., et al. (2011). *Astrophysical J.* 734, 71. doi:10.1088/0004-637x/734/1/71

Opher, M., Liewer, P. C., Gombosi, T. I., Manchester, W. B., DeZeeuw, D. L., Sokolov, I., et al. (2003). Probing the edge of the solar system: Formation of an unstable jet-sheet. *Astrophysical J.* 591, L61–L65. doi:10.1086/376960

Opher, M., Loeb, A., Drake, J., and Gabor, T. (2020). A small and round heliosphere suggested by magnetohydrodynamic modelling of pick-up ions. *Nat. Astron.* 4, 675–683. doi:10.1038/s41550-020-1036-0

Parker, E. N. (1961). The stellar-wind regions. *Astrophysical J.* 134, 20. doi:10.1086/147124

Pesses, M. E., Eichler, D., and Jokipii, J. R. (1981). Cosmic ray drift, shock wave acceleration, and the anomalous component of cosmic rays. *Astrophysical J.* 246, L85. doi:10.1086/183559

Pogorelov, N. V., Borovikov, S. N., Heerikhuisen, J., and Zhang, M. (2015). The heliotail. *Astrophysical J. Lett.* 812 (1), 7. article id. L6. doi:10.1088/2041-8205/812/1/L6

Powell, E., Opher, M., Toth, G., Tenishev, V., Michael, A., Kornbleuth, M., et al. (2021). *AGU fall meeting abstracts, SH15F-2075*.

Richardson, J. D., Cummings, A., Stone, E. C., Burlaga, L., and Opher, M. (2021). Using magnetic flux conservation to determine heliosheath speeds. *Astrophysical J. Lett.* 919, L28. doi:10.3847/2041-8213/ac27b1

Richardson, J. D., Kasper, J. C., Wang, C., Belcher, J. W., and Lazarus, A. J. (2008). Cool heliosheath plasma and deceleration of the upstream solar wind at the termination shock. *Nature* 454 (7200), 63–66. doi:10.1038/nature07024

Roelof, E. (2021). *Voyager SSG*.

Scherer, K., and Ferreira, S. E. S. (2005). A heliospheric hybrid model: Hydrodynamic plasma flow and kinetic cosmic ray transport. *Astrophys. Space Sci. Trans.* 1, 17–27. doi:10.5194/astra-1-17-2005

Schwadron, N. A., and McComas, D. J. (2013). *Astrophysical J. Lett.* 778, L33. doi:10.1088/2041-8205/778/2/l33

Shalchi, A., Bieber, J. W., and Matthaeus, W. H. (2004). Nonlinear guiding center theory of perpendicular diffusion in dynamical turbulence. *Astrophysical J.* 615, 805–812. doi:10.1086/424687

Shrestha, B. L., Zirnstein, E. J., Heerikhuisen, J., and Zank, G. P. (2021). Strength of the termination shock inferred from the global distributed energetic neutral atom flux from IBEX. *The Astrophy. J. Supplement Series* 254 (2), 14.

Smith, C. W., Aggarwal, P., Argall, M. R., Burlaga, L. F., Bzowski, M., Cannon, B. E., et al. (2017). Observations of low-frequency magnetic waves due to newborn interstellar pickup ions using ACE, Ulysses, and Voyager data. *J. Phys. Conf. Ser.* 900 (1), 012018. article id.

Smith, C. W., Isenberg, P. A., Matthaeus, W. H., and Richardson, J. D. (2006). Turbulent heating of the solar wind by newborn interstellar pickup protons. *Astrophysical J.* 638, 508–517. doi:10.1086/498671

Swisdak, M., Drake, J. F., and Opher, M. (2013). *Astrophysical J. Lett.* 774, L8. doi:10.1088/2041-8205/774/1/l8

Toth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., and De Zeeuw, D. L. (2005). Space weather modeling framework: A new tool for the space science community. *J. Geophys. Res. Space Phys.* 110 (12). CiteID A1222.

Toth, G., van der Holst, B., Sokolov, I. V., De Zeeuw, D. L., Gombosi, T. I., Fang, F., et al. (2012). Adaptive numerical algorithms in space weather modeling. *J. Comput. Phys.* 231 (3), 870–903. doi:10.1016/j.jcp.2011.02.006

van der Holst, B., Sokolov, I. V., Meng, X., Jin, M., Manchester IV, W. B., Toth, G., et al. (2014). ALFVÉN WAVE SOLAR MODEL (AWSoM): Coronal heating. *Astrophysical J.* 782, 81. doi:10.1088/0004-637x/782/2/81

Vertesi, J. A. (2020). *Shaping science: Organizations, decisions, and culture on NASA's teams*. University of Chicago Press.

Wang, B., Zank, G. P., Shrestha, B., Kornbleuth, M., and Opher, M. (2023). Relating energetic ion spectra to energetic neutral atoms. *Astrophysical J.* 944, 198. submitted. doi:10.3847/1538-4357/acb437

Wang, C., and Richardson, J. D. (2004). Interplanetary coronal mass ejections observed by Voyager 2 between 1 and 30 AU. *J. Geophys. Res. Space Phys.* 109, A06104. Issue A6, CiteID A06104. doi:10.1029/2004ja010379

Washimi, H., Tanaka, T., and Zank, G. P. (2017). Time-varying heliospheric distance to the heliopause. *Astrophysical J. Lett.* 846, L9. article id. doi:10.3847/2041-8213/aa8556

Washimi, H., Zank, G. P., Hu, Q., Tanaka, T., and Munakata, K. (2015). *Astrophysical J.* 809, 16. doi:10.1088/0004-637x/809/1/16

Williams, L. L., and Zank, G. P. (1994). Effect of magnetic field geometry on the wave signature of the pickup of interstellar neutrals. *J. Geophys. Res.* 99 (10), 19229. doi:10.1029/94ja01657

Zank, G. P., Adhikari, L., Hunana, P., Shiota, D., Bruno, R., and Telloni, D. (2017a). Theory and transport of nearly incompressible magnetohydrodynamic turbulence. *Astrophysical J.* 835, 147. doi:10.3847/1538-4357/835/2/147

Zank, G. P., Du, S., and Hunana, P. (2017b). The origin of compressible magnetic turbulence in the very local interstellar medium. *Astrophysical J.* 842, 114. doi:10.3847/1538-4357/aa7685

Zank, G. P., Heerikhuisen, J., Pogorelov, N. V., Burrows, R., and McComas, D. (2010). *Astrophysical J.* 708, 1092–1106. doi:10.1088/0004-637x/708/2/1092

Zank, G. P., Hunana, P., Mostafavi, P., le Roux, J. A., Li, G., Webb, G. M., et al. (2015). Diffusive shock acceleration and reconnection acceleration processes. *Astrophysical J.* 814–2. doi:10.1088/0004-637X/814/2/137

Zank, G. P. (1999). Interaction of the solar wind with the local interstellar medium: A theoretical perspective. *Space Sci. Rev.* 89 (4), 413–688. doi:10.1023/a:1005155601277

Zank, G. P., Nakanotani, M., and Webb, G. M. (2019). Compressible and incompressible magnetic turbulence observed in the very local interstellar medium by voyager 1. *Astrophysical J.* 887, 116. doi:10.3847/1538-4357/ab528c

Zank, G. P., Nakanotani, M., Zhao, L. L., Du, S., Adhikari, L., Che, H., et al. (2021). Flux ropes, turbulence, and collisionless perpendicular shock waves: High plasma beta case. *Astrophysical J.* 913, 127. doi:10.3847/1538-4357/abf7c8

Zank, G. P., Pauls, D. H., and Williams, L. L. (1996a). Interaction of the solar wind with the local interstellar medium: A multifluid approach. *J. Geophys. Res. (Space)* 101 (21), 639.

Zank, G. P., Pauls, H. L., Cairns, I. H., and Webb, G. M. (1996b). Interstellar pickup ions and quasi-perpendicular shocks: Implications for the termination

shock and interplanetary shocks. *J. Geophys. Res.* 101, 457–477. doi:10.1029/95ja02860

Zank, G. P. (2014). Transport processes in space physics and astrophysics. *Lect. Notes Phys.* 877. doi:10.1007/978-1-4614-8480-6

Zhao, L.-L., Zank, G. P., and Adhikari, L., *Astrophysical J.*, 879:012001, doi:10.1088/1742-6596/1332/1/0120012019).A nearly incompressible turbulence-driven solar wind model,

Zhao, L.-L., Zank, G. P., and Burlaga, L. F. (2020). Turbulence in the very local interstellar medium (VLISM). *Astrophysical J.* 900, 166. doi:10.3847/1538-4357/ababa2, no.

Zieger, B., Opher, M., Toth, G., and Florisnki, V. (2020). Dispersive fast magnetosonic waves and shock-driven compressible turbulence in the inner heliosheath. *J. Geophys. Res.* 125 (10). article id. E28393.

Zieger, B., Opher, M., Toth, G., Decker, R., and Richardson, J. (2015). Constraining the pickup ion abundance and temperature through the multifluid reconstruction of the Voyager 2 termination shock crossing. *J. Geophys. Res.* 120, 7130–7153. doi:10.1002/2015ja021437

Zirnstein, E. J., McComas, D. J., Kumar, R., Elliott, H., Szalay, J., Olkin, C., et al. (2018). *In situ* observations of preferential pickup ion heating at an interplanetary shock. *PRL* 121, 075102. doi:10.1103/physrevlett.121.075102

Zwan, B. J., and Wolf, R. A. (1976). Depletion of solar wind plasma near a planetary boundary. *J. Geophys. Res.* 81, 1636–1648. doi:10.1029/ja081i010p01636

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 19 June 2023
DOI 10.3389/fspas.2023.1130743

Check for updates

OPEN ACCESS

EDITED BY
Konstantinos Dialynas,
Academy of Athens, Greece

REVIEWED BY
Ryuki Hyodo,
Japan Aerospace Exploration Agency (JAXA), Japan
Akos Kereszturi,
Hungarian Academy of Sciences (MTA), Hungary

*CORRESPONDENCE
Zita Martins,
zita.martins@tecnico.ulisboa.pt

†These authors share senior authorship

RECEIVED 23 December 2022
ACCEPTED 09 June 2023
PUBLISHED 19 June 2023

CITATION
Miranda C, Patel M, Berberan-Santos MN, Hormigo T, Correia ACM, Pedras B and Martins Z (2023), The importance of Phobos simulants: a review on our current knowledge.
Front. Astron. Space Sci. 10:1130743.
doi: 10.3389/fspas.2023.1130743

COPYRIGHT
© 2023 Miranda, Patel, Berberan-Santos, Hormigo, Correia, Pedras and Martins. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The importance of Phobos simulants: a review on our current knowledge

Catarina Miranda[1,2], Manish Patel[3], Mário Nuno Berberan-Santos[2,4], Tiago Hormigo[5], Alexandre C. M. Correia[6,7], Bruno Pedras[2,4†] and Zita Martins[1*†]

[1]Centro de Química Estrutural, Departamento de Engenharia Química, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal, [2]Institute for Bioengineering and Biosciences, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal, [3]School of Physical Sciences, The Open University, Milton Keynes, United Kingdom, [4]Associate Laboratory i4HB—Institute for Health and Bioeconomy, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal, [5]Spin.Works, Lisboa, Portugal, [6]CFisUC, Departamento de Física, Universidade de Coimbra, Coimbra, Portugal, [7]IMCCE, Observatoire de Paris, PSL Université, Paris, France

Phobos, a satellite of Mars, was successfully studied by flyby, orbiter, and landing missions to the Red Planet, but several questions remain about its origin, composition, and relationship to Mars. It is suggested that Phobos is either a captured body from the asteroid belt or the outer Solar System (capture scenario), or a consequence of re-accreted ejecta from Mars (*in situ* formation/giant impact). So far, Phobos has been characterized by its two spectral units - blue and red - with different compositional restrains. The red unit represents most of the surface, while the blue unit is focused on the Stickney crater and surroundings. In the absence of samples returned from this satellite, simulant regolith must be studied to infer various proprieties, and complement *in situ* studies. To date, there are three simulants of this satellite: Phobos-1C, Phobos Captured Asteroid-1 (PCA-1), and Phobos Giant Impact-1 (PGI-1). Since Phobos may have a Mars-like composition, terrestrial analogues of Mars should also be analysed. The data retrieved from the various assays performed with these planetary field analogues may be used as a database to complement future space missions to Phobos, but, ultimately, the composition of Phobos will have to be analysed by a sample-return mission.

KEYWORDS

Mars, Phobos, regolith, minerals, space missions

1 Introduction

The formation and evolution of the Solar System began around 4.56 billion years ago, due to the gravitational collapse of a molecular cloud. Rocky planets such as Mercury, Venus, Earth, and Mars formed in distinct stages (Chambers, 2004). Mars appears to have partly similar characteristics to Earth, however, its satellite system comprises two small moons (Phobos and Deimos). Phobos is the biggest Martian Moon, and it orbits closer to the primary body than any other Moon in the Solar System.

Many questions about Phobos remain unanswered, including: *What is the composition and origin of Phobos? What is the relationship between Phobos and Deimos? What is the internal structure of these moons and the mineralogy of the surface? Will the composition of Phobos help to tell the story of the Martian past? Are there biosignatures on the surface of*

Phobos? The origin of Phobos remains unresolved (Craddock, 2011; Rosenblatt, 2011; Canup and Salmon, 2018), and may provide insights into the formation of the terrestrial planets. The above questions ultimately rely on the acquisition of regolith specimens, and state-of-the-art laboratory equipment, with future sample return missions presenting growing interest. The preparation and analysis of specimens in terrestrial laboratories allows the optimization for a particular study, which can sometimes be impossible to do *in situ*. Terrestrial analyses can also be replicated many times, and verified by diverse and independent analytical methods, that can be unaffordable or impossible for remote or robotic techniques (Murchie et al., 2014). In the meantime, and in the absence of samples of regolith returned from Phobos to the terrestrial laboratories, analogous or simulant regolith are fundamental to prepare future sample return missions.

Finally, the study of Phobos may also provide an opportunity for future human exploration of the Mars system (Murchie et al., 2014), as the exploration of Phobos is not only driven by planetary science or the quest to understand its origin. Phobos also holds strategic importance in developing a forward base for human exploration on Mars (Muscatello et al., 2012). With the increasing interest in taking humans to Mars, scientists and engineers are also considering using Phobos as a staging area for a trip to the Red Planet, or even a teleoperation centre for robots on the surface of Mars, to avoid long-latency (Nallapu et al., 2020). Ultimately, Phobos ensures the continuity of the exploration program to Mars.

In this review, the different reasons why it is relevant to study Phobos will be covered. The origin of the Martian moons will be covered in Section 2 while Section 3 describes how the Mars-Phobos system may inform about the geological history of Mars. Section 4 will cover the space exploration of Phobos, including past, present, and future space missions and how these will inform on the characterization of Phobos. Finally, Section 5 will describe the fundamental role and mineral composition of different regolith simulants of Phobos.

2 Origin of the Martian moons

After years of exploring Mars, the origin and evolution of Phobos and Deimos remain unknown. Understanding their origin may provide constraints on the formation and early dynamical environment of Mars (see, e.g., Ramsley and Head, 2021). To explain the origin of the Martian Moon system, several hypothesis have been presented: the recycling hypothesis, the hypothesis that Phobos and Deimos were once a single large Moon that later split, the hypothesis in which the moons are captured small bodies from the asteroid belt or outer Solar System (capture scenario), or the hypothesis in which they are reaccreted impact ejecta from Mars (*in situ* formation or giant impact scenario).

Hesselbrock and Minton (2017) proposed the recycling hypothesis, in which Phobos and Deimos would have accreted from a ring of debris, with the orbit of Phobos slowly decaying and eventually split up to form a new ring of debris around Mars. This new ring would deposit most debris onto Mars, and the leftover ring material would accrete into a new Moon. This recycling hypothesis was challenged by Liang and Hyodo (2023), which showed that particles of rp ~10 cm may be removed by the Poynting–Robertson (PR) effect over 10^9 years, and smaller particles would have been removed faster. Liang and Hyodo (2023) indicated that the PR effect alone would have not been able to remove the particles in the disk, and another process would have been needed. Without that, the hypothesis of Hesselbrock and Minton (2017) would not be accurate.

Bagheri et al. (2021) used geophysical and tidal-evolution modelling of a Mars–satellite system to propose another hypothesis, in which Phobos and Deimos were once a single large Moon that later disintegrated into two. This hypothesis was challenged by Hyodo et al. (2022), who indicated that those authors did not study the impact itself, and therefore the outcome was not demonstrated (i.e., that a single Moon splits only into two fragments), and that the gravitational interactions and collisions during a moon–moon close encounter were neglected.

The capture scenario proposes that the moons were asteroids from the main belt, subsequently captured, under suitable conditions, by the gravitational attraction of Mars. This hypothesis arises from remote-sensing observations of the moons that present similarities with numerous low-albedo and small-sized asteroids (Burns, 1978; Pollack et al., 1979; Rosenblatt, 2011). Phobos presents numerous grooves, which also recall some asteroids, and are claimed to have been formed during the capture by Mars (Pollack and Burns, 1977; Hunten, 1979; Thomas et al., 1979). To note that recently, Cheng et al. (2022) have proposed a tidal fracturing model to explain the numerous grooves, with their analysis supporting a layered heterogeneous structure for Phobos with possible underlying failure-induced fractures. However, significant areas showing absences and anomalies were observed, which were not consistent with the tidal fracturing model. Escaping ejecta imprinted Phobos with linear, low-velocity crater chains (sesquinary catenae) matching the morphology of prominent features (grooves) that do not match a tidal model (Murray and Heggie, 2014; Nayak and Asphaug, 2016; Cheng et al., 2022).

Further support for the capture scenario, the resemblance between the surface spectra of Phobos and D-type carbonaceous material points to the condensation of this material in the solar nebula at a distance far away from Mars (Murchie et al., 1991; Yamamoto et al., 2018), which later moved inwards to the inner Solar System (Rosenblatt, 2011). Nevertheless, this scenario is weakened by some ambiguities in the interpretation of the reflectance spectra and requires high tidal dissipation rates inside the moons to account for their current orbits around Mars (Rosenblatt, 2011). According to Fraeman et al. (2014), in this scenario, a possible composition of the surface implies the presence of primitive phyllosilicate-rich carbonaceous chondrites or anhydrous, olivine- and pyroxene-dominated carbonaceous chondrites.

In recent years, the idea that Phobos may result from a collision between an asteroid and Mars has gained popularity. In this scenario, the collision would have formed a disc of debris around the planet, and it has been proposed that these moons formed from the outer edge of such disk, while massive inner moons spiralled inward and were lost (Ronnet et al., 2016; Rosenblatt et al., 2016; Canup and Salmon, 2018; Changela et al., 2022). It is thought that the Martian surface was heated to 3,000–6,000 K immediately after the massive impact, while the building blocks of Phobos, including volatile-rich vapor, were heated to 2,000 K, depleting the present Moon from volatiles (Hyodo et al., 2017). Like the Earth, Mars has

too much angular momentum to be explained by the accretion of many small bodies (Dones and Tremaine, 1993a). Studies suggest that a planetesimal with 0.02 Mars masses must have collided with Mars early in its history (Dones and Tremaine, 1993b). This collision would have been able to produce the largest basins on Mars (Borealis, Elysium, Utopia, and Hellas) (Craddock, 2011; Canup and Salmon, 2018), and may have been significant enough to impact Mars at a velocity to vaporize rock, ejecting large amounts of material into orbit (Craddock, 2011). Other lines of evidence point to the low mass and high porosity of Phobos that may be explained by a composition of loosely aggregated material from an accretion disk (Craddock, 2011; Rosenblatt, 2011). Due to the inferred presence of phyllosilicates on Phobos, which are probably derived from Mars, the silica composition also supports this hypothesis (Giuranna et al., 2011). It is considered that Mars-derived phyllosilicates are probably significant sources of phyllosilicate intake on the Moon. The physical properties (orbital eccentricity and inclination) of the satellites, which previous scenarios could not explain, are the natural result of accretion from a circum-planetary disk (Craddock, 2011; Ronnet et al., 2016). In this rationale, disk material would be heated to dehydrate OH-bearing minerals (Lange and Ahrens, 1982), possibly indicating a dry endogenic composition. If a significant impact occurred, the source materials would be mostly Martian, and the refractory elemental composition of Phobos would be Mars-like (Fraeman et al., 2014). Notwithstanding, since the composition of the debris may also come from the impactor, which could be of carbonaceous composition, the *in situ* formation hypothesis also supports the carbonaceous composition of Phobos (Rosenblatt, 2011). As also discussed by Pignatale et al. (2018), and Ronnet et al. (2016), the spectral properties of Phobos are in accordance with an impact-generated accretion disk formation. A more conclusive answer to the origin of Phobos will come from more *in situ* measurements or sample return missions, in which state-of-the-art technology will help constrain the various scenarios (see Section 4.3).

3 Martian past

As one of the major terrestrial planets in our the Solar System, Mars contributes to our understanding of the evolution of habitable planets. Throughout its history, Mars has experienced numerous asteroidal impacts from various locations, producing ejecta from its surface. A fraction of this debris is delivered to Earth as meteorites, but another fraction is delivered to Phobos and Deimos and mixed within the surface regolith. Hyodo et al. (2019) reported that the materials delivered to Phobos are physically and chemically different from the Martian meteorites retrieved on Earth, as they are less shocked (<5 GPa) and more fragile. Impacts on the surface of Mars were frequent, and it was already suggested that the regolith of Phobos might hold up to 250 ppm of this material (Ramsley and Head, 2013). As Phobos has been receiving this ejecta representative of the surface of Mars over geological time, Phobos may hold remarking information of the entire history of the planet that would have been altered by the changing geological processes on Mars. If life ever existed on Mars, biosignatures could be carried to Phobos. This could be investigated in the future by the Mars Sample Return (MSR) mission led by NASA-ESA, which will return Mars material from the Jezero crater, and the Japan Aerospace Exploration Agency (JAXA)-led Martian Moons eXploration (MMX) mission, which will collect material from the Martian Moon Phobos (Hyodo and Usui, 2021).

Besides asteroidal impacts in the Mars-Phobos system, atoms and molecules that escape the planetary atmosphere can also be embedded and preserved inside the surface of the satellite (Nénon et al., 2021). Although the age of the surface of Phobos may be uncertain (Schmedemann et al., 2014; Ramsley and Head, 2017; Hu et al., 2020), it is believed that atmospheric ions may have bombarded some regolith during the pre-Noachian epoch of Mars, where water is thought to have existed (Fassett and Head, 2008; Warner et al., 2010). Phobos currently shows its nearside to Mars, but it is likely that the Stickney Crater impact desynchronized the rotation of Phobos (Ramsley and Head, 2017). After the pre-impact tidal lock was restored, Phobos either returned to the original pre-impact longitude or was reoriented into a tidal lock ~180°. Therefore, atmospheric flux embedded on the top of Phobos regolith may reflect the different periods of time of the geological history after the Stickney Crater impact. Due to its orbit and based on *in situ* ion observations, Phobos presents a space-weathering dichotomy between the nearside and farside (Nénon et al., 2021). The bombardment by Martian ions can lead to the transformation of an initially crystalline material at a given depth into an amorphous material, creating amorphous rims on regolith grains (Nénon et al., 2021). Sampling the Phobos regolith would provide information about the atmospheric composition of Mars, just like Martian meteorites have (Bogard and Garrison, 1998; Filiberto and Schwenzer, 2019).

4 Observing and exploring Phobos

Phobos was discovered by the astronomer Asaph Hall in 1877, with the collaboration of his wife and mentor, Angeline Stickney Hall (Hall, 1877). However, the first mission to observe this Moon from a closer distance was almost 100 years later. Since the first photograph, taken in 1969 (Masursky et al., 1972), various missions, including failed ones, attempted to study Phobos. Although the majority was aimed for Mars, and there are still no samples from the regolith, several questions have been answered over the past 50 years. Flyby, orbiter, and lander missions that studied Mars were also able to characterize Phobos successfully (Figure 1). In recent years, due to the possibility of having humans on Mars in the 2040s (Deutsch et al., 2018), Phobos has been viewed as a potential staging area for future human exploration of Mars. In this section, a brief history of the exploration of Phobos, including past and present missions, along with the main findings essential for the characterization of Phobos, will be reported. The future of the research concerning Phobos will also be discussed.

4.1 Past and present space missions

To address the distinct hypotheses and other aspects of its history, Phobos has been studied via spacecraft-based thermal, imaging, spectroscopy, and gravity measurements. Most missions that observed Phobos were flyby, orbiter, or lander/rover destined to

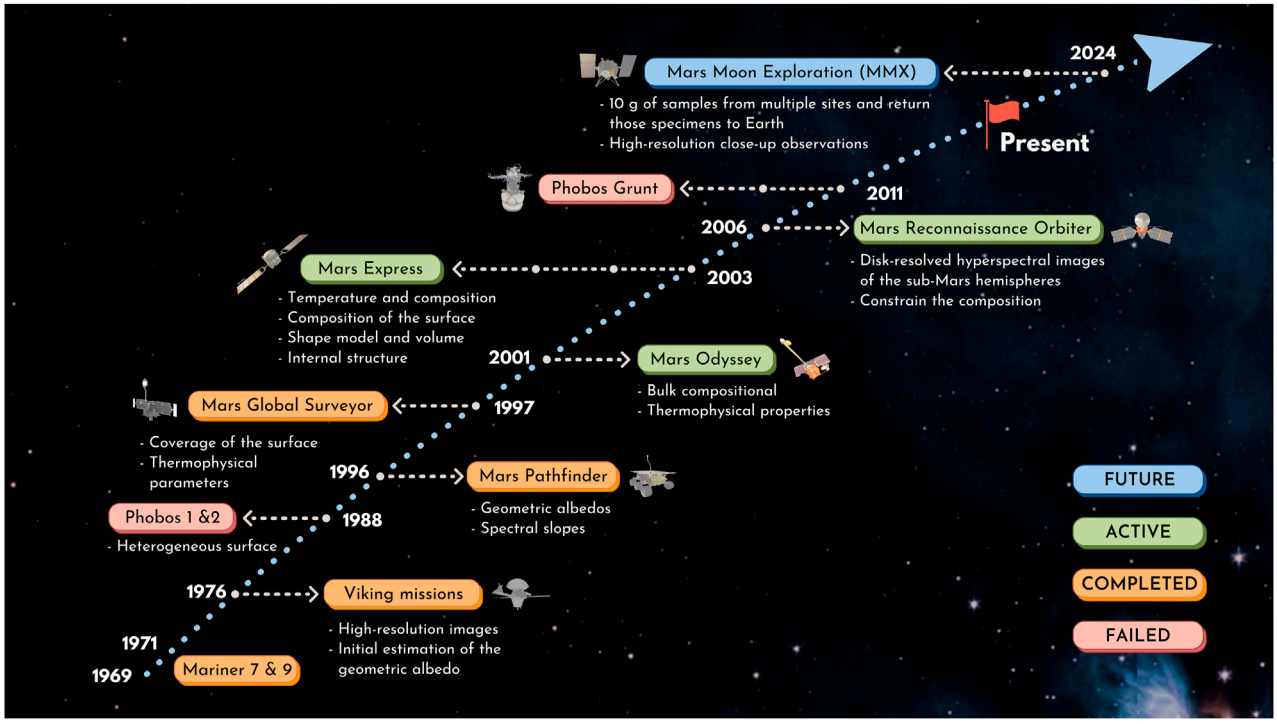


FIGURE 1
Timeline of the exploration of Phobos with corresponding achievements. Failed missions are represented in red, completed missions in orange, still active missions in green, and future mission in blue. ©Image by Catarina Miranda.

study Mars (Siddiqi, 2018 and Figure 1). Nevertheless, as the interest in exploring Phobos grew, missions were explicitly developed for investigating this satellite (e.g., Marov et al., 2004; Nakamura et al., 2021).

Phobos was first photographed by some early probes to Mars as flyby targets of opportunity, beginning with the Mariner 7 in 1969, followed by Mariner 9 spacecraft, launched in 1971 (Masursky et al., 1972). Although the observation of Phobos was not initially scheduled, the final proposed orbits for the Mariner 9 mission were sufficiently close to use the high-resolution B-camera (Pollack et al., 1972), infrared (IR) radiometer (Gatley et al., 1974), and ultraviolet (UV) spectrometer onboard (Chase and Hat, 1972). A total of 214 images of Phobos and Deimos, some as close as 1,200 km were recovered (Duxbury et al., 2014). The subsequent missions to explore Phobos were the Viking missions, launched in 1975, which placed two orbiters around Mars and two landers on the Martian surface in 1976. During the Viking primary mission, from June to November 1976, the cameras on the two orbiters took high-resolution images (<500 m) of Phobos, completing its surface coverage and leading to the discovery of a few unusual surface features, including a global network of grooves (Duxbury and Veverka, 1977; Veverka and Duxbury, 1977). In the extended mission phase, close flybys within ~100 km were performed. The infrared thermal mappers (IRTMs) onboard the orbiters carried extensive thermal infrared observations of the Martian satellite at various ranges, observing geometries, and wavelengths, allowing the deduction of the near-surface thermophysical properties (Lunine et al., 1982). A sequence of pictures taken from the narrow-band blue and red diodes, and one of the three infrared imaging (in this case IR2) diodes onboard the Viking lander also enabled an initial estimation of the geometric albedo of Phobos (Pollack et al., 1978). Named after their target, Phobos 1 and Phobos 2 missions, consisting of an orbiter and lander spacecraft, were sent in 1988 and lost prematurely, with Phobos 1 lost on the way to Mars. Nevertheless, before losing contact nearly an hour before it was due for a touch-down, Phobos 2 observed parts of the surface with three multispectral sensors, covering the wavelength range 0.33–3.16 μm (Murchie and Erard, 1996). Thirty-seven images during three (700 km) encounters were retrieved, showing the presence of more grooves (Avanesov et al., 1989) and a heterogeneous surface (Murchie and Erard, 1996), consisting of two spectral units (see Section 3.3 Characterization of Phobos).

Almost 10 years later, on the surface of Mars, the 1996 Mars Pathfinder Rover was also able to make multispectral measurements of Phobos, using the Imager to Mars Pathfinder (IMP) onboard, showing consistent results with previous data (Murchie et al., 1999; Thomas et al., 1999). The satellite was observed on two occasions (Murchie et al., 1999), covering the wavelength range from 440 nm to 1 μm. The spectra retrieved were used to derive the geometric albedos and spectral slopes of the objects (Thomas et al., 1999). The IMP data also contributed to tantalizing but inconclusive evidence for the presence of some type of Fe mineral having an absorption near 700 nm (Murchie et al., 1999).

Over 350 resolved observations of Phobos were performed by the 1997 Mars Global Surveyor (MGS). As MGS went into the orbit of Mars, close flybys of Phobos were targeted. Data

retrieved from the scientific remote sensing payload: Thermal Emission Spectrometer (TES), Mars Orbiter Camera (MOC), and Mars Orbiter Laser Altimeter (MOLA), provided coverage of the surface of Phobos (Duxbury et al., 2014), allowing insights into the variations of thermophysical parameters across the surface (Smith et al., 2018). In the following year, the first Japanese planetary spacecraft (Ichiro Kawaguchi, 1999) named NOZOMI was launched. This mission was expected to take pictures of Mars and its moons. However, the spacecraft did not reach Mars, and therefore did not return any data about Mars and/or its moons, only providing valuable data about the deep space environment (Siddiqi, 2002). In October 2001, the 2001 Mars Odyssey spacecraft performed a low, circular orbit around Mars and is still operating Clique ou toque aqui para introduzir texto. Although it was not possible to observe Phobos right away, the Thermal Emission Imaging System (THEMIS) was able to study the satellite in the following months. THEMIS enabled the characterization and mapping of the bulk compositional and thermophysical properties of Phobos. This instrument collected images in the thermal infrared (TIR) and visible (VIS) wavelengths, observing the shadow of Phobos in February 2012 (Duxbury et al., 2014), and the full Moon near the closest approach on 19 September 2017 (Bandfield et al., 2018). The retrieved TIR images revealed new spectral and thermophysical properties, possibly indicating significant compositional heterogeneity (see Section 3.3 Characterization of Phobos) (Bandfield et al., 2018).

The Mars Express (MEx) mission has been orbiting Mars since 24 December 2003, approaching Phobos at distances less than1000 km, at about every 500 orbits (5 months) (Witasse et al., 2014). MEx has an orbit that extends beyond Phobos, providing data for the side that is not facing Mars (Duxbury et al., 2014). The instrumentation onboard comprises the High-Resolution Stereo Camera (HRSC), imaging spectrometer Observatoire pour la Minéralogie, l'Eau, les Glaces et l'Activité (OMEGA), Planetary Fourier Spectrometer (PFS), UV-IR spectrometer Spectroscopy for the Investigation of the Characteristics of the Atmosphere of Mars (SPICAM), radar instrument Mars Advanced Radar for Subsurface and Ionosphere Sounding (MARSIS), plasma package Analyzer of Space Plasmas and EneRgetic Atoms (ASPERA-3), and the Mars Radio Science experiment (MaRS) (Witasse et al., 2014). These instruments were designed to study Mars but are very well suited for the observation of Phobos. In fact, MEx mapped 90% of the surface, allowing images in 3-D and colour to be taken, including high-resolution images of the far side of Phobos, adding substantially to the science of Phobos. Other achievements include the study of the temperature and composition of the surface with the IR and UV-VIS-IR spectrometers, respectively, the investigation of the internal structure of the satellite and detection of backscattered solar wind protons from the surface of Phobos, enlightening the regolith-solar wind interaction, and the refinement of the ephemerides (Witasse et al., 2014). Data retrieved also improved the shape model and volume of Phobos, constrained the composition of the surface, allowed new geologic interpretation for the grooves, and contributed to the still debatable dating of the surface age of Phobos (Witasse et al., 2014). Several authors have attempted to age-dating Phobos by counting craters (see, e.g., Schmedemann et al., 2014). However, the overprinting of secondary craters inside the Stickney crater does not allow a feasible crater age-dating (Ramsley and Head, 2017). This gives an indication of a history of several impacts on Phobos, but it does not allow to obtain information on the present-day surface age.

FIGURE 2
VIS/NIR colour composite images of Phobos from HiRISE instrument onboard MRO. The biggest crater, Stickney is on the left edge of the image. The image is in the HiRISE catalog as PSP_007769_9010_IR. Credit NASA/JPL-Caltech/UArizona; https://www.uahirise.org/releases/phobos/.

Another still operating spacecraft is the Mars Reconnaissance Orbiter (MRO), which went into its low, near-circular orbit in August 2006. For calibration purposes, Phobos and Deimos were observed as point sources against star fields by the Navigation Camera during the approach (Thomas et al., 2011). The hyperspectral infrared Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) (Murchie et al., 2007; Fraeman et al., 2012; Pajola et al., 2018) and the High-Resolution Imaging Science Experiment (HiRISE) (Thomas et al., 2011) onboard were able to image Phobos (see Figure 2). These instruments are expected to add new observations of the satellite periodically. CRISM has two high spatial resolution spectrometers with a spectral resolution of 6.55 nm/channel, whereby one channel is in the VIS/NIR region (from 0.4 to 1.0 mm), and the other in the IR region (1.0–3.9 mm) (Murchie et al., 2007). This instrument has acquired disk-resolved hyperspectral images of the sub-Mars hemispheres of Phobos, and along HiRISE, helped to constrain the knowledge on the composition of Phobos (Murchie et al., 2007; Thomas et al., 2011; Pajola et al., 2018).

The last mission sent with the ability to study in detail Phobos was the follow-up to Phobos 2 mission, named Phobos-Grunt. This spacecraft was launched on 9 November 2011 and intended to return up to 200 g of regolith to Earth (Galimov, 2010; Murchie et al., 2014). However, soon after launch, the spacecraft failed, and the probe re-entered the atmosphere of the Earth in January 2012 (Murchie et al., 2014). Presently, the investigation of Phobos relies

on the still operating orbiter missions (Mars Odyssey, MEx, and MRO) and surface observations during the solar transits by the Mars Surface Laboratory Curiosity, and Interior Exploration using Seismic Investigations, Geodesy and Heat Transport (InSight) mission (Duxbury et al., 2014; Stähler et al., 2020).

4.2 Future exploration of Phobos

Various missions already had the opportunity to observe Phobos, while others remain at the mission concept stage. While missions that remain at the concept stage were never launched, information is still gained: a team of scientists and engineers develop their concept around a question and idea, and their proposal is evaluated for scientific merit and feasibility, including among others, power consumption, mass, operations to be carried out and instruments to be used. At this point, the concept of the proposal is fine-tuned, so even when a space mission concept is not selected and approved, the information that is gained in this process will inform efforts in other future space projects, in terms of science, payload and engineering, therefore reducing risk and guarantying a cost-effective proposal. The proposed Phobos mission concepts can be divided into three different goals: sample return, imaging and spectroscopy, and high-resolution imaging. Sample return missions include the low-cost Discovery mission Aladdin (Mueller et al., 2003), Martian Moon Sample Return (MMSR) (Michel et al., 2011), PHOOTPRINT (Galimov, 2010), and HALL (Lee et al., 2010) missions. On the other hand, Mars Phobos and Deimos Survey (M-PADS) (Ball et al., 2009), Phobos Close Rendevous Observation Sensing Satellite (PCROSS) (Bellerose and Andrews, 2012), Phobos Surveyor (Lang et al., 2012), PANDORA (Raymond et al., 2015), and CubeSat missions Low Orbit Geology Imaging CubeSat (LOGIC) (Thangavelautham et al., 2017), and Perseus (Nallapu et al., 2020) were proposed as imaging and spectroscopy missions, and NASA Discovery mission Phobos And Deimos and Mars Environment (PADME) (Lee et al., 2014) as a high-resolution mission.

Currently, the only approved mission targeting Phobos is the JAXA mission named MMX (see Figure 1), which is planned for launch in 2024 (Ulamec et al., 2019). MMX focuses on revealing the origin of the Martian moons, and making progress in the study of planetary system formation and primordial material transport around the border between the inner- and outer parts of the early Solar System. In addition, MMX aims to understand the physical processes in the circum-Martian environment and atmosphere of Mars, improving the knowledge about the evolution of the Martian moons and the surface environmental transition of the Red Planet (Kuramoto et al., 2018). As such, the main spacecraft will make high-resolution close-up observations of both moons from orbit, and a small rover will be delivered to the surface of Phobos. More than 10 g of samples from multiple sites will be collected by the main lander, and the MMX spacecraft will later return those specimens to Earth. Once in the laboratory, scientists will analyse the regolith, measure the mineralogical composition of the surface, and determine the thermal properties of the surface material (Kuramoto et al., 2018). The mineralogical composition of the specimens will be scrutinized with chronological analysis, unravelling the origin and cosmochemical nature of Phobos in the context of planet formation and the evolution of the Martian system. The operation period will be shorter than 2.5 years (Hyodo et al., 2019), where rover operations are foreseen for at least 100 days in 2026. This mission is supposed to return in 2028, arriving on Earth in 2029 (Kuramoto et al., 2018).

MMX will be crucial to finally determine the origin of Phobos, combining close-up spectroscopic observations of the mineralogical and chemical composition of bedrock material of Phobos, with sample analyses, such as petrological, chemical and isotopic characterization (Kuramoto et al., 2022). A capture origin will be supported if the Phobos bedrock material is enriched in hydrous minerals and organic matter, therefore being similar to carbonaceous meteorites. The capture scenario is further supported if a sign of internal ice is detected (either through water molecules emerging from the surface of Phobos and/or internal density heterogeneity). Alternatively, if the Phobos samples are volatile-poor, and igneous characteristics are found, then a giant impact is supported (Kuramoto et al., 2022).

4.3 Characterization of Phobos

The Martian Moon Phobos is an irregularly shaped and low-density body (Pätzold et al., 2014), whose surface can be characterized as an arena for various geologic processes. The significant landforms detectable on the surface are craters and grooves (Basilevsky et al., 2014). From available spectroscopic data, Phobos features two distinct units (e.g., (Thomas et al., 1999); (Murchie and Erard, 1996); (Thomas et al., 2011): the "blue unit", which follows Stickney crater and continues eastward for about 5–7 km (proximal ejecta), and the "red unit", corresponding to the remaining surface (Pieters et al., 2014). UV-VIS albedo patterns, the relationship between color and albedo, thermal IR spectra, and the UV continuum slope spectral properties can distinguish both units, possibly constraining the mineralogy of the surface materials. VIS to NIR spectra acquired are commonly used to establish, or at least confine, the mineral composition of Solar System bodies through the identification of highly diagnostic mineral absorption bands (Pieters et al., 2014).

As a result of its proximity with Mars and slightly eccentric orbit, Phobos experiences downslope movements, altering the global distribution of dynamic slopes (Crites et al., 2018). This mass movements play an important role in surface modification by exposing fresh sub-surface material. In fact, there is a correlation between mass wasting characteristics and regions where the dynamic slope increased because of the tidal action (Shi et al., 2016). On the "blue unit", on the eastern ridge of Stickney (Shingareva and Kuzmin, 2001; Shi et al., 2016; Crites et al., 2018).

Thermal infrared measurements performed by the PFS/MEx (Giuranna et al., 2011), TES/MGS (Giuranna et al., 2011; Ebel and Brook, 2015), THEMIS/Mars Odyssey (Bandfield et al., 2018) and CRISM/MRO (Fraeman et al., 2014) indicated the presence of phyllosilicates (which are the primary phase in carbonaceous chondrite meteorites) on parts of Phobos. Although it was initially interpreted to be a signature associated with the "blue unit" (Giuranna et al., 2011), the visual correlation of the footprints of the

spectra of phyllosilicates with CRISM colour variations indicates a better match with the "red unit" (Fraeman et al., 2014). According to (Fraeman et al., 2014), the "blue unit" is spectrally featureless in the thermal infrared, as it is in the near-infrared. Spectral analyses from PFS and TES also point to the presence of feldspars/feldspathoids. Notwithstanding, there are differences compared to the spectra observed in the laboratory, and classes of chondrites are excluded, as they do not present a good agreement with the spectral features observed (Giuranna et al., 2011). The THEMIS continuum-removed spectra hit for the presence of material with identical SiO_2 content to pyroxene minerals or basaltic rocks, as opposed to more silicic or ultramafic compositions (Bandfield et al., 2018).

Two focal absorption features in the VIS/NIR region could be differentiated at 0.65 μm and 2.8 μm. The broad absorption centered at 0.5 μm was highlighted by CRISM but reexamined observations from IMP/Pathfinder (Murchie et al., 1999), Phobos-2 (Murchie and Erard, 1996), and OMEGA/MEx (Fraeman et al., 2012) are also in agreement. Cronstedtite (the most common phyllosilicate in CM chondrites) and nontronite (Fe-bearing phyllosilicate) were intended to match this feature but proved inadequate. Moreover, Fe^{2+} electronic absorptions diagnostic of olivine and pyroxene were not detected. However, a mixture of iron particles ranging in size from a few nanometres to a few micrometres provided an alternative spectral match (Fraeman et al., 2014). Without the thermally emitted component, a shallower metal-OH absorption band in the "blue unit" compared to the "red unit" was detected at 2.8 μm. The strength, position, and shape of both features are similar to characteristics seen on red-sloped, low-albedo primitive asteroids. Therefore, evidence that desiccated phyllosilicates may be present on the surface of Phobos. As an alternative to explaining the presence of highly desiccated Fe-phyllosilicate minerals, the idea that space weathering-related processes may create Fe and OH on the surface was also suggested to justify the features observed by CRISM (Fraeman et al., 2014).

The studies already performed helped constrain the possible mineral composition of Phobos, pointing to the plausible presence of phyllosilicates, specifically desiccated Fe-phyllosilicates (e.g., Nontronite), feldspars/feldspathoids, pyroxene minerals, or basaltic rocks. The future investigation should depend on samples from the regolith and comparison with the already mentioned minerals.

5 Simulant regolith and mineralogical composition

In the absence of specimens from the regolith of Phobos here on Earth, analogous or simulant regolith are an useful solution for inferring properties and complement *in situ* studies in the future. Composition, size, shape distribution, and packing density are the four major characteristics needed to specify a simulant (Rickman et al., 2013). Since there are no specimens from Phobos, none of the characteristics are explicitly known. As such, specifications for a Phobos simulant must be based on spectroscopy and thermophysical properties or other measurements and informed judgment. So far, three Phobos regolith simulants have been developed: Phobos Captured Asteroid-1 (PCA-1), Phobos Giant Impact-1 (PGI-1) (Landsman et al., 2021), and Phobos-1C (Rickman et al., 2016). It is still not clear which hypothesis, i.e., either a capture scenario or *in situ* formation/giant impact explains the origin of Phobos. Furthermore, in this last scenario the composition of Phobos would result from reaccreted impact ejecta from Mars, therefore it is important to study both Phobos and Mars analogues (for example, JSC Mars-1). There are still various questions regarding the Mars-Phobos system, and by studying both regoliths, one can complement this intriguing riddle. The composition of Phobos-1C, PCA-1, PGI-1, and JSC Mars-1 analogues are described in Table 1.

Phobos-1C is a compositional simulant, for which the Tagish Lake meteorite was taken as an estimation for the target composition, customized for the effects of space weathering at the surface. This simulant is composed of the lunar soil simulant JSC-1A (Ray et al., 2010), pseudo-agglutinate, antigorite, and gilsonite. The first two components provide glass with microcrystals and mineral grains, while antigorite was used to represent phyllosilicates found in the meteorite and gilsonite to mimic the complex organic constituents (Rickman et al., 2016). PCA-1 and PGI-1 are also compositional simulants based on the inferred mineralogy of Phobos (Landsman et al., 2021). These regoliths manifest both hypotheses for the formation of Phobos. PCA-1 is based on the scenario that Phobos is a captured primitive asteroid, where the primitive carbonaceous CI meteorite Orgueil (Metzger et al., 2019) was used as a reference material. PGI-1 was developed under the assumption that Phobos is probably a mixture of the upper mantle of Mars and a compositionally primitive impactor. In fact, numerical simulations suggest that the Martian crust could have also been mixed because of a giant impact (Hyodo et al., 2017; Pignatale et al., 2018). As such, the composition of PGI-1 is 43% simplified Mars upper mantle material (olivine and pyroxene), and 57% CI simulant (Landsman et al., 2021). Both simulants have already been studied, specifically their geotechnical properties, and tested in the context of spacecraft hardware and mission operations for the upcoming MMX mission. In the latter studies, the regoliths help constraining the designs regarding the sampling and landing mechanisms (D'Amore et al., 2019; Miyamoto et al., 2018). Using the same recipe as PCA-1 and PGI-1, but sourced in Japan, Miyamoto et al. (2018) developed the "University of Tokyo Phobos Simulant-Impact Based" (UTPS-IB) and the "University of Tokyo Phobos Simulant-Tagish Based" (UTPS-TB).

JSC Mars-1 is a Hawaiian altered volcanic soil consisting of a mixture of ash particles (Allen et al., 1998a). This regolith has been used in various Mars-oriented research studies (Sternovsky et al., 2002; Kral et al., 2004; Garry et al., 2006). Fortunately, this analogue has already been well characterized by Allen et al. (1998a) and Moroz et al. (2009). According to these authors, a relatively featureless ferric absorption edge can be detected in the visible. There is also an indication of a ferric absorption band in the 800–1,000 nm region and a relatively flat absorption in the near-IR spectrum. X-Ray diffraction (XRD) spectra of JSC Mars-1 are dominated by peaks corresponding to Ca-feldspar and minor magnetite (Ti-magnetite), but olivine and pyroxene could also be evidenced. Nevertheless, the spectra from the JSC Mars-1 do not present evidence for a composition rich in phyllosilicates (Allen et al., 1998).

TABLE 1 Mineral composition of the different simulant of Phobos (Phobos Captured Asteroid-1 (PCA-1), Phobos Giant Impact-1 (PGI-1), and Phobos-1C), and Mars (JSC Mars-1). JSC-1A (+) is a component of Phobos-1C.

Simulants	Mineral composition	wt%	References
Phobos-1C	JSC-1A (+)	46	Rickman et al. (2016)
	Antigorite	35	
	Gilsonite	4	
PCA-1/PGI - 1	Antigorite	62.0/35.3	Landsman et al. (2021)
	Attapulgite	Not defined	
	Epsomite	6.0/3.4	
	Olivine	7.0/35.3	
	Pyrite	6.5/3.7	
	Vermiculite	Not defined	
	Organics (sub bituminous coal)	5.0/2.9	
	Pyroxene	–/17.2	
	Magnetite	13.5/7.7	
JSC Mars-1	Ca-feldspar	Not defined	Allen et al. (1998b)
	Ti- Magnetite		
	Olivine		
	Pyroxene		
JSC-1A (+)	Plagioclase	37.1	Ray et al. (2010)
	Olivine	9.0	
	Cr-spinel	1.1	
	Ti-magnetite	0.4	
	K-silicate	1.4	
	Sulfide	1.0	
	Albite	0.3	
	Quartz	0.2	
	Chlorite	0.1	
	Apatite	<0.1	
	Clinopyroxene	<0.1	

6 Conclusion

After more than 50 years of studying Phobos, this Moon remains a mystery. Intriguing questions regarding the origin, composition and relation with the history of Mars are still unanswered. From the two main hypothesis that prevail, capture scenario or giant impact, the latter gained popularity in recent years. As such, in the absence of regolith from this satellite, and taking into consideration that the composition of Phobos may be similar to Mars, it is also important to consider Martian analogues in different regolith studies. Analogue regoliths from Phobos have already been developed and are key elements to complement *in situ* studies in the future. Phobos simulants may be used to support future space missions to test protocols and payload instrumentation, as well as to establish and better understand Phobos properties. The Phobos analogues PCA-1, PGI-1, and Phobos-1C, along the Martian analogues (i.e., JSC-1) may be crucial to infer properties for future space missions, such as the MMX. The JAXA mission will study Phobos, and hopefully will return regolith samples back to Earth by 2029. These samples will later be studied by state-of-the-art technology, and will finally provide detailed information on the physical and chemical properties of the Phobos regolith, and its origin.

Author contributions

AC, ZM, TH, and MB-S contributed to conception and design of the study. CM wrote the first draft of the manuscript,

with contribution by BP and ZM. All authors contributed to the article and approved the submitted version.

Funding

CM was financed by Fundo Europeu de Desenvolvimento Regional (FEDER) funds through the COMPETE 2020 - Operational Programme for Competitiveness and Internationalisation (POCI), and by Portuguese funds through FCT—Fundação para a Ciência e a Tecnologia in the framework of the project POCI-01-0145-FEDER-029932 (PTDC/FIS-AST/29932/2017). Centro de Química Estrutural (CQE) acknowledges the financial support of FCT—Fundação para a Ciência e Tecnologia through grants UIDB/00100/2020 and UIDP/00100/2020. Institute of Molecular Sciences acknowledges the financial support of FCT—Fundação para a Ciência e Tecnologia through grant LA/P/0056/2020. Institute for Bioengineering and Biosciences (iBB) acknowledges the financial support of FCT—Fundação para a Ciência e Tecnologia through grants UIDB/04565/2020 and UIDP/04565/2020. The Associate Laboratory Institute for Health and Bioeconomy (i4HB) acknowledges the financial support of FCT—Fundação para a Ciência e Tecnologia through grant LA/P/0140/2020. Centro de Física da Universidade de Coimbra (CFisUC) acknowledges the financial support of FCT—Fundação para a Ciência e Tecnologia through grants UIDB/04564/2020 and UIDP/04564/2020. The Open University acknowledges funding from the UK Space Agency through grants ST/V005332/1 and ST/V002295/1.

Acknowledgments

We are grateful to three referees for helpful comments.

Conflict of interest

TH was employed by Spin.Works.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Allen, C. C., Jager, K. M., Morris, R. v., Lindstrom, D. J., Lindtsrom, M. M., and Lockwood, J. P. (1998a). Martian soil stimulant available for scientific, educational study. *Eos, Trans. Am. Geophys. Union* 79 (34), 405–409. doi:10.1029/98EO00309

Allen, C. C., Morris, R. V., Jager, K. M., Golden, D. C., Lindstrom, D. J., Lindstrom, M. M., et al. (1998b). "Martian regolith simulant JSC MARS-1," in 29th Annual Lunar and Planetary Science Conference, Houston, TX,, March 16-20, 1998.

Avanesov, G. A., Bonev, B. I., Kempe, F., Bazilevsky, A. T., Boycheva, V., Chikov, K. N., et al. (1989). Television observations of Phobos. *Nature* 341 (6243), 585–587. doi:10.1038/341585a0

Bagheri, A., Khan, A., Efroimsky, M., Kruglyakov, M., and Giardini, D. (2021). Dynamical evidence for Phobos and Deimos as remnants of a disrupted common progenitor. *Nat. Astron* 5, 539–543. doi:10.1038/s41550-021-01306-2

Ball, A. J., Price, M. E., Walker, R. J., Dando, G. C., Wells, N. S., and Zarnecki, J. C. (2009). Mars Phobos and Deimos Survey (M-PADS) - a martian moons orbiter and Phobos lander. *Adv. Space Res.* 43 (1), 120–127. doi:10.1016/j.asr.2008.04.007

Bandfield, J. L., Piqueux, S., Glotch, T. D., Shirley, K. A., Duxbury, T. C., and Edwards, C. S. (2018). "Mars Odyssey themis observations of Phobos: New spectral and thermophysical measurements," in 49th Lunar and Planetary Science Conference, Woodlands, TX, March 19-23, 2018, 7–8.

Basilevsky, A. T., Lorenz, C. A., Shingareva, T. V., Head, J. W., Ramsley, K. R., and Zubarev, A. E. (2014). The surface geology and geomorphology of Phobos. *Planet. Space Sci.* 102 (C), 95–118. doi:10.1016/J.PSS.2014.04.013

Bellerose, J., and Andrews, D. (2012). Phobos close rendevous observation sensing satellite A introduction: PCROSS is a robotic pre-cursor mission. *Concepts Approaches MarsExploration* 1679, 4180.

Bogard, D. D., and Garrison, D. H. (1998). Relative abundances of argon, krypton, and xenon in the Martian atmosphere as measured in Martian meteorites. *Geochimica Cosmochimica Acta* 62 (10), 1829–1835. doi:10.1016/S0016-7037(98)00116-1

Burns, J. A. (1978). The dynamical evolution and origin of the Martian moons. *Vistas Astronomy* 22 (2), 193–210. doi:10.1016/0083-6656(78)90015-6

Canup, R., and Salmon, J. (2018). Origin of Phobos and Deimos by the impact of a vesta-to-ceres sized body with Mars. *Sci. Adv.* 4 (4), eaar6887. doi:10.1126/sciadv.aar6887

Chambers, J. E. (2004). Planetary accretion in the inner solar system. *Earth Planet. Sci. Lett.* 223 (3–4), 241–252. doi:10.1016/J.EPSL.2004.04.031

Changela, H. G., Chatzitheodoridis, E., Antunes, A., Beaty, D., Bouw, K., Bridges, J. C., et al. (2022). Mars: New insights and unresolved questions. *Int. J. Astrobiol.* 20 (6), 394–426. doi:10.1017/S1473550421000276S

Chase, J. R., Hat, H., Stewart, A. I., and Lane, A. L. (1972). Mariner 9 ultraviolet spectrometer experiment: Initial results. *Science* 175 (4019), 309–312. doi:10.1126/science.175.4019.309

Cheng, B., Asphaug, E., Ballouz, R.-L., Yu, Y., and Baoyin, H. (2022). Numerical simulations of drainage grooves in response to extensional fracturing: Testing the Phobos groove formation model. *Planet. Sci. J.* 3 (11). doi:10.3847/PSJ/ac8c33

Craddock, R. A. (2011). Are Phobos and Deimos the result of a giant impact? *Icarus* 211 (2), 1150–1161. doi:10.1016/j.icarus.2010.10.023

Crites, S., Kawakatsu, Y., Fujimoto, M., Ballouz, R.-L., and Baresi, N. (2018). The shifting sands of Phobos: Space weathering and mass wasting on Mars' moon. *Cosp* 42, B1.1-62–18. doi:10.1016/J.PSS.2014.02.008KALUNA

D'Amore, M., Maturilli, A., Miyamoto, H., Niihara, T., Grott, M., Knollenberg, J., et al. (2019). "Phobos regolith simulant for MMX mission: Spectral measurement for remote target identification and deconvolution system training," in 50th Lunar and Planetary Science Conference, Woodlands, TX, March 18-22, 2019.

Deutsch, A. N., Head, J. W., Ramsley, K. R., Pieters, C. M., Potter, R. W. K., Palumbo, A. M., et al. (2018). Science exploration architecture for Phobos and Deimos: The role of Phobos and Deimos in the future exploration of Mars. *Adv. Space Res.* 62 (8), 2174–2186. doi:10.1016/j.asr.2017.12.017

Dones, L., and Tremaine, S. (1993a). On the origin of planetary spins. *Icarus* 103 (1), 67–92. doi:10.1006/icar.1993.1059

Dones, L., and Tremaine, S. (1993b). Why does the Earth spin forward? *Science* 259 (5093), 350–354. doi:10.1126/science.259.5093.350

Duxbury, T. C., and Veverka, J. (1977). Viking imaging of Phobos and Deimos: An overview of the primary mission. *J. Geophys. Res.* 82 (28), 4203–4211. doi:10.1029/JS082i028p04203

Duxbury, T. C., Zakharov, A. V., Hoffmann, H., and Guinness, E. A. (2014). Spacecraft exploration of Phobos and Deimos. *Planet. Space Sci.* 102 (C), 9–17. doi:10.1016/j.pss.2013.12.008

Ebel, D. S., and Brook, S. (2015). "Spectral properties of Phobos from the Mars global surveyor thermal emission spectrometer: Evidence for water and carbonate," in 46th Lunar and Planetary Science Conference, Woodlands, Texas, March 16–20, 2015, 5–6.

Fassett, C. I., and Head, J. W. (2008). Valley network-fed, open-basin lakes on Mars: Distribution and implications for Noachian surface and subsurface hydrology. *Icarus* 198 (1), 37–56. doi:10.1016/j.icarus.2008.06.016

Filiberto, J., and Schwenzer, S. P. (2019). "Introduction to volatiles in the martian crust," in *Volatiles in the martian crust* (Amsterdam, Netherlands: Elsevier), 1–12.

Fraeman, A. A., Arvidson, R. E., Murchie, S. L., Rivkin, A., Bibring, J. P., Choo, T. H., et al. (2012). Analysis of disk-resolved OMEGA and CRISM spectral observations of Phobos and Deimos. *J. Geophys. Res. E Planets* 117 (10), 1–11. doi:10.1029/2012JE004137

Fraeman, A. A., Murchie, S. L., Arvidson, R. E., Clark, R. N., Morris, R. v., Rivkin, A. S., et al. (2014). Spectral absorptions on Phobos and Deimos in the visible/near infrared wavelengths and their compositional constraints. *Icarus* 229, 196–205. doi:10.1016/j.icarus.2013.11.021

Galimov, E. M. (2010). Phobos sample return mission: Scientific substantiation. *Sol. Syst. Res.* 44 (1), 5–14. doi:10.1134/S0038094610010028

Garry, J. R. C., Loes ten Kate, I., Martins, Z., Nørnberg, P., and Ehrenfreund, P. (2006). Analysis and survival of amino acids in Martian regolith analogs. *Meteorit. Planet. Sci.* 41 (3), 391–405. doi:10.1111/j.1945-5100.2006.tb00470.x

Gatley, I., Kieffer, H., Miner, E., and Neugebauer, G. (1974). Infrared observations of PHOBOS from mariner 9. *Astrophysical J.* 190, 497–503. doi:10.1086/152902

Giuranna, M., Roush, T. L., Duxbury, T., Hogan, R. C., Carli, C., Geminale, A., et al. (2011). Compositional interpretation of PFS/MEx and TES/MGS thermal infrared spectra of Phobos. *Planet. Space Sci.* 59 (13), 1308–1325. doi:10.1016/J.PSS.2011.01.019

Hall, A. (1877). The satellites of Mars. *Mon. Not. Roy. Astron. Soc.* 28, 206.

Hesselbrock, A., and Minton, D. (2017). An ongoing satellite–ring cycle of Mars and the origins of Phobos and Deimos. *Nat. Geosci.* 10, 266–269. doi:10.1038/ngeo2916

Hu, X., Oberst, J., and Willner, K. (2020). Equipotential figure of Phobos suggests its late accretion near 3.3 Mars radii. *Geophys. Res. Lett.* 47, 1–7. doi:10.1029/2019GL085958

Hunten, D. M. (1979). Capture of Phobos and Deimos by photoatmospheric drag. *Icarus* 37 (1), 113–123. doi:10.1016/0019-1035(79)90119-2

Hyodo, R., Genda, H., Charnoz, S., and Rosenblatt, P. (2017). On the impact origin of Phobos and Deimos. I. Thermodynamic and physical aspects. *Astrophysical J.* 845 (2), 125. doi:10.3847/1538-4357/AA81C4

Hyodo, R., Genda, H., Sekiguchi, R., Madeira, G., and Charnoz, S. (2022). Challenges in forming Phobos and Deimos directly from a splitting of an ancestral single moon. *Planet. Sci. J.* 3, 204. doi:10.3847/PSJ/ac88d2

Hyodo, R., Kurosawa, K., Genda, H., Usui, T., and Kazuhisa, F. (2019). Challenges in forming Phobos and Deimos directly from a splitting of an ancestral single moon. *Planet. Sci. J.* 3, 204–206. doi:10.3847/PSJ/ac88d2

Hyodo, R., and Usui, T. (2021). Searching for life on Mars and its moons: Sample-return missions will look for extraterrestrial life and biomarkers on Mars and Phobos. *Science* 373 (6556), 742. doi:10.1126/science.abj1512

Ichiro Kawaguchi, J. (1999). "On the lunar and heliocentric gravity assist experienced in the planet-B ("NOZOMI")," in Fourth International Symposium on Space Flight Dynamics, Foz do Iguaçu, Brazil, February 8–12, 1999.

Kral, T. A., Bekkum, C. R., and McKay, C. P. (2004). Growth of methanogens on a Mars soil simulant. *Orig. Life Evol. Biosphere* 34 (6), 615–626. doi:10.1023/B:ORIG.0000043129.68196.5F

Kuramoto, K., Kawakatsu, Y., Fujimoto, M., Araya, A., Barucci, M. A., Genda, H., et al. (2022). Martian moons exploration MMX: Sample return mission to Phobos elucidating formation processes of habitable planets. *Earth Planets Space* 74 (12), 12–31. doi:10.1186/s40623-021-01545-7

Kuramoto, K., Kawakatsu, Y., Fujimoto, M., Genda, H., Imamura, T., Kameda, S., et al. (2018). "Martian moons exploration (MMX) conceptual study update," in 49th Lunar and Planetary Science Conference, Woodlands, TX, March 19-23, 2018, 1–2.

Landsman, Z. A., Schultz, C. D., Britt, D. T., Peppin, M., Kobrick, R. L., Metzger, P. T., et al. (2021). Phobos regolith simulants PGI-1 and PCA-1. *Adv. Space Res.* 67 (10), 3308–3327. doi:10.1016/j.asr.2021.01.024

Lang, J. J., Baker, J. D., Castillo-Rogez, J. C., Mcelrath, T. P., Piacentine, J. S., and Snyder, J. S. (2012). *Phobos surveyor - Phobos exploration using two small solar electric propulsion spacecraft*. Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration.

Lange, M. A., and Ahrens, T. J. (1982). The evolution of an impact-generated atmosphere. *Icarus* 51 (1), 96–120. doi:10.1016/0019-1035(82)90031-8

Lee, P., Benna, M., Britt, D., Colaprete, A., Davis, W., Delory, G., et al. (2014). PADME (Phobos and Deimos and Mars environment): A proposed NASA discovery mission to investigate the two moons of Mars. Available at: https://ntrs.nasa.gov/api/citations/20150001913/downloads/20150001913.pdf.

Lee, P., Veverka, J., Bel-Lerose, J., Boucher, M., Boynton, J., Braham, S., et al. (2010). "Hall: A Phobos and Deimos sample return mission," in 41st Lunar and Planetary Science Conference, Woodlands, TX, March 1-5, 2010.

Liang, Y., and Hyodo, R. (2023). Giga-year dynamical evolution of particles around Mars. *Icarus* 391, 115335. doi:10.1016/j.icarus.2022.115335

Lunine, J. I., Neugebauer, G., and Jakosky, B. M. (1982). Infrared observations of Phobos and Deimos from viking. *J. Geophys. Res.* 87 (12), 10297. doi:10.1029/jb087ib12p10297

Marov, M. Y., Avduevsky, V. S., Akim, E. L., Eneev, T. M., Kremnev, R. S., Kulikov, S. D., et al. (2004). Phobos-grunt: Russian sample return mission. *Adv. Space Res.* 33 (12), 2276–2280. doi:10.1016/S0273-1177(03)00515-5

Masursky, H., Batson, R. M., Mccauley, J. F., Soderblom, L. A., Wildey, R. L., Carr, M. H., et al. (1972). Mariner 9 television reconnaissance of Mars and its satellites: Preliminary results. *Science* 175 (4019), 294–305. doi:10.1126/science.175.4019.294

Metzger, P. T., Britt, D. T., Covey, S., Schultz, C., Cannon, K. M., Grossman, K. D., et al. (2019). Measuring the fidelity of asteroid regolith and cobble simulants. *Icarus* 321, 632–646. doi:10.1016/J.ICARUS.2018.12.019

Michel, P., Agnolon, D., Brucato, J., Gondet, B., Korablev, O., Koschny, D., et al. (2011). "Mmsr - a study for a martian moon sample return mission," in EPSC-DPS Joint Meeting 2011, Nantes, France, 2-7 October 2011, 849.

Miyamoto, H., Niihara, T., Wada, K., Ogawa, K., Baresi, N., Abell, P., et al. (2018). "Phobos environment model and regolith simulant for MMX mission," in 49th Lunar and Planetary Science Conference, Woodlands, TX, March 19-23, 2018, 1–2.

Moroz, L. v., Basilevsky, A. T., Hiroi, T., Rout, S. S., Baither, D., van der Bogert, C. H., et al. (2009). Spectral properties of simulated impact glasses produced from martian soil analogue JSC Mars-1. *Icarus* 202 (1), 336–353. doi:10.1016/j.icarus.2009.02.007

Mueller, J. T., Guo, Y., Von Mehlem, U. I., and Cheng, A. F. (2003). Aladdin mission concept. *Acta Astronaut.* 52 (2–6), 211–218. doi:10.1016/S0094-5765(02)00159-5

Murchie, S., Arvidson, R., Bedini, P., Beisser, K., Bibring, J. P., Bishop, J., et al. (2007). Compact reconnaissance imaging spectrometer for Mars (CRISM) on Mars reconnaissance orbiter (MRO). *J. Geophys. Res. E Planets* 112 (5), E05S03–57. doi:10.1029/2006JE002682

Murchie, S., and Erard, S. (1996). Spectral properties and heterogeneity of Phobos from measurements byPhobos 2. *Icarus* 123 (1), 63–86. doi:10.1006/icar.1996.0142

Murchie, S. L., Britt, D. T., Head, J. W., Pratt, S. F., Fisher, P. C., Zhukov, B. S., et al. (1991). Color heterogeneity of the surface of Phobos: Relationships to geologic features and comparison to meteorite analogs. *J. Geophys. Res. Solid Earth* 96 (4), 5925–5945. doi:10.1029/90JB02354

Murchie, S. L., Britt, D. T., and Pieters, C. M. (2014). The value of Phobos sample return. *Planet. Space Sci.* 102, 176–182. doi:10.1016/j.pss.2014.04.014

Murchie, S., Thomas, N., Britt, D., Herkenhoff, K., and Bell, J. F. (1999). Mars pathfinder spectral measurements of Phobos and Deimos: Comparison with previous data. *J. Geophys. Res. E Planets* 104 (E4), 9069–9079. doi:10.1029/98JE02248

Murray, J. B., and Heggie, D. C. (2014). Character and origin of Phobos' grooves. *Planet. Space Sci.* 102, 119–143. doi:10.1016/J.PSS.2014.03.001

Muscatello, A. C., Mueller, R., et al.Muscatello, A. C., Mueller, R., Sanders, G. B., and Larson, W. E. (2012). "Phobos and Deimos sample collection and prospecting missions for science and ISRU," in Concepts and Approaches for Mars Exploration, Houston, TX, June 12-14, 2012, 4296.

Nakamura, T., Ikeda, H., Kouyama, T., Nakagawa, H., Kusano, H., Senshu, H., et al. (2021). Science operation plan of Phobos and Deimos from the MMX spacecraft. *Earth, Planets Space* 73, 227–253. doi:10.21203/RS.3.RS-414828/V1

Nallapu, R. T., Dektor, G., Kenia, N., Uglietta, J., Ichikawa, S., Herreras-Martinez, M., et al. (2020). Trajectory design of Perseus: A CubeSat mission concept to Phobos. *Aerospace* 7 (12), 179. doi:10.3390/AEROSPACE7120179

Nayak, M., and Asphaug, E. (2016). Sesquinary catenae on the Martian satellite Phobos from reaccretion of escaping ejecta. *Nat. Commun.* 7 (1), 12591–12598. doi:10.1038/ncomms12591

Nénon, Q., Poppe, A. R., Rahmati, A., and McFadden, J. P. (2021). Implantation of Martian atmospheric ions within the regolith of Phobos. *Nat. Geosci.* 14 (2), 61–66. doi:10.1038/s41561-020-00682-0

Pajola, M., Roush, T., Dalle Ore, C., Marzo, G. A., and Simioni, E. (2018). Phobos MRO/CRISM visible and near-infrared (0.5–2.5 μm) spectral modeling. *Planet. Space Sci.* 154, 63–71. doi:10.1016/j.pss.2018.02.016

Pätzold, M., Andert, T., Jacobson, R., Rosenblatt, P., and Dehant, V. (2014). Phobos: Observed bulk properties. *Planet. Space Sci.* 102 (C), 86–94. doi:10.1016/j.pss.2014.01.004

Pieters, C. M., Murchie, S., Thomas, N., and Britt, D. (2014). Composition of surface materials on the moons of Mars. *Planet. Space Sci.* 102 (C), 144–151. doi:10.1016/j.pss.2014.02.008

Pignatale, F. C., Charnoz, S., Rosenblatt, P., Hyodo, R., Nakamura, T., and Genda, H. (2018). On the impact origin of Phobos and Deimos. III. Resulting composition from different impactors. *Astrophysical J.* 853, 118. doi:10.3847/1538-4357/aaa23e

Pollack, J. B., and Burns, J. A. (1977). An origin by capture for the martian satellites? *Bull. Am. Astron. Soc.* 9, 518–519.

Pollack, J. B., Burns, J. A., and Tauber, M. E. (1979). Gas drag in primordial circumplanetary envelopes: A mechanism for satellite capture. *Icarus* 37 (3), 587–611. doi:10.1016/0019-1035(79)90016-2

Pollack, J. B., Veverka, J., Noland, M., Sagan, C., Hartmann, W. K., Duxbury, T. C., et al. (1972). Mariner 9 television observations of Phobos and Deimos. *Icarus* 17 (2), 394–407. doi:10.1016/0019-1035(72)90007-3

Pollack, J. B., Veverka, J., Pang, K., Colburn, D., Lane, A. L., and Ajello, J. M. (1978). Multicolor observations of Phobos with the viking lander cameras: Evidence for a carbonaceous chondritic composition. *Science* 199, 66–69. doi:10.1126/science.199.4324.66

Ramsley, K. R., and Head, J. W. (2013). Mars impact ejecta in the regolith of Phobos: Bulk concentration and distribution. *Planet. Space Sci.* 87, 115–129. doi:10.1016/J.PSS.2013.09.005

Ramsley, K. R., and Head, J. W. (2021). The origins and geological histories of Deimos and Phobos: Hypotheses and open questions. *Space Sci. Rev.* 217, 86. doi:10.1007/s11214-021-00864-1

Ramsley, K. R., and Head, J. W. (2017). The Stickney Crater ejecta secondary impact crater spike on Phobos: Implications for the age of Stickney and the surface of Phobos. *Planet. Space Sci.* 138, 7–24. doi:10.1016/J.PSS.2017.02.004

Ray, C. S., Reis, S. T., Sen, S., and O'Dell, J. S. (2010). JSC-1A lunar soil simulant: Characterization, glass formation, and selected glass properties. *J. Non-Crystalline Solids* 356 (44–49), 2369–2374. doi:10.1016/j.jnoncrysol.2010.04.049

Raymond, C. A., Diniega, S., and Prettyman, T. H. (2015). "Pandora - discovering the origin of the moons of Mars (a proposed Discovery mission)," in AGU Fall Meeting Abstracts, San Francisco, December 14-18, 2015, 11B–2101B.

Rickman, D., Edmunson, J., and McLemore, C. (2013). Functional comparison of lunar regoliths and their simulants. *J. Aerosp. Eng.* 26 (1), 176–182. doi:10.1061/(ASCE)AS.1943-5525.0000223

Rickman, D., Patel, M. R., Pearson, V. K., Wilson, S., and Edmunson, J. (2016). A carbonaceous chondrite based simulant of Phobos. *Earth Space*, 576–583. doi:10.1061/9780784479971.054

Ronnet, T., Vernazza, P., Mousis, O., Brugger, B., Beck, P., Devouard, B., et al. (2016). Reconciling the orbital and physical properties of the martian moons. *Astrophysical J.* 828, 109. doi:10.3847/0004-637X/828/2/109

Rosenblatt, P., Charnoz, S., Dunseath, K. M., Terao-Dunseath, M., Trinh, A., Hyodo, R., et al. (2016). Accretion of Phobos and Deimos in an extended debris disc stirred by transient moons. *Nat. Geosci.* 9 (8), 581–583. doi:10.1038/ngeo2742

Rosenblatt, P. (2011). The origin of the Martian moons revisited. *Astronomy Astrophysics Rev.* 19 (1), 44. doi:10.1007/s00159-011-0044-6

Schmedemann, N., Michael, G. G., Ivanov, B. A., Murray, J. B., and Neukum, G. (2014). The age of Phobos and its largest crater, Stickney. *Planet. Space Sci.* 102 (C), 152–163. doi:10.1016/j.pss.2014.04.009

Shi, X., Oberst, J., Willner, K., Shi, X., Oberst, J., and Willner, K. (2016). Mass wasting on Phobos triggered by an evolving tidal environment. *GeoRL* 43 (24), 371–412. doi:10.1002/2016GL071650

Shingareva, T. V., and Kuzmin, R. O. (2001). Mass-wasting processes on the surface of Phobos. *SoSyR* 35 (6), 431–443. doi:10.1023/a:1013082711274

Siddiqi, A. A. (2018). Beyond Earth: A chronicle of deep space exploration. National aeronautics and space administration (NASA), office of communications, NASA history division. Available at: https://www.nasa.gov/connect/ebooks/beyond_earth_detail.html (Accessed September 28, 2018).

Siddiqi, A. A. (2002). Deep space chronicle A chronology of deep space and planetary probes 1958-2000. *Monogr. Aerosp. Hist.* 24, 173–176.

Smith, N. M., Edwards, C. S., Mommert, M., Trilling, D. E., and Glotch, T. D. (2018). "Mapping the thermal inertia of Phobos using MGS-TES observations and thermophysical modeling," in 49th Lunar and Planetary Science Conference, Woodlands, TX, March 19-23, 2018, 1–2.

Stähler, S. C., Widmer-Schnidrig, R., Scholz, J.-R., Driel, M. V., Mittelholz, A., Hurst, K., et al. (2020). Geophysical observations of Phobos transits by InSight. *Geophys. Res. Lett.* 47 (19), e2020GL089099. doi:10.1029/2020GL089099

Sternovsky, Z., Robertson, S., Sickafoose, A., Colwell, J., and Horányi, M. (2002). Contact charging of lunar and Martian dust simulants. *J. Geophys. Res. Planets* 107 (11), 15-1–15-8. doi:10.1029/2002JE001897

Thangavelautham, J., Asphaug, E., Dektor, G., Kenia, N., Uglietta, J., Ichikawa, S., et al. (2017). "An interplanetary CubeSat mission to Phobos," in Lunar and Planetary Science Conference, Woodlands, TX, March 20-24, 2017.

Thomas, N., Britt, D. T., Herkenhoff, K. E., Murchie, S. L., Semenov, B., Keller, H. U., et al. (1999). Observations of Phobos, Deimos, and bright stars with the imager for Mars pathfinder. *J. Geophys. Res. E Planets* 104 (4), 9055–9068. doi:10.1029/98JE02555

Thomas, N., Stelter, R., Ivanov, A., Bridges, N. T., Herkenhoff, K. E., and McEwen, A. S. (2011). Spectral heterogeneity on Phobos and Deimos: HiRISE observations and comparisons to Mars pathfinder results. *Planet. Space Sci.* 59 (13), 1281–1292. doi:10.1016/j.pss.2010.04.018

Thomas, P. C., Veverka, J., Bloom, A., and Duxbury, T. C. (1979). Grooves on Phobos: Their distribution, morphology and possible origin. *J. Geophys. Res.* 84, 8457–8477. doi:10.1029/JB084iB14p08457

Ulamec, S., Michel, P., Grott, M., Böttger, U., Hübers, H., Murdoch, N., et al. (2019). "A rover for the JAXA MMS mission to Phobos," in 70th International Astronautical Congress, Mmx, Washington, DC, October 21 – 2, 2019, 1–8.

Veverka, J., and Duxbury, T. C. (1977). Viking observations of Phobos and Deimos: Preliminary results. *J. Geophys. Res.* 82 (28), 4213–4223. doi:10.1029/JS082i028p04213

Warner, N., Gupta, S., Lin, S.-Y., Kim, J.-R., Muller, J.-P., and Morley, J. (2010). Late Noachian to Hesperian climate change on Mars: Evidence of episodic warming from transient crater lakes near Ares Vallis. *J. Geophys. Res. Planets* 115 (6), E06013. doi:10.1029/2009JE003522

Witasse, O., Duxbury, T., Chicarro, A., Altobelli, N., Andert, T., Aronica, A., et al. (2014). Mars express investigations of Phobos and Deimos. *Planet. Space Sci.* 102, 18–34. doi:10.1016/j.pss.2013.08.002

Yamamoto, S., Watanabe, S., and Matsunaga, T. (2018). Space-weathered anorthosite as spectral D-type material on the martian satellites. *Geophys. Res. Lett.* 45 (3), 1305–1312. doi:10.1002/2017GL076612

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Original Research
PUBLISHED 27 June 2023
DOI 10.3389/fspas.2023.1139120

Check for updates

OPEN ACCESS

EDITED BY
Matthew Parry,
University of Otago, New Zealand

REVIEWED BY
Geert Verdoolaege,
Ghent University, Belgium
Reinaldo Roberto Rosa,
National Institute of Space Research (INPE), Brazil

*CORRESPONDENCE
Héctor J. Hortúa,
hector.hortua@escuelaing.edu.co

RECEIVED 06 January 2023
ACCEPTED 12 June 2023
PUBLISHED 27 June 2023

CITATION
Hortúa HJ, García LÁ and Castañeda C. L (2023), Constraining cosmological parameters from N-body simulations with variational Bayesian neural networks.
Front. Astron. Space Sci. 10:1139120.
doi: 10.3389/fspas.2023.1139120

COPYRIGHT
© 2023 Hortúa, García and Castañeda C.. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Constraining cosmological parameters from N-body simulations with variational Bayesian neural networks

Héctor J. Hortúa[1,2]*, Luz Ángela García[3] and Leonardo Castañeda C.[4]

[1]Departamento de Matemáticas, Grupo Signos, Universidad El Bosque, Bogotá, Colombia, [2]Maestría en Ciencia de Datos, Universidad Escuela Colombiana de Ingeniería Julio Garavito Bogotá, Bogotá, Colombia, [3]Universidad ECCI, Bogotá, Colombia, [4]Observatorio Astronómico Nacional, Universidad Nacional de Colombia, Bogotá, Colombia

Introduction: Methods based on deep learning have recently been applied to recover astrophysical parameters, thanks to the ability of these techniques to capture information from complex data. One of these schemes is the approximate Bayesian neural network (BNN), which has demonstrated to yield a posterior distribution into the parameter space that is extremely helpful for uncertainty quantification. However, modern neural networks tend to produce overly confident uncertainty estimates and introduce bias when applying BNNs to data.

Method: In this work, we implement multiplicative normalizing flows (MNFs), a family of approximate posteriors for the parameters of BNNs with the purpose of enhancing the flexibility of the variational posterior distribution, to extract Ω_m, h, and σ_8 from the QUIJOTE simulations. We compared the latter method with the standard BNNs and the Flipout estimator.

Results: We have found that the use of MNFs consistently outperforms the standard BNNs with a percent difference in the mean squared error of 21{%}, in addition to high-accuracy extraction of σ_8 (r^2 = 0.99), with precise and consistent uncertainty estimates.

Discussions: These findings imply that MNFs provide a more realistic predictive distribution closer to the true posterior, mitigating the bias introduced by the variational approximation and allowing us to work with well-calibrated networks.

KEYWORDS

cosmology, N-body simulations, parameter estimation, artificial intelligence, deep neural networks

1 Introduction

Cosmological simulations offer one of the most powerful ways to understand the initial conditions of the Universe and improve our knowledge of fundamental physics (Stefano and Kravtsov, 2011). These also open the possibility to fully explore the growth of structure in both the linear and non-linear regimes. Currently, the concordance cosmological model,Λ-CDM, gives an accurate description of most of the observations from the early to late stages of the Universe using a set of few parameters (Dodelson, 2003). Recent observations from

the Cosmic Microwave Background (CMB) have provided such accurate estimation for the cosmological parameters and prompted tension with respect to local scale measurements, along with a well-known degeneracy on the total non-relativistic matter density parameters (Tinker et al., 2011; Planck et al., 2020; Yusofi and Ramzanpour, 2022). Conventionally, the way to capture information from astronomical observations is to compare summary statistics from data against theory predictions. However, two major difficulties arise with this approach: first, it is not well understood what kind of estimator (or at which degree of approximation of order statistic is better to extract the maximum information from observations). In fact, the most common choice is the power spectrum (PS), which has shown to be powerful for making inference (Dodelson, 2003). However, It is well-known that PS is not able to fully characterize the statistical properties of non-Gaussian density fields, yielding that it would not be suitable for upcoming large scale structure (LSS) or 21-cm signals which are highly non-Gaussian (Hamann et al., 2010; Mesinger et al., 2011; Abdalla et al., 2022). Then, PS would miss relevant information if only this statistic is used for parameter recovery (Gillet et al., 2019). Second, cosmologists will require to store and process a large number of data, which is computationally expensive. Clearly, advanced computational frameworks along with new perspectives on data collection, storage, and analysis must be developed in order to interpret these observations (Dvorkin et al., 2022).

In recent years, artificial intelligence (AI) and deep neural networks (DNNs) have emerged as promising tools to tackle the aforementioned difficulties in the cosmological context due to their capability for learning relationships between variables in complex data, outperform traditional estimators, and handle the demanding computational needs in astrophysics and cosmology (Dvorkin et al., 2022). These standard DNNs have been used on a variety of tasks because of their potential to solve inverse problems. Nonetheless, they are prone to overfitting due to the excessive number of parameters to be adjusted and the lack of explanations of their predictions at given instances (Guo et al., 2017). The latter is crucial for cosmological analysis where assessing robustness and reliability of the model predictions is imperative. This problem can be addressed by endowing DNNs with probabilistic properties that permit quantifying posterior distributions on their outcomes and provide them with predictive uncertainties. One of these approaches is the use of Bayesian neural networks (BNNs) that comprise probabilistic layers to capture uncertainty over the network parameters (weights) and are trained using Bayesian inference (Chang, 2021). Several works have utilized BNNs in cosmological scenarios where the combination of DNNs (through convolutional neural networks, CNNs) and probabilistic properties allows building models adapted to non-Gaussian data without requiring *a priori* choice summary statistics (Ravanbakhsh et al., 2017; Gillet et al., 2019; Lazanu, 2021; Wang et al., 2022), along with quantifying predictive uncertainties (Hortúa et al., 2020b; a; Hortua, 2021; Mancarella et al., 2022; List et al., 2020; Wagner-Carena et al., 2021). Indeed, BNNs permit inferring posterior distributions instead of point estimates for the weights. These distributions capture the parameter uncertainty, and by subsequently integrating them, we acquire uncertainties related to the network outputs. Nevertheless, obtaining the posterior distributions is an intractable task, and approximate techniques such as a variational inference (VI) must be used to put them into practice (Graves, 2011; Gunapati et al., 2022). Despite the approximate posterior distribution over the weights employed in VI providing fast inference computation, they can also introduce a degree of bias depending on how complex (or simple) the choice of the approximate distribution family is (Charnock et al., 2020). This issue yields overconfident uncertainty predictions and an unsatisfactory closeness with respect to the true posterior. Hortúa et al. (2020a) and Hortua (2021) included normalizing flows on the top of BNNs to give the joint parameter distribution more flexibility. However, that approach is not implemented into the Bayesian framework, preserving the bias.

In this paper, we attempt to enhance the flexibility of the approximate posterior distribution over the weights of the network by employing multiplicative normalizing flows (MNFs), resulting in accurate uncertainty estimates provided by BNNs. We apply this approach to N-body simulations taken from the QUIJOTE dataset (Villaescusa-Navarro et al., 2020) to show how BNNs can not only take advantage of non-Gaussian signals without requiring specifying the summary statistic (such as PS) but also increase the complexity of the posterior as they yield much larger performance improvements. This paper is organized as follows: Section 2 offers a summary of the BNNs framework and a detailed description of normalizing flow implementation. Section 3 describes the dataset and analysis tools used in this paper. Numerical implementation and configuration for BNNs are described in Section 4. Section 5 presents the results we obtained by training BNNs taking into account different approaches in addition to the inference of cosmological parameters. It also outlines the calibration diagrams to determine the accuracy of the uncertainty estimates. Finally, Section 6 draws the main conclusions of this work and possible further directions to the use of BNNs in cosmology.

2 Variational Bayesian neural networks

Here, we go into detail about BNNs and their implementation to perform parameter inference. We start with a brief introduction, before focusing on improving the variational approximation. We remind the reader to refer to Abdar et al. (2021); Gal (2016); and Graves (2011) for further details.

2.1 Approximate BNNs

The goal of BNNs is to infer the posterior distribution $p(w|\mathcal{D})$ over the weights w of the network after observing the data $\mathcal{D} = (X, Y)$. This posterior can be obtained from Bayes law: $p(w|\mathcal{D}) \sim p(\mathcal{D}|w)p(w)$, given a likelihood function $p(\mathcal{D}|w)$ and a prior on the weights $p(w)$. Once the posterior has been computed, the probability distribution on a new test example $x^\star$ is given by the following equation:

$$p\left(y^\star|x^\star,\mathcal{D}\right) = \int_w p\left(y^\star|x^\star,w\right)p\left(w|\mathcal{D}\right)dw, \tag{1}$$

where $p\left(y^\star|x^\star,w\right)$ is the predictive distribution for a given value of the weights. For neural networks, however, computing the exact posterior is intractable; so, one must resort to approximate BNNs

for inference (Gal, 2016). A popular method to approximate the posterior is variational inference (Graves, 2011). Let $q(w|\theta)$ be a family of simple distributions parameterized by θ. So, the goal of VI is to select a distribution $q(w|\theta^{\star})$ such that $\theta^{\star}$ minimizes $KL[q(w|\theta)p(w|\mathcal{D})]$, $KL[\cdot\|\cdot]$ being the Kullback–Leibler divergence. This minimization is equivalent to maximizing the evidence lower bound (ELBO) (Gal, 2016).

$$\mathrm{ELBO}(\theta) = \mathbb{E}_{q(w|\theta)}[\log p(Y|X,w)] - \mathrm{KL}[q(w|\theta)p(w)], \tag{2}$$

where $\mathbb{E}_{q(w|\theta)}[\log p(Y|X,w)]$ is the expected log-likelihood with respect to the variational posterior, and $\mathrm{KL}[q(w|\theta)\|p(w)]$ is the divergence of the variational posterior from the prior. We observe from Eq. 2 that the KL divergence acts as a regularizer forcing the variational posterior to move toward the modes of the prior. A common choice for the variational posterior is a product of independent (i.e., mean-field) Gaussian distributions, one distribution for each parameter w in the network (Abdar et al., 2021).

$$q(w|\theta) = \prod_{ij} \mathcal{N}\left(w; \mu_{ij}, \sigma_{ij}^2\right), \tag{3}$$

i and j being the indices of the neurons from the previous layer and the current layer, respectively. Applying the reparametrization trick, we arrive at $w_{ij} = \mu_{ij} + \sigma_{ij}{}^{*}\epsilon_{ij}$, where ϵ_{ij} is drawn from a standard normal distribution. Furthermore, if the prior is also a product of independent Gaussians, the KL divergence between the prior and the variational posterior be computed analytically, which makes this approach computationally efficient.

2.1.1 Flipout

In case where sampling from $q(w|\theta)$ is not fully independent for different examples in a mini-batch, we will obtain gradient estimates with high variance. The Flipout method provides an alternative to decorrelate the gradients within a mini-batch by implicitly sampling pseudo-independent weights for each example (Wen et al., 2018). The method requires two assumptions about the properties of $q(w|\theta)$: symmetric with respect to zero, and the weights of the network are independent. Under these assumptions, the distribution is invariant to element-wise multiplication by a random sign matrix $\hat{r}$, i.e., $\widehat{w} = w \circ \hat{r}$, implies that $w \sim q(w) \approx \widehat{w} \sim \widehat{q}(\widehat{w})$. Therefore, the marginal distribution over gradients computed for individual examples will be identical to the distribution computed using shared weights samples. Hence, Flipout achieves much lower variance updates when averaging over a mini-batch than shared perturbation techniques Wen et al. (2018). We will validate this approach experimentally by comparing it against MNFs.

2.2 Uncertainty in BNNs

BNNs offer a groundwork to incorporate from the posterior distribution both, the uncertainty inherent to the data (aleatoric uncertainty) and the uncertainty in the model parameters due to a limited amount of training data (epistemic uncertainty) (Kiureghian and Ditlevsen, 2009). Following the work of Hortúa et al. (2020b), let us assume that the top of the BNNs consists of a mean vector $\mu \in \mathbb{R}^N$ and a covariance matrix $\Sigma \in \mathbb{R}^{N(N+1)/2}$[1]. Feeding a fixed input $x^{\star}$, T through the network, we obtain its mean μ_t and covariance matrix Σ_t. Then, an estimator for the approximate predictive covariance can be written as follows:

$$\widehat{\mathrm{Cov}}\left(y^{*}, y^{*}|x^{*}\right) \approx \underbrace{\frac{1}{T}\sum_{t=1}^{T}\Sigma_t}_{\text{Aleatoric}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}(\mu_t - \bar{\mu})(\mu_t - \bar{\mu})^{\mathrm{T}}}_{\text{Epistemic}}, \tag{4}$$

with $\bar{\mu} = \frac{1}{T}\sum_{t=1}^{T}\mu_t$. It should be noted that in case Σ is diagonal and $\sigma^2 = \mathrm{diag}(\Sigma)$, the last equation reduces to the results obtained in the work of Kendall and Gal (2017) and Kwon et al. (2018).

$$\widehat{\mathrm{Var}}\left(y^{*}|x^{*}\right) \approx \underbrace{\frac{1}{T}\sum_{t=1}^{T}\sigma_t^2}_{\text{Aleatoric}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}(\mu_t - \bar{\mu})^2}_{\text{Epistemic}}. \tag{5}$$

In this scenario, BNNs can be used to learn the correlations between the outputs and produce estimates of their uncertainties. Unfortunately, the uncertainty computed from Eq. 4 and Eq. 5 tends to be miscalibrated, i.e., the predicted uncertainty (taking into account both epistemic and aleatoric uncertainties) is underestimated and prevents a robust detection of uncertain predictions at inference (Guo et al, 2017). Therefore, calibration diagrams along with methods to jointly calibrate aleatoric and epistemic uncertainties must be employed before inferring predictions from BNNs according to the work of Laves et al. (2020). We come back to this point in Section 5.

2.3 Multiplicative normalizing flows

As mentioned previously, the most common family for the variational posterior used in BNNs is the Gaussian mean-field distributions defined in Eq. 3. This simple distribution is unable to capture the complexity of the true posterior. Therefore, we expect that by increasing the complexity of the variational posterior, we gain significant performance due to the fact that we are now able to sample from a complicated distribution that more closely resembles the true posterior. Certainly, transforming the variational posterior must be followed with fast computations and still being numerically tractable. We now describe in detail the MNF method that provides flexible posterior distributions in an efficient way by employing auxiliary random variables and normalizing flows proposed by Louizos and Welling (2017). MNFs propose that the variational posterior can be expressed as an infinite mixture of distributions.

$$q(w|\theta) = \int q(w|z,\theta)\, q(z|\theta)\, dz, \tag{6}$$

where θ is the learnable posterior parameter and $z \sim q(z|\theta) \equiv q(z)$[2] is a vector with the same dimension of the input layer, which plays the role of an auxiliary latent variable. Moreover, allowing for local reparametrizations, the variational posterior for fully connected layers becomes a modification of Eq. 3 written as follows:

$$w \sim q(w|z) = \prod_{ij} \mathcal{N}\left(w; z_i\mu_{ij}, \sigma_{ij}^2\right). \tag{7}$$

1 The targets $y \in \mathbb{R}^N$.

2 The parameter θ will be omitted in this section for clarity of notation.

It should be noted that by enhancing the complexity of $q(z)$, we can increase the flexibility of the variational posterior. This can be done using normalizing flows since the dimensionality of z is much lower than the weights. Starting from samples $z_0 \sim q(z_0)$ from fully factorized Gaussian Eq. 3, a rich distribution $q(z_K)$ can be obtained by applying a successively invertible K-transformations f_K on z_0.

$$z_K = \mathrm{NF}(z_0) = f_K \circ \cdots \circ f_1(z_0);$$
$$\log q(z_K) = \log q(z_0) - \sum_{k=1}^{K} \log \left| \det \frac{\partial f_k}{\partial z_{k-1}} \right|. \tag{8}$$

Unfortunately, the KL divergence in Eq. 2 becomes generally intractable as the posterior $q(w)$ is an infinite mixture as shown in Eq. 6. This is also addressed in the work of Ranganath et al. (2016) by evoking Bayes law $q(z_K)q(w|z_K) = q(w)q(z_K|w)$ and introducing an auxiliary distribution $r(z_K|w,\phi)$ parameterized by ϕ, with the purpose of approximating the posterior distribution of the original variational parameters $q(z_K|w)$ to further lower bound the KL divergence term. Therefore, the KL divergence term can be bounded as follows:

$$\begin{aligned}
-KL[q(w)p(w)] &= -\mathbb{E}_{q(w)}\left[\log\left(\frac{q(w)}{p(w)}\right)\right] \\
&\geq -\mathbb{E}_{q(w)}\left[\log\left(\frac{q(w)}{p(w)}\right) + KL[q(z_K|w)r(z_K|w,\phi)\right] \\
&= -\mathbb{E}_{q(w)}\left[\log\left(\frac{q(w)}{p(w)}\right) + \mathbb{E}_{q(z_K|w)} \right. \\
&\quad \left. \times\left[\log\left(\frac{q(z_K|w)}{r(z_K|w,\phi)}\right)\right]\right] \\
&= -\mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z_K|w)}\left[\log\left(\frac{q(w)}{p(w)}\right)\right]\right. \\
&\quad \left. + \mathbb{E}_{q(z_K|w)}\left[\log\left(\frac{q(z_K|w)}{r(z_K|w,\phi)}\right)\right]\right] \\
&= -\mathbb{E}_{q(w,z_K)}\left[\log\left(\frac{q(w)}{p(w)}\right) + \log\left(\frac{q(z_K|w)}{r(z_K|w,\phi)}\right)\right] \\
&= \mathbb{E}_{q(w,z_K)}[-\log(q(w)q(z_K|w)) \\
&\quad + \log r(z_K|w,\phi) + \log p(w)] \Rightarrow \\
-KL[q(w)p(w)] &\geq \mathbb{E}_{q(w,z_K)}[-KL[q(w|z_K)p(w)] \\
&\quad + \log q(z_K) + \log r(z_K|w,\phi)],
\end{aligned} \tag{9}$$

where we have taken into account that $\mathrm{KL}[P\|Q] \geq 0$, and the equality is satisfied if $P = Q$. In the last line, the first term can be analytically computed since it will be the KL divergence between two Gaussian distributions, while the second term is given by the normalizing flow generated by f_K (in Eq. 8). Finally, the auxiliary posterior term is parameterized by inverse normalizing flows as follows (Touati et al., 2018):

$$z_0 = \mathrm{NF}^{-1}(z_K) = g_1^{-1} \circ \cdots \circ g_K^{-1}(z_K);$$
$$\log r(z_K|w,\phi) = \log r(z_0|w,\phi) + \sum_{k=1}^{K} \log \left| \det \frac{\partial g_k^{-1}}{\partial z_k} \right|, \tag{10}$$

where one can parameterize g_K^{-1} as another normalizing flow. In the paper by Louizos and Welling (2017), the authors also propose a flexible parametrization of the auxiliary posterior given as follows:

$$z_0 \sim r(z_K|w,\phi) = \prod_i \mathcal{N}(z_0; \tilde{\mu}_i(w,\phi), \tilde{\sigma}_i^2(w,\phi)). \tag{11}$$

We will use the parameterization of the mean $\tilde{\mu}$ and the variance $\tilde{\sigma}^2$ as in the original paper and the masked RealNVP (Dinh et al., 2017) as the choice of normalizing flows. However, it is crucial to keep in mind that even if we use normalizing flows which are known to be a very powerful and expressive type of density estimator, these models also overfit with little training data, leading to poor out-of-distribution (OOD) input data (Nalisnick et al. (2019); Kirichenko et al. 2020).

3 N-body simulation dataset

In this work, we work with 2000 simulated hypercubes taken from the QUIJOTE project (Villaescusa-Navarro et al., 2020). They were created using the TreePM code Gadget-III (Springel, 2005), and their initial conditions were generated at $z = 127$ using 2LPT according to the work of Scoccimarro (1998). The set chosen for this work is made of standard simulations with different random seeds with the intention of emulating the cosmic variance. Each instance corresponds to a three-dimensional distribution of the density field with size 64^3 voxels. The following cosmological parameters are varied across the dataset: the *matter density* $\Omega_m \in [0.1, 0.5]$, *baryon density* $\Omega_b \in [0.03, 0.07]$, *reduced Hubble constant* $h \in [0.5, 0.9]$, *primordial spectral index* $n_s \in [0.8, 1.2]$, and *matter fluctuation amplitude* $\sigma_8 \in [0.6, 1.0]$, while neutrino mass ($M_\nu = 0$eV) and the equation of state parameter ($w = -1$) are kept fixed. The dataset was split into training (70{%}), validation (10{%}), and test (20{%}), while hypercubes were logarithmically transformed and the cosmological parameters normalized between 0 and 1. In this paper, we will build BNNs with the ability to predict three out of five aforementioned parameters, Ω_m, σ_8, and h.

4 BNN implementation

We will consider three different BNN architectures based on the discussion presented in Section 2: standard BNNs (prior and variational posterior defined as a mean-field normal distribution) **[sBNNs]**; BNNs with a Flipout estimator **[FlipoutBNNs]**; and BNNs with MNFs **[VBNNs]**. Our pipelines are implemented using TensorFlow v:2.9 and TensorFlow-probability v:0.19 (Abadi et al., 2015). All BNNs designed in this paper are comprised of three parts. First, all experiments start with a 64^3 -voxel input layer corresponding to the normalized 3D density field followed by the fully convolutional ResNet-18 backbone as it is presented schematically in Table 1. All the ResBlocks are fully pre-activated and their representation can be seen in Figure 1. The repository classification model 3D was used to build the backbone of BNNs (Solovyev et al., 2022). Subsequently, the second part of BNNs represents the stochasticity of the network. This comprised of just one layer, and it depends on the type of BNN used. For *sBNNs*, we employ the dense variational layer which uses variational inference to fit an approximate posterior to the distribution over both the kernel matrix and the bias terms. Here, we use posterior and prior (no-trainable) normal distributions. Experiments with *FlipoutBNNs*, for instance, are made via a Flipout dense layer where the mean-field normal distribution is also utilized to parameterize the distributions. These two layers are already implemented in the package TF-probability (Abadi et al., 2015). On the other hand, for VBNNs, we have adapted the class DenseMNF implemented

TABLE 1 Configuration of the backbone BNNs used for all experiments presented in this paper.

ResNet-18 backbone		
Layer name	Input shape	Output shape
Batch norm	(N_{batch}, 64,64,64,3)	(N_{batch}, 64,64,64,3)
3D convolutional	(N_{batch}, 70,70,70,3)	(N_{batch}, 32,32,32,64)
Batch norm + ReLU	(N_{batch}, 32,32,32,64)	(N_{batch}, 32,32,32,64)
Max pooling 3D	(N_{batch}, 34,34,34,64)	(N_{batch}, 16,16,16,64)
Batch norm + ReLU	(N_{batch}, 16,16,16,64)	(N_{batch}, 16,16,16,64)
ResBlock 1	[(N_{batch}, 16, 16, 16, 64) (N_{batch}, 16, 16, 16, 64)]	(N_{batch}, 16,16,16,64)
Batch norm + ReLU	(N_{batch}, 16,16,16,64)	(N_{batch}, 16,16,16,64)
ResBlock 2	[(N_{batch}, 16, 16, 16, 64) (N_{batch}, 8, 8, 8, 128)]	(N_{batch}, 8,8,8,128)
Batch norm + ReLU	(N_{batch}, 8,8,8,128)	(N_{batch}, 8,8,8,128)
ResBlock 3	[(N_{batch}, 8, 8, 8, 128) (N_{batch}, 4, 4, 4, 256)]	(N_{batch}, 4,4,4,256)
Batch norm + ReLU	(N_{batch}, 4,4,4,256)	(N_{batch}, 4,4,4,256)
ResBlock 4	[(N_{batch}, 4, 4, 4, 256) (N_{batch}, 2, 2, 2, 512)]	(N_{batch}, 2,2,2,512)
Batch norm + ReLU	(N_{batch}, 2,2,2,512)	(N_{batch}, 2,2,2,512)
Global avg pooling	(N_{batch}, 2,2,2,512)	(N_{batch}, 512)

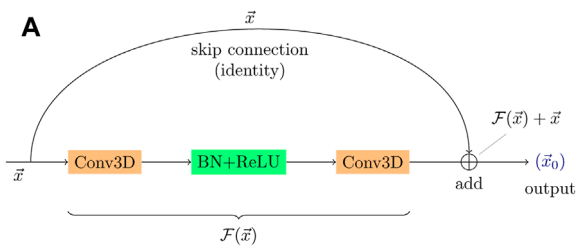


Illustration of the first skip connection in a residual block.

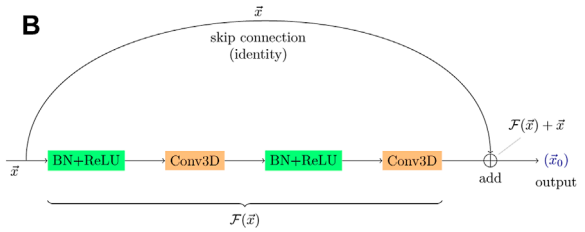


Illustration of the second skip connection in the residual block.

FIGURE 1
Each ResBlock includes both skip connection configurations. **(A)** ResBlock starts with this configuration applied to the input tensor. **(B)** Output of the previous configuration is fed into this connection.

in the repositories TF-MNF-VBNN and MNF-VBNN (Louizos and Welling, 2017) to our model. Here, we use 50 layers for the masked RealNVP NF, and the maximum variance for layer weights is around unity. Finally, the last part of all BNN accounts for the output of the network, which depends on the aleatoric uncertainty parameterization. We use a 3D multivariate Gaussian distribution with nine parameters to be learned (three means μ for the cosmological parameters and six elements for the covariance matrix Σ).

The loss function to be optimized during training is given by ELBO 2, where the second term is associated with the negative log-likelihood (NLL)

$$-\text{NLL} \sim \frac{1}{2}\log|s\cdot\Sigma| + \frac{1}{2}(\mathbf{y}-\mu)^{\top}(s\cdot\Sigma)^{-1}(\mathbf{y}-\mu) \tag{12}$$

averaged over the mini-batch. The scalar variable s is equal to one during the training process, and it becomes a trainable variable during post-training to recalibrate the probability density function (Hortúa et al., 2020b; Laves et al., 2020). The optimizer used to minimize the objective function is the Adam with first and second moments exponential decay rates of 0.9 and 0.999, respectively Kingma and Ba (2014). The learning rate starts from 10^{-3}, and it will be reduced by a factor of 0.8 in case any improvement has not been observed after 10 epochs. Furthermore, we have applied a warm-up period for which the model turns on progressively the KL term in Eq. 2. This is achieved by introducing a β variable in the ELBO, i.e., $\beta\cdot\text{KL}[q(w|\theta)p(w)]$; so, this parameter starts being equal to 0 and grows linearly to 1 during 10 epochs (Sønderby et al., 2016). BNNs were trained with 32 batches and early stopping callback to avoid overfitting. The infrastructure put in place by the Google Cloud Platform (GCP) uses NVIDIA-Tesla-T4 of 16 GB GDDR6 in an N1 machine series shared-core.

4.1 Metrics

We compare all BNN results in terms of performance, i.e., the precision of their predictions for the cosmological parameters quantified through mean square error (MSE), ELBO, and plotting the true vs. predicted values with its coefficient of determination. Moreover, it is important to quantify the quality of the uncertainty estimates. One of the ways to diagnose the quality of the uncertainty estimates is through reliability diagrams. Following the work of Laves et al. (2020) and Guo et al. (2017), we can define a perfect calibration of regression uncertainty as follows:

$$\mathbb{E}_{\widehat{\sigma}^2}\left[\text{abs}\left[\left(\|\boldsymbol{y}-\boldsymbol{\mu}\|^2 \mid \widehat{\sigma}^2=\alpha^2\right)-\alpha^2\right]\right] \quad \forall\{\alpha^2\in\mathbb{R}\mid\alpha^2\geq 0\}, \tag{13}$$

abs [.] being the absolute value function. Hence, the predicted uncertainty $\widehat{\sigma}^2$ is partitioned into K bins with equal width, and the variance per bin is defined as follows:

$$\text{var}(B_k) \coloneqq \frac{1}{|B_k|}\sum_{i\in B_m}\frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{\mu}_{i,n}-\boldsymbol{y}_i\right)^2, \tag{14}$$

with N stochastic forward passes. On the other hand, the uncertainty per bin is defined as follows:

$$\text{uncert}(B_k) \coloneqq \frac{1}{|B_k|}\sum_{i\in B_k}\widehat{\sigma}_i^2. \tag{15}$$

TABLE 2 Metrics for the test set for all BNN architectures. High UCE values indicate miscalibration. MSE and ELBO are computed only over the cosmological parameters.

Metrics	FlipoutBNNs					VBNNs					sBNNs				
	Ω_m	σ_8	h	$\Omega_m h^2$	$\sigma_8\Omega_m^{0.25}$	Ω_m	σ_8	h	$\Omega_m h^2$	$\sigma_8\Omega_m^{0.25}$	Ω_m	σ_8	h	$\Omega_m h^2$	$\sigma_8\Omega_m^{0.25}$
MSE	0.063					0.057					0.190				
ELBO	20.85					19.71					31.57				
r^2	0.82	0.98	0.2	0.03	0.93	0.85	0.99	0.4	0.56	0.95	0.75	0.85	0.01	0.23	0.80
UCE	0.109	8.10	0.26			0.0008	0.0008	0.010			>1.0				

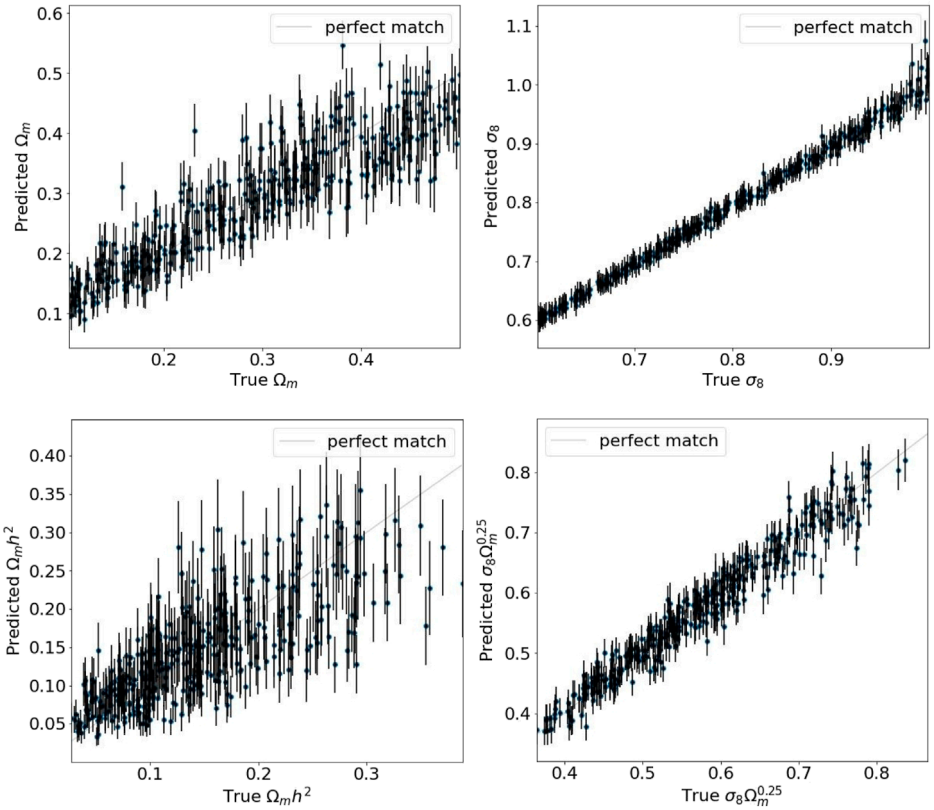


FIGURE 2
Plots of true vs. predicted values provided by the best experiment VBNNs, for Ω_m, σ_8, and some derivative parameters. Points are the mean of the predicted distributions, and error bars stand for the heteroscedastic uncertainty associated with epistemic plus aleatoric uncertainties at 1σ.

With these two quantities, we can generate reliability diagrams to evaluate the quality of the estimated uncertainty via plotting var(B_k) vs. uncert(B_k). In addition, we can compute the expected uncertainty calibration error (UCE) in order to quantify the miscalibration

$$\mathrm{UCE} := \sum_{k=1}^{K} \frac{|B_k|}{m} \left|\mathrm{var}\left(B_k\right) - \mathrm{uncert}\left(B_k\right)\right|, \tag{16}$$

with the number of inputs m and set of indices B_k of inputs, for which the uncertainty falls into the bin k. A more general approach proposed by Hortúa et al. (2020b) consists in computing the expected coverage probabilities defined as the x% of samples for which the true value of the parameters falls in the x%-confidence region defined by the joint posterior. Clearly, this option is more precise since it captures higher-order statistics through the full posterior distribution. However, for simplicity, we will follow the UCE approach.

5 Analysis and results of parameter inference with BNNs

In this section, we discuss the results obtained by comparing three different versions of BNNs, the one with MNFs, the standard BNN, and the third one using Flipout as an estimator. The results reported in this section were computed on the test dataset. Table 2 shows the metrics obtained for each BNN approach.

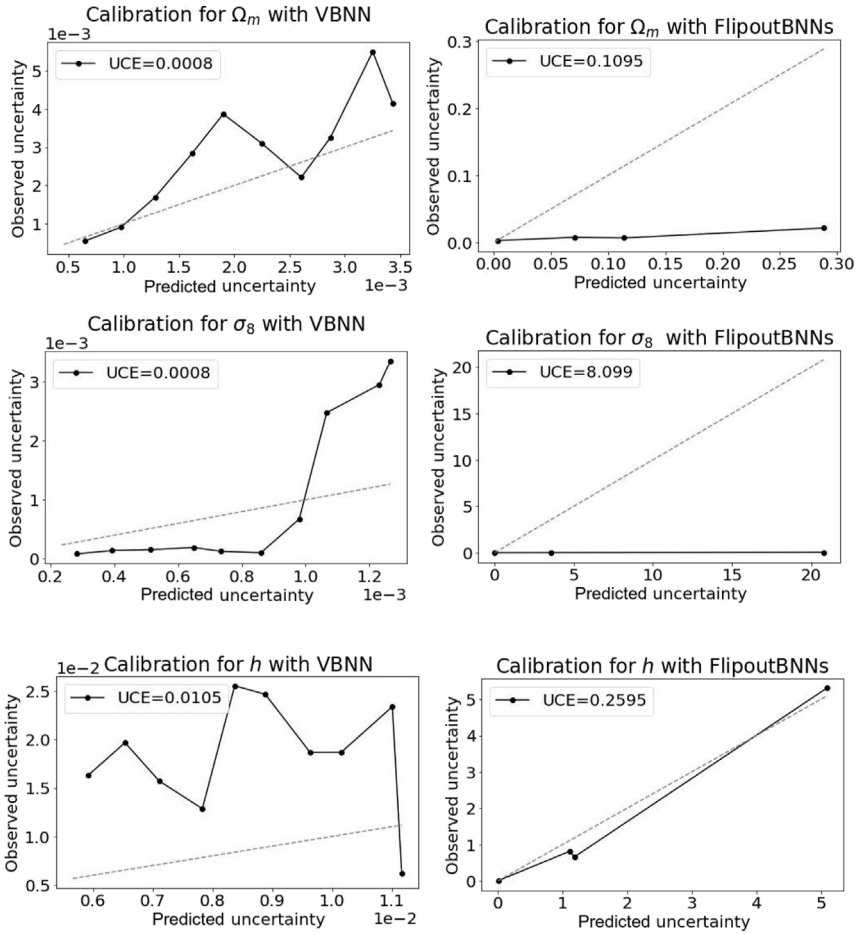


FIGURE 3
Calibration diagrams for the best experiments, VBNNs and FlipoutBNNs. The lower the UCE value, the higher the calibration of the model. Dashes lines stand for the perfect calibration, so the discrepancy to this identity curve reveals miscalibration.

TABLE 3 Parameter 95% intervals taken from the parameter constraint contours (Figure 4) from one example of the QUIJOTE test dataset using VBNNs and FlipoutBNNs.

Parameter	95% limits VBNNs	95% limits FlipoutBNNs	True value
Ω_m	$0.47^{+0.10}_{-0.10}$	$0.45^{+0.11}_{-0.11}$	0.495
σ_8	$0.697^{+0.038}_{-0.038}$	$0.699^{+0.059}_{-0.060}$	0.699
h	$0.81^{+0.17}_{-0.17}$	$0.78^{+0.20}_{-0.19}$	0.800
$\sigma_8\Omega_m^{0.25}$	$0.577^{+0.051}_{-0.052}$	$0.573^{+0.063}_{-0.064}$	0.587
$\Omega_m h^2$	$0.31^{+0.19}_{-0.18}$	$0.573^{+0.063}_{-0.064}$	0.317

As mentioned, MSE, ELBO, and r^2 provide good estimates for determining the precision of the model, while UCE measures the miscalibration. It should be noted that VBNNs outperform all experiments taking into account not only the average error but also the precision for each cosmological parameter along with a good calibration in its uncertainty predictions. Followed by VBBNs, we have the FlipoutBNNs; although this approach yields good cosmological parameter estimation, it underestimates their uncertainties. Therefore, VBNNs avoid the application of an extra post-training step in the machine learning pipeline related to calibration. It should be noted that in all experiments, h becomes hardly predicted for all models. Figure 2 displays the true against the predicted values for Ω_m, ω_m (instead of h), σ_8, and the degeneracy direction defined as $\sigma_8\Omega_m^{0.25}$. Error bars report the epistemic plus aleatoric uncertainties predicted by BNNs, which illustrate the advantages of these probabilistic models where the certainty prediction of the model is captured instead of traditional DNNs where only point estimates are present. This uncertainty was taken from the diagonal part of the covariance matrix.

5.1 Calibration metrics

In Figure 3, we analyze the quality of our uncertainty measurement using calibration diagrams. We show the predicted uncertainty vs. observed uncertainty from our model on the test dataset.

Better performing uncertainty estimates should correlate more accurately with the dashed lines. We can see that estimating

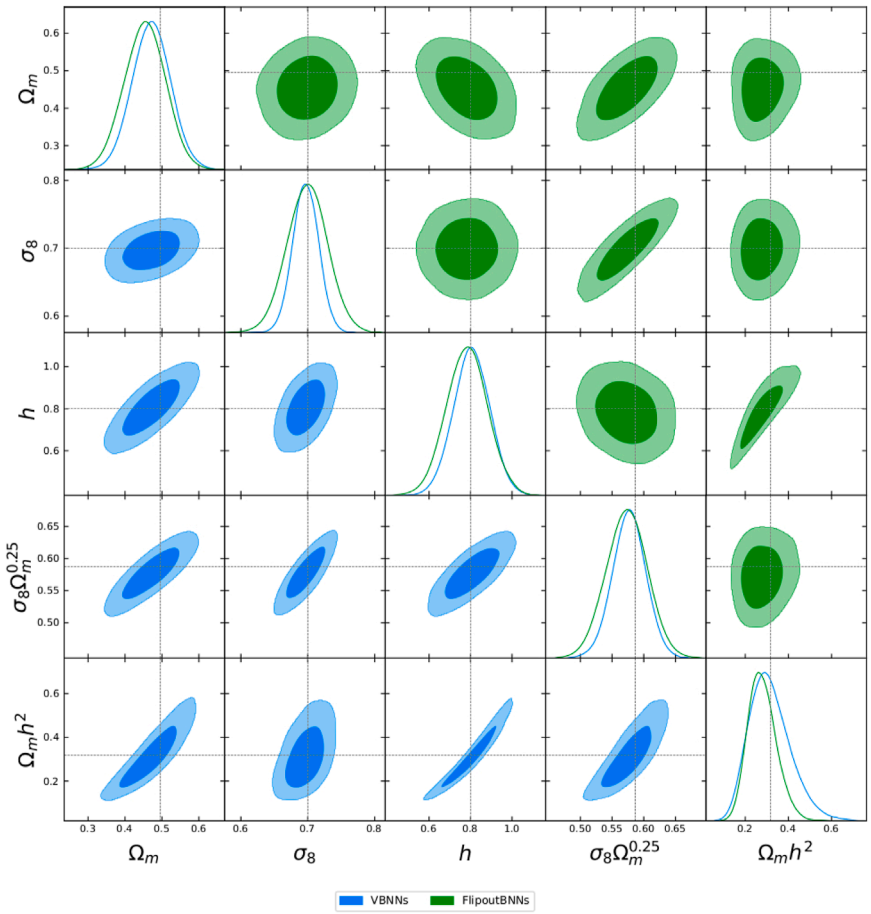


FIGURE 4
68% and 95% parameter constraint contours from one example of the QUIJOTE test dataset using VBNNs and FlipoutBNNs. The diagonal plots are the marginalized parameter constraints, and the dashed lines stand for the true values. We derive these posterior distributions using GetDist (Lewis, 2019).

uncertainty from VBNNs reflects the real uncertainty better. Furthermore, the UCE value for VBNNs is much lower than the one obtained by FlipoutBNN, which also implies how reliable this model according to their predictions is. It should be noted that even if we partitioned the variance into $K = 10$ bins with equal width, FlipoutBNNs and sBNNs yield underestimate uncertainties (many examples concentrate in lower bin values); for this reason, we see that while VBNNs supply all 10 samples in the calibration plots, for the others, we have just 3–4 of them. Next, we employed the σ-scaling methodology for calibrating the FlipoutBNN predictions (Laves et al., 2020). At doing so, we optimize only the loss function described in Eq. **12** where all parameters related to the BNNs were frozen, i.e., the only trainable parameter was s. After training, we got $s \sim 0.723$, reducing UCE only up to 10{%}, and the number of samples in the calibration diagrams enlarged to 4–5. This minor performance enhancement means that σ-scaling is not suitable to calibrate all BNNs, and alternative re-calibration techniques must be taken into account in order to build reliable intervals. At this point, we have noticed the advantages of working with methods that lead to networks already well-calibrated after the training step (Hortúa et al., 2020a).

5.2 Joint analysis for cosmological parameters

In order to show the parameter intervals and contours from the N-body simulations, we randomly choose an example from the test set with true values shown in Table 3. The two-dimensional posterior distribution of the cosmological parameters is shown in Figure 4, and the parameter 95% intervals are reported in Table 3. It should be noted that VBNNs provide considerably tighter and well constraints on all parameters with respect to the sBNNs reported by Hortua (2021). Most importantly, this technique also offers the correlation among parameters and the measurement of how reliable the model is in their predictions.

6 Conclusion

N-body simulations offer one of the most powerful ways to understand the initial conditions of the Universe and improve our knowledge of fundamental physics. In this paper, we used the QUIJOTE dataset to show how convolutional DNNs capture non-Gaussian patterns without requiring a specifying summary

statistic (such as PS). Additionally, we have shown how we can build probabilistic DNNs to obtain uncertainties that generate reliable predictions. One of the main goals of this paper was also to report the degree of improvement achieved by BNNs when we integrate with techniques such as MNFs to enhance the variational posterior complexity. We found that VBNNs not only provide considerably tighter and well constraints on all cosmological parameters as we observed in Figure 4 but also yield well-calibrated estimate uncertainties as shown in Figure 3. Although the approach used in this paper was based only on the Bayesian multilayer perceptron (B-MLP), this method is also expanded to 2D convolutions, B-CNN (Louizos and Welling, 2017). Hence, these results can be applied to other probes such as galaxy photometric redshift (CfA Redshift Survey) and large spectroscopic surveys (SDSS, DES, or LSST). Building those B-MLP and B-CNNs for additional extensions to Λ-CDM will be beneficial to reveal the presence of new physics in cosmological datasets. Nevertheless, some limitations in this research include simple prior assumptions (mean-field approximations), lower resolution in the simulations, and the absence of additional calibration techniques. These restrictions will be analyzed in detail in a future paper.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found at: https://quijote-simulations.readthedocs.io/en/latest/.

Author contributions

HH: conceptualization, methodology, and software. LG: data collection, validation, writing—original draft preparation, and cosmological analysis. LC: conceptualization and cosmological analysis. All authors contributed to the article and approved the submitted version.

Funding

This paper is based on work supported by the Google Cloud Research Credits program with the award GCP19980904. LC was supported by patrimonio autónomo fondo Nacional de financiamiento para la ciencia y la tecnología y la innovacion Francisco José de Caldas (Minciencias Colombia) grant no. 110685269447 RC-80740-465-2020 projects 69723.

Acknowledgments

HH acknowledges support from créditos educación de doctorados nacionales y en el exterior- Colciencias and the grant provided by the Google Cloud Research Credits program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems Software available from tensorflow.org.*

Abdalla, E., Abell'an, G. F., Aboubrahim, A., Agnello, A., Akarsu, O., Akrami, Y., et al. (2022). Cosmology intertwined: A review of the particle physics, astrophysics, and cosmology associated with the cosmological tensions and anomalies. *J. High Energy Astrophysics* 34, 49–211. doi:10.1016/j.jheap.2022.04.002

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* 76, 243–297. doi:10.1016/j.inffus.2021.05.008

Chang, D. T. (2021). *Bayesian neural networks: Essentials*. doi:10.48550/ARXIV.2106.13594

Charnock, T., Perreault-Levasseur, L., and Lanusse, F. (2020). *Bayesian neural networks*. doi:10.48550/ARXIV.2006.01490

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). "Density estimation using real NVP," in *International conference on learning representations*.

Dodelson, S. (2003). *Modern cosmology*. Academic Press, Elsevier Science.

Dvorkin, C., Mishra-Sharma, S., Nord, B., Villar, V. A., Avestruz, C., Bechtol, K., et al. (2022). *Machine learning and cosmology*. doi:10.48550/ARXIV.2203.08056

Gal, Y. (2016). *Uncertainty in deep learning*. Ph.D. thesis, University of Cambridge.

Gillet, N., Mesinger, A., Greig, B., Liu, A., and Ucci, G. (2019). Deep learning from 21-cm tomography of the cosmic dawn and reionization. *Mon. Notices R. Astronomical Soc.* 484, 282–293. doi:10.1093/mnras/stz010

Graves A. (Editor) (2011). *Practical variational inference for neural networks* (Curran Associates, Inc.), 24.

Gunapati, G., Jain, A., Srijith, P. K., and Desai, S. (2022). Variational inference as an alternative to mcmc for parameter estimation and model selection. *Publ. Astronomical Soc. Aust.* 39, e001. doi:10.1017/pasa.2021.64

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On calibration of modern neural networks," in *Proceedings of the 34th international conference on machine learning - volume 70 (JMLR.org)* (ICML), 17, 1321–1330.

Hamann, J., Hannestad, S., Lesgourgues, J., Rampf, C., and Wong, Y. Y. (2010). Cosmological parameters from large scale structure - geometric versus shape information. *J. Cosmol. Astropart. Phys.* 2010, 022. doi:10.1088/1475-7516/2010/07/022

Hortua, H. J. (2021). *Constraining cosmological parameters from n-body simulations with Bayesian neural networks*. doi:10.48550/ARXIV.2112.11865

Hortúa, H. J., Malagò, L., and Volpi, R. (2020a). Constraining the reionization history using Bayesian normalizing flows. *Mach. Learn. Sci. Technol.* 1, 035014. doi:10.1088/2632-2153/aba6f1

Hortúa, H. J., Volpi, R., Marinelli, D., and Malagò, L. (2020b). Parameter estimation for the cosmic microwave background with Bayesian neural networks. *Phys. Rev. D.* 102, 103509. doi:10.1103/physrevd.102.103509

Kendall, A., and Gal, Y. (2017). *What uncertainties do we need in Bayesian deep learning for computer vision?*

Kingma, D. P., and Ba, J. (2014). *Adam: A method for stochastic optimization*. doi:10.48550/ARXIV.1412.6980

Kirichenko, P., Izmailov, P., and Wilson, A. G. (2020). "Why normalizing flows fail to detect out-of-distribution data," in *Proceedings of the 34th international conference on neural information processing systems* (Red Hook, NY, USA: Curran Associates Inc.). NIPS'20.

Kiureghian, A. D., and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Struct. Saf.* 31, 105–112. doi:10.1016/j.strusafe.2008.06.020

Kwon, Y., Won, J.-H., Joon Kim, B., and Paik, M. (2018). in *International conference on medical imaging with deep learning*, 13.

Laves, M.-H., Ihler, S., Fast, J. F., Kahrs, L. A., and Ortmaier, T. (2020). "Well-calibrated regression uncertainty in medical imaging with deep learning," in *Medical imaging with deep learning*.

Lazanu, A. (2021). Extracting cosmological parameters from n-body simulations using machine learning techniques. *J. Cosmol. Astropart. Phys.* 2021, 039. doi:10.1088/1475-7516/2021/09/039

Lewis, A. (2019). *GetDist: A Python package for analysing Monte Carlo samples.*

List, F., Rodd, N. L., Lewis, G. F., and Bhat, I. (2020). Galactic center excess in a new light: Disentangling the gamma-ray sky with Bayesian graph convolutional neural networks. *Phys. Rev. Lett.* 125, 241102. doi:10.1103/physrevlett.125.241102

Louizos, C., and Welling, M. (2017). "Multiplicative normalizing flows for variational Bayesian neural networks," in *Proceedings of the 34th international conference on machine learning - volume 70 (JMLR.org)* (ICML'17), 2218–2227.

Mancarella, M., Kennedy, J., Bose, B., and Lombriser, L. (2022). Seeking new physics in cosmology with Bayesian neural networks: Dark energy and modified gravity. *Phys. Rev. D.* 105, 023531. doi:10.1103/PhysRevD.105.023531

Mesinger, A., Furlanetto, S., and Cen, R. (2011). 21cmfast: A fast, seminumerical simulation of the high-redshift 21-cm signal. *Mon. Notices R. Astronomical Soc.* 411, 955–972. doi:10.1111/j.1365-2966.2010.17731.x

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. (2019). "Do deep generative models know what they don't know?," in *International conference on learning representations*.

Planck, C., Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., et al. (2020). Planck 2018 results. *Astronomy Astrophysics* 641, A6. doi:10.1051/0004-6361/201833910

Ranganath, R., Tran, D., and Blei, D. M. (2016). "Hierarchical variational models," in *Proceedings of the 33rd international conference on international conference on machine learning - volume 48 (JMLR.org)* (ICML'16), 2568–2577.

Ravanbakhsh, S., Oliva, J., Fromenteau, S., Price, L. C., Ho, S., Schneider, J., et al. (2017). *Estimating cosmological parameters from the dark matter distribution*. doi:10.48550/ARXIV.1711.02033

Scoccimarro, R. (1998). Transients from initial conditions: A perturbative analysis. *Mon. Notices R. Astronomical Soc.* 299, 1097–1118. doi:10.1046/j.1365-8711.1998.01845.x

Solovyev, R., Kalinin, A. A., and Gabruseva, T. (2022). 3d convolutional neural networks for stalled brain capillary detection. *Comput. Biol. Med.* 141, 105089. doi:10.1016/j.compbiomed.2021.105089

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. (2016). "Ladder variational autoencoders," in *Proceedings of the 30th international conference on neural information processing systems* (Red Hook, NY, USA: Curran Associates Inc). NIPS'16, 3745–3753.

Springel, V. (2005). The cosmological simulation code gadget-2. *Mon. Notices R. Astronomical Soc.* 364, 1105–1134. doi:10.1111/j.1365-2966.2005.09655.x

Stefano, B., and Kravtsov, A. (2011). Cosmological simulations of galaxy clusters. *Adv. Sci. Lett.* 4, 204–227. doi:10.1166/asl.2011.1209

Tinker, J. L., Sheldon, E. S., Wechsler, R. H., Becker, M. R., Rozo, E., Zu, Y., et al. (2011). Cosmological constraints from galaxy clustering and the mass-to-number ratio of galaxy clusters. *Astrophysical J.* 745, 16. doi:10.1088/0004-637x/745/1/16

Touati, A., Satija, H., Romoff, J., Pineau, J., and Vincent, P. (2018). *Randomized value functions via multiplicative normalizing flows*. doi:10.48550/ARXIV.1806.02315

Villaescusa-Navarro, F., Hahn, C., Massara, E., Banerjee, A., Delgado, A. M., Ramanah, D. K., et al. (2020). The quijote simulations. *Astrophysical J. Suppl. Ser.* 250, 2. doi:10.3847/1538-4365/ab9d82

Wagner-Carena, S., Park, J. W., Birrer, S., Marshall, P. J., Roodman, A., and Wechsler, R. H. (2021). Hierarchical inference with Bayesian neural networks: An application to strong gravitational lensing. *Astrophysical J.* 909, 187. doi:10.3847/1538-4357/abdf59

Wang, B. Y., Pisani, A., Villaescusa-Navarro, F., and Wandelt, B. D. (2022). *Machine learning cosmology from void properties*. doi:10.48550/ARXIV.2212.06860

Wen, Y., Vicol, P., Ba, J., Tran, D., and Grosse, R. (2018). *Flipout: Efficient pseudo-independent weight perturbations on mini-batches*. doi:10.48550/ARXIV.1803.04386

Yusofi, E., and Ramzanpour, M. A. (2022). *Cosmological constant problem and h_0 tension in void-dominated cosmology*. doi:10.48550/ARXIV.2204.12180

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 05 July 2023
DOI 10.3389/fspas.2023.1124317

Check for updates

OPEN ACCESS

EDITED BY
Matthew Parry,
University of Otago, New Zealand

REVIEWED BY
Massimo Della Valle,
Astronomical Observatory of Capodimonte (INAF), Italy
Nial Tanvir,
University of Leicester, United Kingdom

*CORRESPONDENCE
Chao Wu,
cwu@nao.cas.cn
Chengzhi Liu,
lcz@cho.ac.cn

RECEIVED 15 December 2022
ACCEPTED 12 June 2023
PUBLISHED 05 July 2023

CITATION
Li M, Kang Z, Wu C, Liu C, Mao J, Li Z, Deng S, Niu B and Jiang P (2023), Measurement methods for gamma-ray bursts redshifts.
Front. Astron. Space Sci. 10:1124317.
doi: 10.3389/fspas.2023.1124317

COPYRIGHT
© 2023 Li, Kang, Wu, Liu, Mao, Li, Deng, Niu, and Jiang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Measurement methods for gamma-ray bursts redshifts

Mengci Li[1,2], Zhe Kang[1], Chao Wu[3]*, Chengzhi Liu[1,4]*, Jirong Mao[5,6], Zhenwei Li[1], Shiyu Deng[1,3], Bingli Niu[1,2] and Ping Jiang[1,2]

[1]Changchun Observatory, National Astronomical Observatories, Chinese Academy of Sciences, Changchun, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China, [4]Key Laboratory of Space Object and Debris Observation, Chinese Academy of Sciences, Nanjing, China, [5]Yunnan Observatories, Chinese Academy of Sciences, Kunming, China, [6]Key Laboratory for the Structure and Evolution of Celestial Objects, Chinese Academy of Sciences, Kunming, China

In the era of multi-messenger astronomy, gamma-ray bursts (GRBs) with known redshifts, especially high-redshift GRBs, are a powerful tool for studying the structure and evolution of the early Universe. We review the background, the history, and the application of measurement methods of GRB redshifts in astronomy. Based on different observation targets, two measurement methods are mainly introduced. One is on GRB afterglow, the other is on GRB host galaxy. There are various processing methods belonging to measurement methods based on afterglow, including spectral measurement method of afterglow and afterglow spectral energy distribution fitting method with improved methods. There are also numerous measurement methods based on host galaxy, such as spectral measurement method of host galaxy, template matching method of host galaxy, some automatic spectroscopic redshift measurement methods, and machine learning methods. We subsequently introduce the principles, effects, and performance of these methods. We enumerate several detection and measurement instruments, which have been used in observation. The characteristics, advantages, disadvantages, and applicability of the GRB redshift measurement methods are summarized and analyzed. Furthermore, we provide a data set of 611 GRBs with measured redshift. The data set has been collected since 1997. Analysis and statistics are presented based on this data set. We summarize the characteristics of GRBs such as location, time, and accuracy. Finally, we introduce Space-based multi-band astronomical Variable Objects Monitor (SVOM) mission dedicated to searching high redshift GRBs. We also introduce the application prospect of various redshift measurement methods in SVOM mission.

KEYWORDS

gamma-ray bursts, redshift, measurement methods, afterglow, host galaxy

1 Introduction

In the era of multi-messenger astronomy, the redshifts of gamma-ray bursts (GRBs) have been widely used in many research fields of astronomy because they can be observed up to $z \approx 8$–9 (Salvaterra et al., 2009; Cucchiara et al., 2011). We note that their redshifts distribution has a peak at $z \sim 2$ (Coward et al., 2013) which approximately corresponds to the distance of the farthest SNe-Ia ever detected (Rodney et al., 2015). Therefore, today we can state,

without fear of contradiction, that GRBs have become an important tool for the study of cosmology as a whole (Salvaterra, 2015). At present, GRB redshift shows its importance in the study of cosmological parameters (Amati et al., 2008; Amati and Della Valle, 2013; Izzo et al., 2015; Moresco et al., 2022), dark energy with GRBs (Muccino et al., 2021), core-collapse supernovae (SNe) and GRB connection (Woosley and Bloom, 2006; Kelly et al., 2008), SNe emerging from the lightcurve afterglows (Della Valle et al., 2006; Melandri et al., 2019), GRBs hosts (Modjaz et al., 2008; Fruchter et al., 2006; Savaglio et al., 2009).

Redshifts are used to derive GRB parameters, such as $E_{p,i}$, E_{iso}, *etc.* At present, the correlations of GRB parameters are often used to measure cosmological parameters Ω_M and Ω_Λ. The results are high-precision and independent by parameters of other cosmological detectors such as SN Ia. Amati and Della Valle (2013) have measured Ω_M with $E_{p,i} - E_{iso}$ correlation and obtained (for a flat Universe) that the Ω_M is in the range 0.14–0.57. The same analysis carried out on a larger sample, found $\Omega_M \sim 0.3$ (Amati and Della Valle, 2013), which is consistent with $\Omega_M = 0.29^{+0.23}_{-0.15}$ obtained by Izzo et al. (2015) by using $E_{\gamma,iso} - E_{x,iso} - E_{pk}$ parameters. More recently Moresco et al. (2022) measured for a flat Universe $\Omega_M = 0.26^{+0.23}_{-0.12}$ using the $E_{p,i} - E_{iso}$ correlation.

GRBs are suitable for studying the possible evolution of the equation from the state of the Universe when we consider the intermediate redshift case. The parameter $w(z)$ of dark energy will be affected by GRB redshift. According to GRB Hubble diagram, the evolution of $w(z)$ has different piecewise formulae in different redshift intervals. When $z \leq 1.2$, Muccino et al. (2021) found that $w = -1$, and the deviation of $w(z)$ was within 1σ. With the increasing of z, the value of $w(z)$ is better determined, but $w(z)$ seems to have a strong deviation more than 2σ.

The GRB redshifts and light curves of afterglows are helpful to study the relationship between SN and GRB. By studying the GRB and SN at different redshifts, it is found that not all SNe connect with GRBs (Woosley and Bloom, 2006). It is predicted that SN associated with GRB has basically no optical flux at $z > 1.2$. For example, the GRB 050525A has the redshift of spectrum $z = 0.606$. The shape and fluxes of the light curve confirmed that SN 2005nc was formed in the late afterglow of GRB 050525A (Della Valle et al., 2006). GRB 171010A and SN 2017htp (Melandri et al., 2019) have the same redshift at $z = 0.33$. This is another example that SN is connected with GRB.

GRB redshifts are helpful to study the physical characteristics of the host population. The spectra of GRB host galaxies with different redshifts are studied, and the characteristics of chemical abundance are obtained (Modjaz et al., 2008). Long GRBs (LGRBs) are more concentrated in the bright regions of the host galaxy than Core Collapse SNe (Fruchter et al., 2006) and seem to share the same locations as SNe-Ic (Kelly et al., 2008). In addition, the host galaxies of LGRBs are much dimmer and more irregular than those of core-collapse supernovae. The results show that LGRBs are related to the massive stars, and may be confine to chemical evolution. Most of the high redshift LGRB hosts are irregular with no bulges (Kelly et al., 2008). The optical-near infrared (NIR) photometry and spectroscopy are used to deduce the star mass, star formation rates (SFRs), dust extinction, and metal abundance of GRB host galaxy (Savaglio et al., 2009).

Some GRB detection instruments are working or being built, such as Gamma-Ray Burst Optical and Near-Infrared Detector (GROND) (Greiner et al., 2008), Swift (Gehrels et al., 2004), Fermi Gamma Ray Space Telescope (Atwood et al., 2009; GLAST Facility Science Team et al., 1999), Space-based multi-band astronomical Variable Objects Monitor (SVOM) (Wei et al., 2016) and so on.

In Section 2, we introduce the history of GRB redshift measurement. The redshift measurements of GRB 970228 and GRB 970508 (the first and second GRBs with known redshift) are emphatically reviewed. In Section 3, we mainly summarize two kinds of common measurement methods of GRB redshift, including the measurement methods based on afterglow and the measurement methods based on host galaxy. The principles, advantages, disadvantages of different methods are introduced. In Section 4, we gather 611 GRBs with redshifts since 1997. We make a statistical analysis of the data, and summarize the characteristics of GRBs, such as location, time and accuracy. We verify the application of different methods introduced above. The advantages and disadvantages are also been presented. In Section 5, we present the summary in this review. We prospect a future space mission SVOM and new GRB redshift measurement methods.

2 The history of GRB redshift measurement

The GRB redshift measurement began in 1997. On 28 February 1997, BeppoSAX satellites detected GRB 970228 and its X-ray afterglow (Costa et al., 1997; Sahu et al., 1997). The optical afterglow of GRB 970228 was also observed to obtain photometric and spectroscopic data by William Herschel Telescope (WHT), Bologna University Telescope (BUT), Rome Astro Physical Observatory (ROA), and other telescopes (Van Paradijs et al., 1997; Galama et al., 1997). The observed photometric data and images are used for theoretical interpretation and modeling analysis (Wijers et al., 1997). Moreover, an extended nebula region was found and determined as the host of GRB 970228 (Galama et al., 1997; Fruchter et al., 1999). Keck II telescope with a low-resolution spectrometer was used to observe host galaxy of GRB 970228 several times. There are two strong emission lines, which are identified as [O II] λ3727 and [O III] λ5007, and a weak emission line [Ne III] λ3869 in the spectrum of the host galaxy. However, the low resolution of the instrument makes it impossible to distinguish the double peaks of [O II] λ3727. The strong OH line in the night sky also interferes with the measurement of $H\beta$ λ4861, [O III] λ4959 and the higher Balmer line. Despite the influence of the above factors, the redshift of GRB 970228 can still be derived as $z = 0.695 \pm 0.003$ by the spectral measurement method of host galaxy. The redshift confirms GRBs lie at cosmological distance (Bloom et al., 2001a).

On 8 May 1997, BeppoSAX satellite detected GRB 970508 (Metzger et al., 1997b; Sahu et al., 1997). Subsequently, its multi-waveband afterglow was observed, and relatively complete afterglow data were derived (Reichart, 1998). Metzger et al. (1997b) estimated the redshift range by spectral measurement from optical afterglow. Spectrum of optical afterglow OT J065349+79163 were obtained

using Keck II 10-m telescope with low-resolution imaging spectrometer. The spectrum contains eight strong absorption lines of Fe II, Mg II and Mg I systems. By calculating the independent redshift for each absorption line, and considering the interference of gas clouds, a strict lower limit for the redshift of the source was set as $z \geq 0.835$. An approximate upper limit ($z \leq 2.3$) was set for the characteristic that no obvious Lyman-α absorption line in the spectrum. To sum up, the redshift range of GRB 970508 is $0.835 \leq z \leq 2.3$. Reichart (1998) used the photometric data of GRB 970508 optical afterglow to fit the optical spectral energy distribution SED, and used χ^2 fitting to get the redshift $z = 1.09^{+0.14}_{-0.41}$. In addition, Metzger et al. (1997a) and Fruchter et al. (1997) predicted that the host galaxy of GRB 970508 is a dark galaxy with $z \geq 0.835$. Bloom et al. (1998) obtain the photometric and spectroscopic data of the host galaxy from Keck II telescope, and measured a more accurate redshift. There is a strong emission line at $\lambda_{obs} = 6839.7\AA$ and a weak emission line at $\lambda_{obs} = 7097.7\AA$ in the spectrum. Then, the two emission lines are identified as [O II] λ3727 and [Ne III] λ3869 lines. Finally, the redshift calculated by weighted average is $z = 0.8349 \pm 0.0003$. The spectral measurement of GRB 970508 also confirmed the cosmological GRB (Metzger et al., 1997b).

Subsequently, more and more GRBs with redshifts have been measured, such as GRB 970828 at $z = 0.9579$ (Djorgovski et al., 2001), GRB 971214 at $z = 3.42$ (Kulkarni et al., 1998), GRB 980326 at $z \sim 1$ (Bloom et al., 1999), *etc.* After the Swift satellite was launched in 2004, the number of measured GRBs redshifts greatly increased (Lin, 2009). High redshift GRBs can be detected. At present, GRB redshift ranges from $z = 0.0085$ (GRB 980425) (Galama et al., 1998) to $z \sim 9.4$ (GRB 090429B) (Cucchiara et al., 2011). Especially, a few GRBs with $z \geq 6$ have been detected, such as GRB 080913 at $z = 6.7$ (Greiner et al., 2009a), GRB 090423 at $z \sim 8.2$ (Tanvir et al., 2009), GRB 120521C at $z \sim 6$ (Laskar et al., 2014), GRB 120923A at $z \sim 7.8$ (Tanvir et al., 2018). High-redshift GRBs are considered as a new and powerful tool to explore the early Universe (Wei et al., 2015). As LGRBs are related to the death of massive stars. They can be born in the interstellar matter (ISM) of primordial galaxies, and unique probes in the re-ionization era (Heintz et al., 2019; Tanvir et al., 2019; Saccardi et al., 2023). In principle, GRB afterglow can be detected at $z \sim 20$ (Ciardi and Loeb, 2000; Gou et al., 2004). The early Universe with z = 6–20 is the crucial period in cosmology at present (Wei et al., 2015).

3 Measurement methods of GRB redshift

Based on different observation targets, two types of measurement methods are mainly introduced. One is on GRB afterglow, the other is on GRB host galaxy. Besides, we also introduce a pseudo-redshift estimation method without redshift measurement.

Based on the principles and characteristics of different measurement methods summarized in Section 3.1–3.3, we summarize the applications of different measurement methods. We also present a comparison of the advantages and disadvantages of each method in Table 1.

3.1 Measurement methods based on afterglow

The GRB prompt emission duration is extremely short, usually between a few milliseconds and a few minutes. GRB afterglow is occurring later than prompt emission and lasts longer time, so is beneficial for target tracking and observation. Among them, X-ray and optical afterglow emission can last for several days to several weeks, and radio afterglow radiation can last for several years (Zhang, 2007). At present, the afterglows used to measure GRB redshifts are generally in optical, NIR and ultraviolet bands. In the BeppoSAX era, GRB redshifts were measured by the emission lines of X-ray afterglow (Antonelli et al., 2000). In principle, Swift/XRT can be used to measure the redshift with Fe line emission (Burrows et al., 2003). However, Swift never measured the redshift by using the X-ray emission line in observation.

The advantage of afterglow measurement is that afterglow is usually brighter than the host galaxy (Curran et al., 2008). The spectrum and photometry data of afterglow can be obtained. Compared with galaxy optical lines, many prominent absorption lines in the afterglow are bluer. Thus, GRB afterglow at higher redshift can be measured in optical spectroscopy of afterglows. For example, Lyman-α absorption line in GRB afterglow is one of the most important features for redshifts measurement in the case of $z > 2$. In the case of $z \sim 3$, there are many prominent interstellar absorption lines of Mg II, Mg I, Ca II, Fe II often seen in optical spectrum. These absorption lines have rest wavelengths in the 2,000–4,000$\AA$ range. By contrast, usually the bluest emission line O II of the GRB host galaxy is at 3,727 $\AA$. So, the common emission lines of the GRB host galaxy are in the optical band at $z = 0$. The emission lines at $z > 1.5$ tends to be beyond the range of most optical spectrometers. On the other hand, the GRB afterglow continuous spectrum is usually described as a simple power law. The GRB afterglow is typically affected by Lyman-α forest and dust extinction, in addition to discrete absorption lines from the host and any intervening absorbers (Curran et al., 2008). Lyman-α and Lyman-break features are prominent features, which are included in the GRB afterglow continuous spectrum (Krühler et al., 2011b). The spectral energy distribution of GRB afterglow, basically described as a smooth power law, is simpler than that of host galaxy. The absorption features of GRB afterglows have a fairly predictable range and less uncertainty. By contrast, galaxy spectra are very complex, which are the superposition of light from stars, nebulae, and the absorption from interstellar medium. At $z \geq\sim 2.5$ where the Lyman-α break passes through the u band, the photometric redshifts measurement for GRBs are more robust than galaxy photometric redshifts measurement. The measurement methods based on GRB afterglow are more suitable for high-redshift GRB determination. It indicates that the optical afterglow is dark (Mirabel and Rodrigues, 2003; Roming et al., 2006a). The optical afterglow measurement can be affected by the local GRB environment, the type of host galaxy, and other related extinction factors (Coward et al., 2008). Thus, it makes an ineffective redshift measurement.

3.1.1 Spectral measurement method of afterglow

GRB Redshift measurement usually relies on optical and NIR spectroscopy in the medium and large telescopes (Tanvir and Jakobsson, 2007). The observation spectrum

TABLE 1 Advantages, disadvantages and applicability of different measurement methods of GRB redshift

Method		Advantages	Disadvantages	Applications
Measurement methods based on afterglow	Spectral measurement method of afterglow	High precision; Afterglow is brighter than the host, high quality and high redshift spectrum can be obtained.	Long exposure time; Affected by noise; If the afterglow decays too fast, it will affect the accuracy.	GRBs with slow afterglow decay, high luminosity; No detected host galaxy.
	Afterglow spectral energy distribution fitting method	The principle is simple; The speed is fast; The exposure time is short; GRBs with low luminosity and high redshift can be measured.	Low precision; The GRBs have great individual differences, so it is difficult to obtain a complete model.	GRBs with high redshift; Dark afterglow; No detected host galaxy.
Measurement methods based on host galaxy	Spectral measurement method of host galaxy	High precision; Long exposure and multiple observations can be allowed.	The host galaxy is dark, which affects the quality of spectral lines; There may be multiple candidates, which will affect the accuracy.	Dark Bursts; Low redshift GRB
	Automatic spectroscopic redshift measurement methods	High precision; Automatic redshift measurement; Fast speed; High efficiency	Some targets have large errors, which can't reach the precision of manual measurement; Template needs to be established; Complex algorithm	Rapidly determine the redshift range of multiple host candidates
	Template matching method of host galaxy	The speed is fast; The exposure time is short; The intra-class difference of galaxies is small, the SED template is relatively standard, dark high-redshift galaxies and outside the spectral measurement limit can be observed.	Low precision; No universal SED template; Multiple candidates, which will affect the accuracy.	Dark Bursts; Dark host galaxies; GRBs with relatively high redshift in host galaxies.
	Machine learning	High precision; Automatic redshift measurement; Fast speed; Good robustness; Strong fault tolerance; High efficiency	A training set that requires spectral redshift; Complex algorithm	Rapidly determine the redshift range of multiple host candidates
Pseudo-redshift estimation method	Luminosity-indictor relationships	The estimation method is independent of observed real redshift; Not affected by the redshift selection effect; Get larger amount of redshift data than obversed.	Affected by correlation; Large errors	Popular in early research; It is suitable for GRBs with strong correlation of photometric indexes; Compared with other methods, it is less used at present.

10.3389/fspas.2023.1124317

is matched with laboratory spectrum. Usually, GRB optical afterglow do not show emission lines.[1] We can compare one observed line of the spectrum to the line in the laboratory spectrum, and the weighted average operation is performed to derive spectroscopic redshift value of the target. The method was applied in GRB 060714 ($z = 2.711$) (Fynbo et al., 2009; Jakobsson et al., 2006a), GRB 080210 ($z = 2.6419$) (Fynbo et al., 2009), GRB 210905A ($z = 6.312$) (Rossi et al., 2022).

The accuracy of the measured redshift is high, up to about 10^{-4}. However, the signal-to-noise ratio of absorption line can affect the accuracy of spectroscopic redshift. The optical afterglow is too fast to have a sufficient bright spectrum. Due to some optical afterglow decaying in a short time, and the brightness is insufficient. Thus, the spectrum with strong absorption may not be obtained.

Spectral measurement method of afterglow is suitable for relatively bright GRBs with slow delay. In fact, as long as one GRB has bright afterglow, when a clear spectrum can be obtained, the very high redshift of target can be measured. For example, GRB 090423 has a high redshift of $z \sim 8$ (Salvaterra et al., 2009; Tanvir et al., 2009; Oates and Cummings, 2009). Due to the bright near-infrared afterglow and no counterpart at optical wavelengths, Tanvir et al. (2009) used the ISAAC spectroscopy and VLT spectroscopy by spectral measurement method of afterglow, as well as the photometric data to derive the redshift of $z = 8.23^{+0.06}_{-0.07}$. Salvaterra et al. (2009) used TNG to derive the spectroscopic redshift of $z = 8.1^{+0.1}_{-0.3}$.

The common instruments for measuring afterglow spectrum include Keck, VLT, GTC and so on. Due to the limitation of the atmosphere, the occurrence time and the position of the source in the sky, in a few cases, the afterglow spectrum can be obtained by using ground near-ultraviolet equipment such as VLT-UVES(Fiore et al., 2005; Vreeswijk et al., 2007). Ultraviolet afterglow can also be detected by Swift/UVOT (Kuin et al., 2009) to measure redshift.

3.1.2 Afterglow spectral energy distribution fitting method (photometric redshift measurement of afterglow)

We used spectral energy distribution (SED) fitting method of GRB afterglow to obtain photometric redshift. The medium and small aperture telescopes are used to perform photometric observations of the afterglow in multi-bands. Then SED fitting method is used to estimate the photometric redshift of target. Some templates are necessary for SED fitting method. The synthetic template is compared with the observed data, so that it can be reasonably extrapolated to the unmeasured redshift range. When χ^2 is the smallest, the redshift of target is considered (Bolzonella et al., 2000). X-ray and optical afterglows are usually used, and the SED shape depends on cooling frequency (ν_c), optical frequency (ν_o) and X-ray frequency (ν_x) (Japelj et al., 2015; Zaninoni et al., 2013). The SED includes two prominent features, Lyman-α and Lyman Break at $121.5(1+z)$ nm and $91.2(1+z)$ nm respectively, and GRB redshift measurement mainly depends on them (Krühler et al., 2011b).

When $\nu_c < \nu_o$ or $\nu_c > \nu_x$, the SED can be described with a single power-law as Eq. 1

$$F_\nu = F_0 \nu^{-\beta_{O,X}} \tag{1}$$

Where ν is the observed frequency, $\beta_{O,X}$ the spectral index and F_0 the normalization.

And, when $\nu_o < \nu_c < \nu_x$, the SED can be described with a broken power-law as Eq. 2

$$F_\nu = F_0 \begin{cases} \nu^{-\beta_O} & \nu \le \nu_c \\ \nu_c^{\Delta\beta}\ \nu^{-\beta_X} & \nu > \nu_c \end{cases} \tag{2}$$

Where $\Delta\beta = \beta_X - \beta_O$.

At present, Hyperz, New Hyperz, *etc.*, are usually used to fit SED (Cucchiara et al., 2011; Afonso et al., 2011; Zaninoni et al., 2013; Rossi et al., 2008). The dust extinction laws in the Milky Way (MW), Small Magellanic Clouds (SMC) and Large Magellanic Clouds (LMC) are added to the synthetic template. This method was applied in GRB 050814 ($z = 5.3 \pm 0.3$) (Jakobsson et al., 2006b), GRB 070802 ($z = 2.47^{+0.18}_{-0.15}$) (Krühler et al., 2008), GRB 080514B ($z = 1.8^{+0.4}_{-0.3}$) (Rossi et al., 2008), GRB 090429B ($z \sim 9.4$) (Cucchiara et al., 2011), GRB 120521C ($z = 6.01^{+0.05}_{-0.09}$) (Laskar et al., 2014).

In order to make the characteristics of the synthetic template approach to the real afterglow, some improved SED fitting methods are proposed. Curran et al. (2008) simulated the absorption effect caused by Lyman forest and host extinction, and used both a genetic algorithm-based routine and Monte Carlo error analysis to simultaneously fit the parameters of interest. Japelj et al. (2015) corrected the MW extinction laws and photoelectric absorption. Littlejohns et al. (2014) increased the host dust extinction and the neutral hydrogen effect along the line of sight. Krühler et al. (2011b) considered several systematic uncertainties, including the uncertainty in the dust redding properties, the opacity of the Lyman-α forest, and the strength of the Damped Lyman-α Absorber (DLA) associated with the GRB. All the above methods are reliably to derive the GRB photometric redshift. The cases are GRB 050904 ($z = 6.61 \pm 0.14$) (Curran et al., 2008), GRB 100905A ($z = 7.88^{+0.75}_{-0.94}$) (Bolmer et al., 2018), GRB 191016A ($z = 3.29 \pm 0.4$) (Smith et al., 2021).

Compared to spectroscopic redshift, the accuracy of photometric redshift is lower. Greiner et al. (2008) proposed that the error of GRB photometric redshift is in $\Delta z \sim 0.3$–0.5. Krühler et al. (2011b) proposed that the best result is $\eta = \frac{\Delta z}{1+z} \sim 0.03$. By comparing GRBs with known spectroscopic redshift, the accuracy of photometric redshift is lower than that of spectroscopic redshift derived by using a low-resolution spectrometer, about 10^2–10^3 error.

Each GRB has certain individual differences. Thus, a standard template is difficult to meet the characteristics of all GRBs. However, photometric measurement is suitable for relatively optical-NIR GRB afterglow.

Some common telescopes for photometric measurements are GROND, Robotic Optical Transient Search Experiment (ROTSE-III), Rapid Eye Mount telescope (REM). In addition, Swift/UVOT are also perform for the photometric measurement. When Lyman-α

1 The point is quoted from anonymous reviewers' comments.

and Lyman-break pass into the filter ranges, accurate GRB photometric redshift can be derived. In other words, with different redshifts, Lyman features enter different filters. For example, redshift sensitivity to Lyman-α cutoff between UVW2 band (central wavelength λ_c = 188 nm) and UVM2 band (λ_c = 217 nm) is $z \sim 1.3$, between UVM2 band and UVW1 band (λ_c = 251 nm) is $z \sim 1.7$, between UVM1 band and U band (λ_c = 345 nm) is $z \sim 2.3$, between U band and B band (λ_c = 439 nm) is $z \sim 3.2$, between B band and V band (λ_c = 544 nm) is $z \sim 3.5$ (Krühler et al., 2011b; Roming et al., 2005).[2] Lyman-break falls from g' band (λ_c = 468.6 nm) to z' band (λ_c = 893.1 nm) at $3.5 < z < 6.5$, and from z' band to J band ($\lambda_c = 1.25\ \mu m$) at $z \sim 8$. Afterglows are only detected in H band ($\lambda_c = 1.65\ \mu m$) or Ks band ($\lambda_c = 2.16\ \mu m$) at $z \sim 10$ (Krühler et al., 2011b; Greiner et al., 2008; Skrutskie et al., 2006; Abazajian et al., 2009).[3]

It appears that targets with different redshift ranges can be measured using different filter bands. The measurable redshift range of Swift/UVOT is $0.5 < z < 3.5$ (Kuin et al., 2009), covering the 170–600 nm wavelength, and the measurable redshift range of GROND is $3.5 \leq z \leq 13$ (Greiner et al., 2008), covering the 360–2,300 nm wavelength. Therefore, multiple devices are often used for combined measurement.

3.2 Measurement methods based on host galaxy

Host galaxy provides important information on the GRB environment. GRB host has the individual differences. For example, the host of GRB 970228 is essentially a face-on late-type blue dwarf galaxy (Bloom et al., 2002), GRB 970508 is in a compact, elongated and blue galaxy (Bloom et al., 2002), GRB 980425 is suspected to be in a nearby spiral galaxy (Conselice et al., 2005), the host galaxy of GRB 981226 appears to have a doubly nucleated morphology (Bloom et al., 2002), GRB 000418 is in a compact galaxy (Bloom et al., 2002).

By the statistical analysis, GRB hosts, which have been observed, are usually young, blue and low-redshift, in addition to low mass and low metallicity (Savaglio et al., 2009) (Perley et al., 2016b). But there are also a few GRBs, especially Dark Bursts, found in massive, metal-rich and high-redshift galaxies (Krühler et al., 2012; Perley et al., 2016b). Due to the correlation between mass and metallicity, GRB host galaxies have lower luminosity and smaller sizes than typical galaxies (Le Floc'h et al., 2003; Fruchter et al., 2006). Bloom et al. (2002) found that the hosts are compact and blue with high central surface brightness, when GRBs are found nearby their host centers. Svensson et al. (2010) found that GRB hosts are always fainter than those of SN counterparts, due to the low metallicity of the GRB host. Moreover, GRB hosts are irregular, small, low mass with more actively star forming. Perley et al. (2016b) found that NIR luminosities of hosts are rapidly increase in low redshift range ($0.5 < z < 1.5$). However, they are almost unchanged in mid-high redshift ($1.5 < z < 5$). The low-redshift hosts are fainter than tracers with star formation uniformly, and the mid-high redshift hosts are similar to tracers. At low redshift, there is almost no GRB in massive galaxies.

Short GRBs (SGRBs) originate from the mergers of compact objects, such as binary neutron stars mergers and neutron star-black holes mergers (Fruchter et al., 2006; Nugent et al., 2022). About 84% of SGRB hosts are star-forming galaxies (Nugent et al., 2022). The hosts of SGRB have SFRs consistent with their stellar masses, which is about 1.44 $M_{\odot} yr^{-1}$. They have usually a wide age range ($\sim 0.1 - 9\ Gyr$) with low redshift (Nugent et al., 2022; Berger et al., 2005; Barthelmy et al., 2005).

LGRBs are usually found in extremely blue galaxies with strong emission lines, indicating that this is an apparent abundance of young and massive stars (Fruchter et al., 2006). LGRB hosts are mainly irregular galaxies with low mass, and they have relatively higher redshift median than SGRB hosts (Lyman et al., 2017). Only a few LGRBs are found in spiral galaxies (Fruchter et al., 2006; Conselice et al., 2005). LGRB hosts with low redshifts have higher concentration and lower luminosity than high redshift hosts (Lyman et al., 2017). LGRBs are centrally concentrated, strongly biased towards the brighter regions in hosts. It is found that LGRB hosts are fainter than SFR-weighted field galaxy populations at the same redshift.

The advantage of host galaxy redshift measurement is that hosts have constant luminosity. Unlike the rapidly fading afterglow, the GRB host galaxy can be detected by increasing the exposure time, so that we can obtain more accurate spectrum and photometric data. Host galaxy can be detected many times in a long-time range, reducing the measurement error. At low redshift, Lyman-break may not enter into the photometric bands of UVOT, so the spectral features of afterglow are not clear. The host galaxy measurement is more suitable for low redshift GRB. For "Dark Burst", where optical afterglow is hardly observed (O'Connor et al., 2022), measurement methods based on host galaxy are important tools to derive redshift.

Some GRBs host galaxies are failed to be detected. It is because some GRBs were kick out from the host galaxies before the bursts (Bloom et al., 2002), or the GRBs have a high redshift and the host galaxy are too dark to be detected. Several candidate galaxies are near the GRB position, so it is difficulty to identify host. In addition, due to the influence of "Redshift Desert", the lack of strong emission lines affects the measurement of GRB with $z = 1$–2 (Fiore et al., 2007).

3.2.1 Spectral measurement method of host galaxy

The principle of host galaxy spectral measurement is the same as the normal galaxy. By comparing the observed lines with spectral lines in the laboratory, the spectroscopic redshift of GRB can be derived. GRB hosts are usually star-forming galaxies, and they show the strong emission lines (Jakobsson et al., 2012; Krühler et al., 2015), such as [O II] λ3727, $H\beta$ [OIII] λ5007 and $H\alpha$, in addition to weak emission lines [Ne III] [NIII], *etc.* Krühler et al. (2015) build a comprehensive sample of GRB host emission line spectrum. In the spectra, the superposition fitting of the Gaussian function to the data is effective to get the fluxes of strong emission lines. For weak emission lines, it is necessary to tie the Gaussian centroid and linewidth to those of strong lines. The method was applied in GRB 050714B ($z = 2.4383$) (Krühler et al., 2015),

2 The bands please see https://swift.gsfc.nasa.gov/about_swift/uvot_desc.html

3 g'r'i'z' filters please see https://classic.sdss.org/dr5/

GRB 060923B ($z = 1.5094$) (Krühler et al., 2015), GRB 080905A ($z = 0.1218$) (Rowlinson et al., 2010), GRB 101224A ($z = 0.4536$) (O'Connor et al., 2022), GRB 180805B ($z = 0.6609$) (O'Connor et al., 2022), GRB 211211A ($z = 0.0763$) (Rastinejad et al., 2022).

Spectral measurement method of host galaxy has high accuracy. GRB host has constant luminosity. Unlike the rapidly fading afterglow, they can be observed with a high-precision spectrometer for long exposure time. However, some host galaxies are faint, so that we cannot detected the spectral lines with high signal-to-noise ratio. Moreover, some GRB host galaxies at high redshift are not detected. Therefore, spectral measurement method of host galaxy is suitable for low redshift GRBs. We can also use this method to detect "Dark Burst" redshift. Some large telescopes, such as VLT/X-shooter, Keck, Magellan, are used for this method.

3.2.2 Automatic spectroscopic redshift measurement methods

Some GRB host galaxies are nearby galaxies. They are blue, low mass and low metallicity. Almost all the observation techniques and data processing techniques for studying other distant galaxies can be used to GRB host galaxies (Antonelli et al., 2000). At present, astronomers manually compare the observed spectral lines to the spectral lines in the laboratory. This way is inefficient.

By Automatic spectroscopic redshift measurement methods, we can automatically measure, identify and verify spectral lines, and use template matching method to derive redshift. There are many methods we can consider adopting. They are cross-correlation (Tonry and Davis, 1979), PCAZ (Glazebrook et al., 1998), pseudo-triangle (Qiu et al., 2002), wavelet technique to measure spectral lines and identify Balmer jump points (Luo and Zhao, 2001), density estimation (Duan et al., 2005), Hough transform (Zhou et al., 2000), Bayesian convolution network (Xu et al., 2006a), similarity measurement (Liu et al., 2008), nonlinear dimension reduction (Xu et al., 2006b), Parzen window (Tu et al., 2012), multi-resolution fusion distance method (Pan et al., 2016).

Compared with the number of normal galaxies, the number of GRB hosts is much less, and they are spread over the sky and randomly distributed. Therefore, it is hard to observe one GRB host from astronomy sky surveys. But automated algorithms still have some advantages. Automated algorithms may improve efficiency compared to traditional processing methods. It was presented that the automatic spectroscopic redshift measurement methods have high accuracy. The 97% goal can get the same result as manual processing (Zhu et al., 2005). These methods have fast processing speed, and can automatically process low signal-noise-ratio spectrum (Pan et al., 2016; Zhu et al., 2005). When GRB host is identified, the automatic algorithm can rapidly and accurately measure the redshift. Moreover, the automatic algorithm can quickly determine the redshift interval of each candidate and make a rough classification when the GRBs have multiple candidate hosts. Thus, one or more unqualified candidates can be quickly eliminated and the observation cost can be reduced.

3.2.3 Template matching method of host galaxy

Template matching method of host galaxy can measure the photometric redshift of target by SED fitting. Compared with GRB afterglow, the physical model of host galaxy is more complex, which should have considered isochron, star spectrum, initial mass function (IMF), star formation history, and chemical evolution (Johnson et al., 2021). At present, some popular codes for SED fitting are GALAXEV (Bruzual and Charlot, 2003), PGase (Maraston, 2005), FSPS (Conroy et al., 2009), BPASS (Eldridge et al., 2017). We can compare the template with the observation data. When the likelihood function is minimum, the photometric redshift is derived. Some Matching methods are Bayes (Salim et al., 2007), Hyperz (Bolzonella et al., 2000), EAZY (Acquaviva et al., 2015), CIGALE (Burgarella et al., 2005), MAGPHYS (Cunha et al., 2008), MCMC (Acquaviva et al., 2012), GalMC (Acquaviva et al., 2011), BEAGLE (Chevallard and Charlot, 2016), BAGPIPES (Carnall et al., 2018), LePHARE (Gupta et al., 2021), Prospector (Johnson et al., 2021). The method was applied in GRB 000214 ($z = 0.49^{+0.05}_{-0.07}$) (Guziy et al., 2005), GRB 191031D ($z = 0.5 \pm 0.2$) (O'Connor et al., 2022), GRB 200411A ($z = 0.6 \pm 0.1$) (O'Connor et al., 2022).

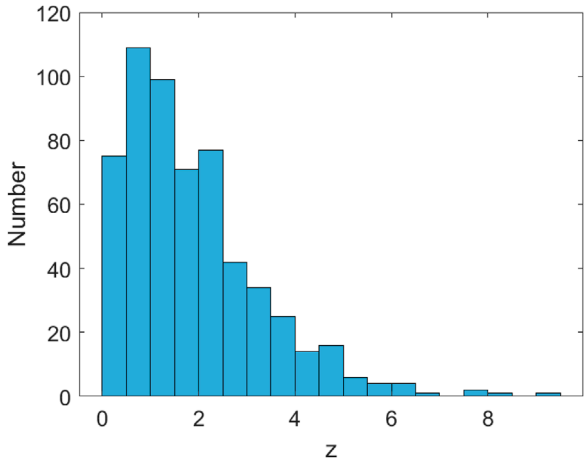


FIGURE 1
The distribution of GRB redshifts.

At present, there is no universal SED template due to the diverse features of galaxies (Gupta et al., 2022). Due to the uncertainty of photometric measurement, the accuracy of template matching method of host galaxy is relatively low. However, some GRB host galaxy are faint, so absorption lines and emission lines cannot be distinguished in some galaxy spectra. Photometric measurement is particularly useful for observing faint and relatively high redshift galaxies.

3.2.4 Machine learning

As long as the host galaxy can be detected, the redshift measurement methods of ordinary galaxies can be applied to the GRB afterglow or host galaxy. Besides template matching method, training set methods are also very common. Training set methods usually use machine learning technology to estimate redshift, which is based on statistical theory. These methods need a training data set, which contains samples of data or images with both spectroscopic and photometric redshifts. Redshift

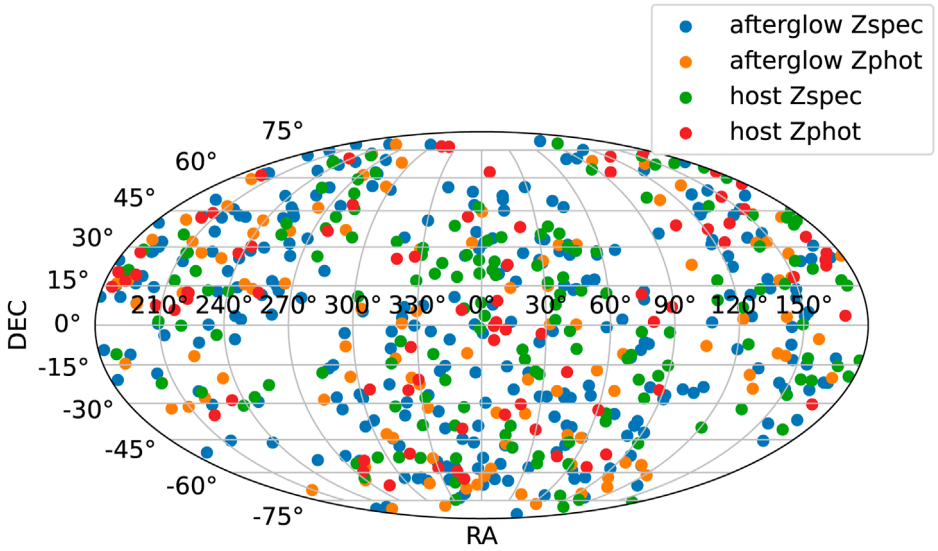


FIGURE 2
Distribution of GRB right ascension (RA) and declination (DEC).

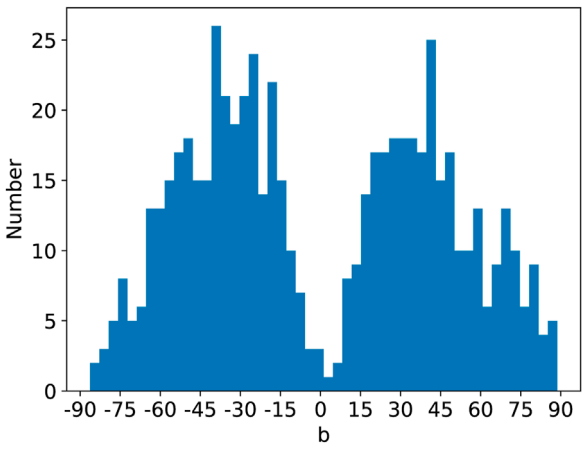


FIGURE 3
The distribution of GRB galactic latitude (b).

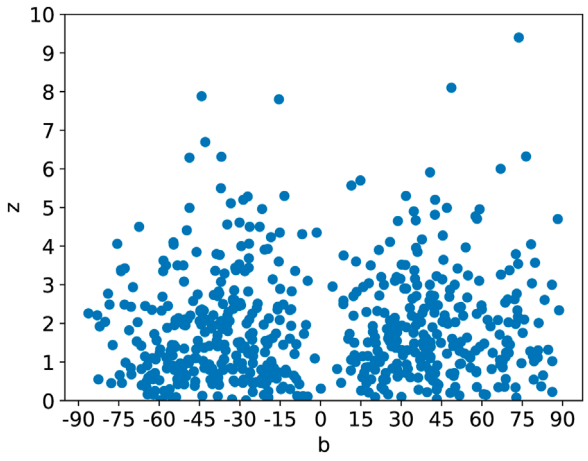


FIGURE 4
Correlation between galactic latitude (b) and redshift (z).

without spectroscopic measurement can be estimated by building a fitting model based on spectrum and photometric data in the training data set. Some typical methods are Polynomial Function Fitting (Connolly et al., 1995), Bayesian model (Benítez, 2000), Support Vector Machine (Wadadekar, 2005), K-Nearest Neighbor (Beck et al., 2016), Kernel Regression (Wang et al., 2008), Random Forest (Carrasco and Brunner, 2013), Active Learning (Han et al., 2015), Artificial Neural Network (Firth et al., 2003) and so on. The latest methods are Ensemble Learning (Cunha and Humphrey, 2022), Deep Capsule Network (Dey et al., 2022) and so on.

Compared to traditional processing methods, machine learning has the advantage of high efficiency. But the sample size of GRB hosts is far smaller than that of the normal galaxies. However, machine learning methods can still be useful in combination with other redshift estimation methods to improve the accuracy and speed of the analysis.

3.3 Pseudo-redshift estimation method

Besides the methods mentioned in section 3.1 and 3.2, pseudo-redshift can also be estimated according to the luminosity-indictor relationships, which contain some GRB observed features such as spectral peak energy, spectral complexity, light variation,

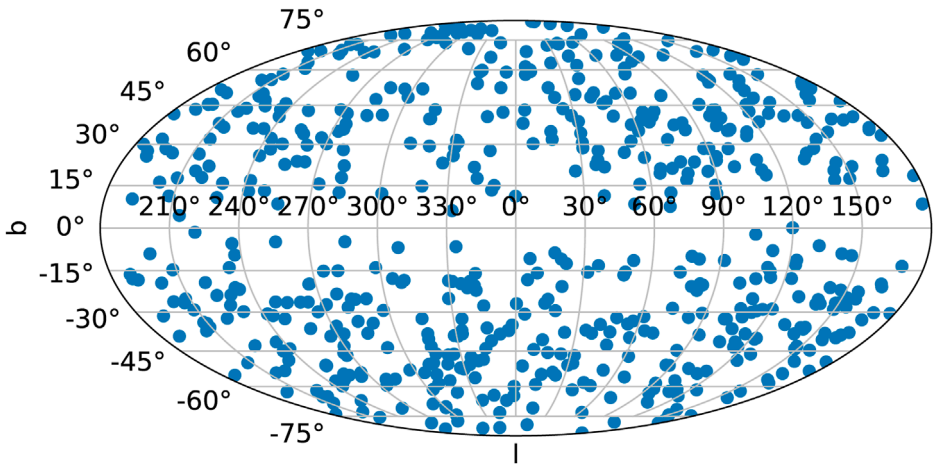


FIGURE 5
Distribution of GRB galactic longitude (l) and galactic latitude (b).

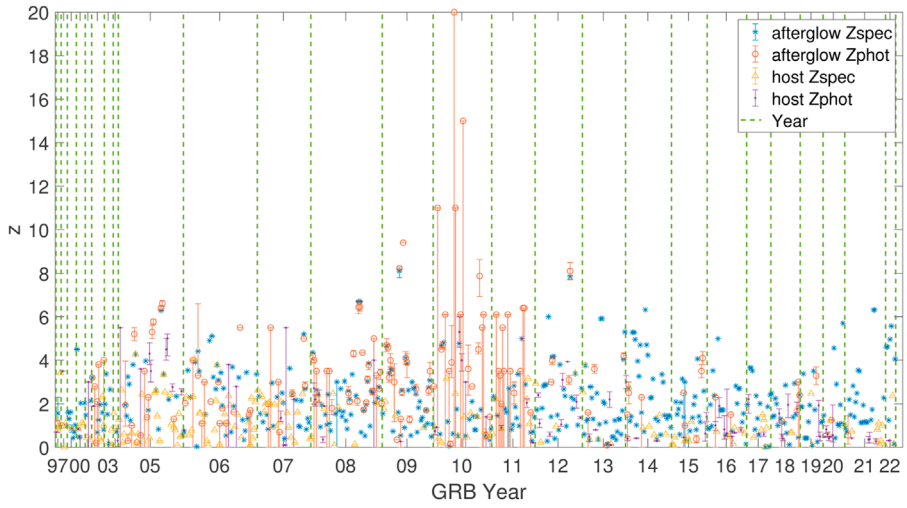


FIGURE 6
Comparison of different measurement methods of redshift. The horizontal axis represents the years of the GRBs, and the vertical axis represents the redshift. GRBs redshifts with known errors are described by error bars. For the GRBs with a certainly range limitation, the bottoms of error bars are connected to the zero-value point.

afterglow correlation, etc (Tan et al., 2013). This redshift estimation method was popularly used in the Compton BATSE era (Lloyd-Ronning et al., 2002). The advantage of this method is that it is completely independent of real redshift measurements in section 3.1 and 3.2, so Pseudo-redshift estimation cannot be affected by redshift selection effects such as redshift desert. Moreover, it can provide a larger amount of redshift data. However, the accuracy of the estimation depends on the tightness of the correlations between the observed features and redshift, and it may not be reliable for all GRBs (Guo, 2020). The method was applied in GRB 080514B ($z = 1.76 \pm 0.30$), and its photometric redshift is $z = 1.8^{+0.4}_{-0.3}$ (Rossi et al., 2008).

4 Statistical characteristics of GRB redshifts measured by different methods

The purpose of this paper is to explore various measurement methods of GRB redshift, so as to help astronomers choose accurate and effective measurement methods and apply them to future exploration. Therefore, in addition to the theoretical study of the algorithm, it is necessary to perform statistics on GRB redshift and show the applicable scope of these measurement methods in practical application.

We collect 611 GRBs with redshifts measured since 1997, as shown in Supplementary Table S2. The afterglow z_{spec} indicates

spectroscopic redshift measured by spectral measurement method of afterglow. The afterglow z_{phot} indicates photometric redshift measured by Afterglow SED fitting method. The host z_{spec} indicates measurement by spectral measurement method of host galaxy. The host z_{phot} indicates measurement by template matching method of host galaxy.

581 GRBs have accurately measured redshifts, with a range from 0.0085 to 9.4. 30 GRBs redshift only have an upper or lower limit. Some GRBs have been measured by several methods or the same method several times, and these GRBs have multiple redshift values. In a few cases, afterglow z_{phot} and host z_{phot} may have large differences in several measurements. One example is GRB 191031D. The photometric redshift methods select the redshift with the minimum χ^2. Although, χ^2 values of the afterglow z_{phot} measurement and host z_{phot} measurement may be similar, the redshift values of each result can have large differences.

The distribution of GRB redshifts is shown in Figure 1. The mean value is $\bar{z} = 1.8820$, the median value is $z_{med} = 1.5200$, and the standard deviation value is $s = 1.4148$.

As shown in Figure 2, GRB is randomly distributed on the celestial sphere. When $RA < 180°$, $-90° < DEC < 0°$, there are 151 GRBs; when $RA < 180°$, $0° < DEC < 90°$, there are 160 GRBs; when $RA > 180°$, $-90° < DEC < 0°$, there are 125 GRBs; when $RA > 180°$, $0° < DEC < 90°$, there are 175 GRBs. In addition, afterglow z_{spec}, afterglow z_{phot}, host z_{spec} and host z_{phot} GRBs are also randomly distributed on the celestial sphere respectly.

As shown in Figures 3, 4, GRBs are less distributed in the low galactic latitudes. It is very difficult to find GRBs in galactic disk, due to the serious dust extinction, reddening and dense star field in galactic disk. In the high galactic latitudes, there are few transients such as GRB. There is no obvious correlation between galactic latitude and redshift. GRB with the highest redshift (GRB 090429B) is $b = 73.72°$.

As shown in Figure 5, except for the low galactic latitude, GRB is uniform distributed on the galactic celestial sphere, which proves that GRB is isotropic.

Comparison of different measurement methods of redshift is shown in Figure 6. It can be seen that the redshift distribution measured in different years is similar. While the dark host galaxy and the detection limit of current equipment, it is difficult to detect the host galaxy with high redshift. With the increase of redshift, the errors of all methods are relatively increased. The accuracy of spectral measurement is usually higher than that of photometric measurement. At present, the highest redshift of GRB 090429B belongs to afterglow z_{phot}, as the afterglow is relatively bright. As GRB afterglow is relatively brighter than host. In order to improve the redshift measurement accuracy, we can combine above methods and take all the advantages of the various methods.

5 Summary and prospect

We introduce the research history and scientific significance of GRB redshift. Two types of measurement methods for GRB redshift are mainly summarized. The two types are the measurement methods based on afterglow and the measurement methods based on host galaxy. We have also outlined the principles, methods, and processes involved in GRB redshift measurement, and summarized the applications of different redshift measurement methods. Finally, we summarized and discussed the distribution of 611 GRBs with redshift measured by various methods.

SVOM mission is dedicated to GRB detection especially for high redshifts thanks to its gamma-ray detector payload, ECLAIRs, is more sensitive in softer the energy range of 4–150 keV (Wei et al., 2016). After ECLAIRs is triggered, SVOM can take follow-up observation quickly with its payload telescope Visible-band Telescope (VT), which can detect GRB at $z > 6.5$, and narrow-field ground-based robotic follow-up telescope (F-GFT and C-GFT) in multiple bands. The F-GFT will take broadband optical/infrared photometry with r and J simultaneous bands, operating in 0.4–1.7 μm. While C-GFT has SDSS g, r, i bands to operate photometry in 0.4–0.95 μm (Wei et al., 2016). The two ground telescopes, which have characteristics of high availability, can perform follow-up observations with, good sensitivity, fast pointing speed, low read-out time, high-precision localization and so on. GFTS can measure redshift from ~2.5 to ~6 with its multi-bands photometry if afterglow of GRB is detected. Lyman-α line is beyond the range of GFTs bands at $z < 2.5$, so the redshifts of these GRBs can be obtained by detecting host galaxies rather than afterglows. SVOM will also cooperate with facilities such as James Webb Space Telescope (JWST) and Extremely Large Telescopes to detect the host galaxies. To detect GRBs with high redshift, SVOM can raise optical-counterpart alerts quickly. Alerts can trigger large ground telescopes to observe afterglow and measure the spectroscopic redshift of GRB. In addition, SVOM has the anti-solar pointing strategy, so large telescopes can observe GRBs at early times, which is also helpful for the detection of GRBs with high redshifts (Wei et al., 2016). To sum up, several measurement methods of GRB redshift can be comprehensively applied to SVOM mission. For example, Wang et al. (2020) considered the polynomial fitting method to simulate the relationship between the B-R colors and spectroscopic redshift. We will consider to use machine learning methods to measure redshift in the future.

Author contributions

CL and CW are responsible for the construction of the overall framework of the article. JM is responsible for literature collection, data collection, theoretical guidance and writing guidance. ML is responsible for literature collection, data collection, experiment design, data analysis and writing. ZK is responsible for theoretical and writing guidance. ZL is responsible for literature collection and theoretical guidance. SD and BN are responsible for data collection. PJ is responsible for writing guidance. In addition, All authors contributed to the article and approved the submitted version.

Funding

This work is supported by the National Natural Science Foundation of China (Grant No. U2031129, U1931133, 12273080, 12203078, 11673062). This work is also supported by the SVOM project, a mission in the Strategic Priority Program on Space Science of CAS. This work is also supported by Yunnan Revitalization Talent Support Program (YunLin Scholar Award).

Acknowledgments

We appreciate all reviewers and editor for their careful review. The comments are very valuable for us to complete the major revision and improve the quality of the manuscript. We appreciate Prof. Jianyan Wei's useful discussion and helps on improving paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2023.1124317/full#supplementary-material

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Allende Prieto, C., An, D., et al. (2009). The seventh data release of the Sloan digital sky survey. *Astrophysical J. Suppl.* 182, 543–558. doi:10.1088/0067-0049/182/2/543

Acquaviva, V., Gawiser, E., and Guaita, L. (2012). SED fitting with MCMC: Methodology and application to large galaxy surveys. *Spectr. Energy Distribution Galaxies - SED 2011* 284, 42–45. doi:10.1017/S1743921312008691

Acquaviva, V., Gawiser, E., and Guaita, L. (2011). Spectral energy distribution fitting with Markov chain Monte Carlo: Methodology and application to z = 3.1 lyα-emitting galaxies. *Astrophysical J.* 737, 47. doi:10.1088/0004-637X/737/2/47

Acquaviva, V., Raichoor, A., and Gawiser, E. (2015). Simultaneous estimation of photometric redshifts and SED parameters: Improved techniques and a realistic error budget. *Astrophysical J.* 804, 8. doi:10.1088/0004-637X/804/1/8

Afonso, P., Greiner, J., Pian, E., Covino, S., Malesani, D., Küpcü Yoldaş, A., et al. (2011). GRB 050502B optical afterglow: A jet-break at high redshift. *Astronomy Astrophysics* 526, A154. doi:10.1051/0004-6361/200913965

Agüí Fernández, J. F., Thöne, C. C., Kann, D. A., de Ugarte Postigo, A., Selsing, J., Schady, P., et al. (2021). *Grb 160410A: The first chemical study of the interstellar medium of a short GRB. arXiv e*-prints , arXiv:2109.13838.

Amati, L., and Della Valle, M. (2013). Measuring cosmological parameters with gamma ray bursts. *Int. J. Mod. Phys. D* 22, 1330028. doi:10.1142/S0218271813300280

Amati, L., Guidorzi, C., Frontera, F., Della Valle, M., Finelli, F., Landi, R., et al. (2008). Measuring the cosmological parameters with the $E_{p,i}$-E_{iso} correlation of gamma-ray bursts. *Mon. Notices R. Astronomical Soc.* 391, 577–584. doi:10.1111/j.1365-2966.2008.13943.x

Andersen, M. I., Hjorth, J., Pedersen, H., Jensen, B. L., Hunt, L. K., Gorosabel, J., et al. (2000). VLT identification of the optical afterglow of the gamma-ray burst GRB 000131 at z=4.50. *Astronomy Astrophysics* 364, L54–L61. doi:10.48550/arXiv.astro-ph/0010322

Antonelli, L. A., Piro, L., Vietri, M., Costa, E., Soffitta, P., Feroci, M., et al. (2000). Discovery of a redshifted iron K line in the X-ray afterglow of GRB 000214. *Astrophysical J.* 545, L39–L42. doi:10.1086/317328

Atwood, W. B., Abdo, A. A., Ackermann, M., Althouse, W., Anderson, B., Axelsson, M., et al. (2009). The large area telescope on the Fermi gamma-ray space telescope mission. *Astrophysical J.* 697, 1071–1102. doi:10.1088/0004-637X/697/2/1071

Atwood, W. B., Baldini, L., Bregeon, J., Bruel, P., Chekhtman, A., Cohen-Tanugi, J., et al. (2013). New fermi-LAT event reconstruction reveals more high-energy gamma rays from gamma-ray bursts. *Astrophysical J.* 774, 76. doi:10.1088/0004-637X/774/1/76

Barthelmy, S. D., Chincarini, G., Burrows, D. N., Gehrels, N., Covino, S., Moretti, A., et al. (2005). An origin for short γ-ray bursts unassociated with current star formation. *Nature* 438, 994–996. doi:10.1038/nature04392

Basa, S., Cuby, J. G., Savaglio, S., Boissier, S., Clément, B., Flores, H., et al. (2012). Constraining the nature of the most distant gamma-ray burst host galaxies. *Astronomy Astrophysics* 542, A103. doi:10.1051/0004-6361/201218882

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., and Csabai, I. (2016). Photometric redshifts for the SDSS data release 12. *Mon. Notices R. Astronomical Soc.* 460, 1371–1381. doi:10.1093/mnras/stw1009

Bedient, J., Pearce, A., Otero, S., and Carvajal, J. (2001). V445 puppis. *International Astronomical Union Circular*, Available at: http://www.cbat.eps.harvard.edu/iauc/07500/07586.html.

Bellm, E. C., Hurley, K., Pal'shin, V., Yamaoka, K., Bandstra, M. S., Boggs, S. E., et al. (2008). Observations of the prompt gamma-ray emission of GRB 070125. *Astrophysical J.* 688, 491–498. doi:10.1086/592136

Benítez, N. (2000). Bayesian photometric redshift estimation. *Astrophysical J.* 536, 571–583. doi:10.1086/308947

Berger, E. (2010). A short gamma-ray burst "No-host" problem? Investigating large progenitor offsets for short GRBs with optical afterglows. *Astrophysical J.* 722, 1946–1961. doi:10.1088/0004-637X/722/2/1946

Berger, E., Fox, D. B., Kulkarni, S. R., Frail, D. A., and Djorgovski, S. G. (2007a). The ERO host galaxy of GRB 020127: Implications for the metallicity of GRB progenitors. *Astrophysical J.* 660, 504–508. doi:10.1086/513007

Berger, E., Fox, D. B., Price, P. A., Nakar, E., Gal-Yam, A., Holz, D. E., et al. (2007b). A new population of high-redshift short-duration gamma-ray bursts. *Astrophysical J.* 664, 1000–1010. doi:10.1086/518762

Berger, E., Kulkarni, S. R., Fox, D. B., Soderberg, A. M., Harrison, F. A., Nakar, E., et al. (2005). Afterglows, redshifts, and properties of*Swift*Gamma-ray bursts. *Astrophysical J.* 634, 501–508. doi:10.1086/491667

Berger, E., Penprase, B. E., Cenko, S. B., Kulkarni, S. R., Fox, D. B., Steidel, C. C., et al. (2006). Spectroscopy of GRB 050505 at z = 4.275: A logN(H I) = 22.1 DLA host galaxy and the nature of the progenitor. *Astrophysical J.* 642, 979–988. doi:10.1086/501162

Bloom, J. S., Djorgovski, S. G., Kulkarni, S. R., and Frail, D. A. (1998). The host galaxy of GRB 970508. *Astrophysical J.* 507, L25–L28. doi:10.1086/311682

Bloom, J. S., Djorgovski, S. G., and Kulkarni, S. R. (2001a). The redshift and the ordinary host galaxy of GRB 970228. *Astrophysical J.* 554, 678–683. doi:10.1086/321398

Bloom, J. S., Frail, D. A., and Sari, R. (2001b). The prompt energy release of gamma-ray bursts using a cosmological [ITAL]k[/ITAL]-Correction. *Astronomical J.* 121, 2879–2888. doi:10.1086/321093

Bloom, J. S., Kulkarni, S. R., Djorgovski, S. G., Eichelberger, A. C., Côté, P., Blakeslee, J. P., et al. (1999). The unusual afterglow of the γ-ray burst of 26 March 1998 as evidence for a supernova connection. *Nature* 401, 453–456. doi:10.1038/46744

Bloom, J. S., Kulkarni, S. R., and Djorgovski, S. G. (2002). The observed offset distribution of gamma-ray bursts from their host galaxies: A robust clue to the nature of the progenitors. *Astronomical J.* 123, 1111–1148. doi:10.1086/338893

Bloom, J. S., Perley, D. A., Chen, H. W., Butler, N., Prochaska, J. X., Kocevski, D., et al. (2007). A putative early-type host galaxy for GRB 060502B: Implications for the progenitors of short-duration hard-spectrum bursts. *Astrophysical J.* 654, 878–884. doi:10.1086/509114

Bolmer, J., Greiner, J., Krühler, T., Schady, P., Ledoux, C., Tanvir, N. R., et al. (2018). Dust reddening and extinction curves toward gamma-ray bursts at $z > 4$. *Astronomy Astrophysics* 609, A62. doi:10.1051/0004-6361/201731255

Bolzonella, M., Miralles, J. M., and Pelló, R. (2000). Photometric redshifts based on standard SED fitting procedures. *Astronomy Astrophysics* 363, 476–492. doi:10.48550/arXiv.astro-ph/0003380

Bruzual, G., and Charlot, S. (2003). Stellar population synthesis at the resolution of 2003. *Mon. Notices R. Astronomical Soc.* 344, 1000–1028. doi:10.1046/j.1365-8711.2003.06897.x

Burgarella, D., Buat, V., and Iglesias-Páramo, J. (2005). Star formation and dust attenuation properties in galaxies from a statistical ultraviolet-to-far-infrared analysis. *Mon. Notices R. Astronomical Soc.* 360, 1413–1425. doi:10.1111/j.1365-2966.2005.09131.x

Burrows, D. N., Hill, J. E., Nousek, J. A., Wells, A. A., Short, A. T., Ambrosi, R. M., et al. (2003). Swift x-ray telescope (XRT). *Soc. Photo-Optical Instrum. Eng. (SPIE) Conf. Ser.* 4851, 1320–1325. doi:10.1117/12.461279

Butler, N. R. (2007). On the early-time X-ray spectra of Swift afterglows. I. Evidence for anomalous soft X-ray emission. *Astrophysical J.* 656, 1001–1018. doi:10.1086/510548

Carnall, A. C., McLure, R. J., Dunlop, J. S., and Davé, R. (2018). Inferring the star formation histories of massive quiescent galaxies with BAGPIPES: Evidence for multiple quenching mechanisms. *Mon. Notices R. Astronomical Soc.* 480, 4379–4401. doi:10.1093/mnras/sty2169

Carrasco, K. M., and Brunner, R. J. (2013). Tpz: Photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Mon. Notices R. Astronomical Soc.* 432, 1483–1501. doi:10.1093/mnras/stt574

Castro-Tirado, A. J., Møller, P., García-Segura, G., Gorosabel, J., Pérez, E., de Ugarte Postigo, A., et al. (2010). Grb 021004: Tomography of a gamma-ray burst progenitor and its host galaxy. *Astronomy Astrophysics* 517, A61. doi:10.1051/0004-6361/200913966

Chen, H.-W. (2012). Near-infrared spectroscopy of gamma-ray burst host galaxies at: Insights into host galaxy dynamics and interpretations of afterglow absorption spectra: NIR spectra of two z ≥ 1.5 GRB hosts. *Mon. Notices R. Astronomical Soc.* 419, 3039–3047. doi:10.1111/j.1365-2966.2011.19944.x

Chevallard, J., and Charlot, S. (2016). Modelling and interpreting spectral energy distributions of galaxies with BEAGLE. *Mon. Notices R. Astronomical Soc.* 462, 1415–1443. doi:10.1093/mnras/stw1756

Chornock, R., Berger, E., Fox, D. B., Lunnan, R., Drout, M. R., Fong, W., et al. (2013). GRB 130606A as a probe of the intergalactic medium and the interstellar medium in a star-forming galaxy in the first gyr after the big bang. *Astrophysical J.* 774, 26. doi:10.1088/0004-637X/774/1/26

Ciardi, B., and Loeb, A. (2000). Expected number and flux distribution of gamma-ray burst afterglows with high redshifts. *Astrophysical J.* 540, 687–696. doi:10.1086/309384

Connolly, A. J., Csabai, I., Szalay, A. S., Koo, D. C., Kron, R. G., and Munn, J. A. (1995). Slicing through multicolor space: Galaxy redshifts from broadband photometry. *Astronomical J.* 110, 2655. doi:10.1086/117720

Conroy, C., Gunn, J. E., and White, M. (2009). The propagation of uncertainties in stellar population synthesis modeling. I. The relevance of uncertain aspects of stellar evolution and the initial mass function to the derived physical properties of galaxies. *Astrophysical J.* 699, 486–506. doi:10.1088/0004-637X/699/1/486

Conselice, C. J., Vreeswijk, P. M., Fruchter, A. S., Levan, A., Kouveliotou, C., Fynbo, J. P. U., et al. (2005). Gamma-ray burst-selected high-redshift galaxies: Comparison to field galaxy populations to z ˜3. *Astrophysical J.* 633, 29–40. doi:10.1086/432829

Costa, E., Frontera, F., Heise, J., Feroci, M., Zand, J., Fiore, F., et al. (1997). Discovery of an X-ray afterglow associated with the γ-ray burst of 28 February 1997. *Nature* 387, 783–785. doi:10.1038/42885

Coward, D. M., Guetta, D., Burman, R. R., and Imerito, A. (2008). Where are the missing gamma-ray burst redshifts? *Mon. Notices R. Astronomical Soc.* 386, 111–116. doi:10.1111/j.1365-2966.2008.13006.x

Coward, D. M., Howell, E. J., Branchesi, M., Stratta, G., Guetta, D., Gendre, B., et al. (2013). The Swift gamma-ray burst redshift distribution: Selection biases and optical brightness evolution at high z? *Mon. Notices R. Astronomical Soc.* 432, 2141–2149. doi:10.1093/mnras/stt537

Cucchiara, A., Levan, A. J., Fox, D. B., Tanvir, N. R., Ukwatta, T. N., Berger, E., et al. (2011). A photometric redshift of z ˜9.4 for GRB 090429B. *Astrophysical J.* 736, 7. doi:10.1088/0004-637X/736/1/7

Cucchiara, A., Prochaska, J. X., Perley, D., Cenko, S. B., Werk, J., Cardwell, A., et al. (2013). Gemini spectroscopy of the short-hard gamma-ray burst GRB 130603B afterglow and host galaxy. *Astrophysical J.* 777, 94. doi:10.1088/0004-637X/777/2/94

Cunha, E. D., Charlot, S., and Elbaz, D. (2008). A simple model to interpret the ultraviolet, optical and infrared emission from galaxies. *Mon. Notices R. Astronomical Soc.* 388, 1595–1617. doi:10.1111/j.1365-2966.2008.13535.x

Cunha, P. A. C., and Humphrey, A. (2022). Photometric redshift-aided classification using ensemble learning. *Astronomy Astrophysics* 666, A87. doi:10.1051/0004-6361/202243135

Curran, P. A., Wijers, R. A. M. J., Heemskerk, M. H. M., Starling, R. L. C., Wiersema, K., and van der Horst, A. J. (2008). Robust photometric redshift determinations of gamma-ray burst afterglows at z ≥ 2. *Astronomy Astrophysics* 490, 1047–1053. doi:10.1051/0004-6361:200810545

Dainotti, M. G., Young, S., Li, L., Levine, D., Kalinowski, K. K., Kann, D. A., et al. (2022). The optical two- and three-dimensional fundamental plane correlations for nearly 180 gamma-ray burst afterglows with Swift/UVOT, RATIR, and the subaru telescope. *Astrophysical J. Suppl. Ser.* 261, 25. doi:10.3847/1538-4365/ac7c64

D'Avanzo, P., Malesani, D., Covino, S., Piranomonte, S., Grazian, A., Fugazza, D., et al. (2009). The optical afterglows and host galaxies of three short/hard gamma-ray bursts. *Astronomy Astrophysics* 498, 711–721. doi:10.1051/0004-6361/200811294

Della Valle, M., Malesani, D., Bloom, J. S., Benetti, S., Chincarini, G., D'Avanzo, P., et al. (2006). Hypernova signatures in the late rebrightening of GRB 050525A. *Astrophysical J.* 642, L103–L106. doi:10.1086/504636

Dey, B., Andrews, B. H., Newman, J. A., Mao, Y.-Y., Rau, M. M., and Zhou, R. (2022). Photometric redshifts from SDSS images with an interpretable deep capsule network. *Mon. Notices R. Astronomical Soc.* 515, 5285–5305. doi:10.1093/mnras/stac2105

Dimple, K., MisraGhosh, A., Arun, K. G., Gupta, R., Kumar, A., Resmi, L., et al. (2022). Grb 210217A: A short or a long GRB? *Astrophysics Astronomy* 43, 39. doi:10.1007/s12036-022-09822-x

Djorgovski, S. G., Frail, D. A., Kulkarni, S. R., Bloom, J. S., Odewahn, S. C., and Diercks, A. (2001). The afterglow and the host galaxy of the dark burst GRB 970828. *Astrophysical J.* 562, 654–663. doi:10.1086/323845

Djorgovski, S. G., Kulkarni, S. R., Bloom, J. S., Goodrich, R., Frail, D. A., Piro, L., et al. (1998). Spectroscopy of the host galaxy of the gamma-ray burst 980703. *Astrophysical J.* 508, L17–L20. doi:10.1086/311729

Duan, F. Q., Wu, F. C., Luo, A. L., and Zhao, Y. H. (2005). Density estimation based model matching method for redshift determination. *Spectrosc. Spectr. Analysis* 2005, 1895–1898. doi:10.3321/j.issn:1000-0593.2005.11.042

Eldridge, J. J., Stanway, E. R., Xiao, L., McClelland, L. A. S., Taylor, G., Ng, M., et al. (2017). Binary population and spectral synthesis version 2.1: Construction, observational verification, and new results. *Astronomical Soc. Aust.* 34, e058. doi:10.1017/pasa.2017.51

Ferrero, P., Klose, S., Kann, D. A., Savaglio, S., Schulze, S., Palazzi, E., et al. (2009). Grb 060605: Multi-wavelength analysis of the first GRB observed using integral field spectroscopy. *Astronomy Astrophysics* 497, 729–741. doi:10.1051/0004-6361/200809980

Fiore, F., D'Elia, V., Lazzati, D., Perna, R., Sbordone, L., Stratta, G., et al. (2005). A flash in the dark: UVES very large telescope high-resolution spectroscopy of gamma-ray burst afterglows. *Astrophysical J.* 624, 853–867. doi:10.1086/429385

Fiore, F., Guetta, D., Piranomonte, S., D'Elia, V., and Antonelli, L. A. (2007). Selection effects shaping the gamma ray burst redshift distributions. *Astronomy Astrophysics* 470, 515–522. doi:10.1051/0004-6361:20077157

Firth, A. E., Lahav, O., and Somerville, R. S. (2003). Estimating photometric redshifts with artificial neural networks. *Mon. Notice R. Astronomical Soc.* 339, 1195–1202. doi:10.1046/j.1365-8711.2003.06271.x

Fong, W., Berger, E., Chornock, R., Margutti, R., Levan, A. J., Tanvir, N. R., et al. (2013). Demographics of the galaxies hosting short-duration gamma-ray bursts. *Astrophysical J.* 769, 56. doi:10.1088/0004-637X/769/1/56

Fong, W., Berger, E., Chornock, R., Tanvir, N. R., Levan, A. J., Fruchter, A. S., et al. (2011). The optical afterglow and z = 0.92 early-type host galaxy of the short GRB 100117A. *Astrophysical J.* 730, 26. doi:10.1088/0004-637X/730/1/26

Fong, W., Laskar, T., Rastinejad, J., Escorial, A. R., Schroeder, G., Barnes, J., et al. (2021). The broadband counterpart of the short GRB 200522A at z = 0.5536: A luminous kilonova or a collimated outflow with a reverse shock? *Astrophysical J.* 906, 127. doi:10.3847/1538-4357/abc74a

Fong, W., Margutti, R., Chornock, R., Berger, E., Shappee, B. J., Levan, A. J., et al. (2016). The afterglow and early-type host galaxy of the short GRB 150101B at z = 0.1343. *Astrophysical J.* 833, 151. doi:10.3847/1538-4357/833/2/151

Fong, W., Nugent, A. E., Dong, Y., Berger, E., Paterson, K., Chornock, R., et al. (2022). Short GRB host galaxies. I. Photometric and spectroscopic catalogs, host associations, and galactocentric offsets. *Astrophysical J.* 940, 56. doi:10.3847/1538-4357/ac91d0

Fruchter, A., Bergeron, L., and Pian, E. (1997). Grb 970508. *International Astronomical Union Circular*. Available at: http://www.cbat.eps.harvard.edu/iauc/06600/06674.html.

Fruchter, A. S., Levan, A. J., Strolger, L., Vreeswijk, P. M., Thorsett, S. E., Bersier, D., et al. (2006). Long γ-ray bursts and core-collapse supernovae have different environments. *Nature* 441, 463–468. doi:10.1038/nature04787

Fruchter, A. S., Pian, E., Thorsett, S. E., Bergeron, L. E., González, R. A., Metzger, M., et al. (1999). The fading optical counterpart of GRB 970228, 6 Months and 1 Year later. *Astrophysical J.* 516, 683–692. doi:10.1086/307136

Fynbo, J. P. U., Jakobsson, P., Prochaska, J. X., Malesani, D., Ledoux, C., de Ugarte Postigo, A., et al. (2009). Low-resolution spectroscopy of gamma-ray burst optical afterglows: Biases in the Swift sample and characterization of the absorbers. *Astrophysical J. Suppl.* 185, 526–573. doi:10.1088/0067-0049/185/2/526

Fynbo, J. P. U., Starling, R. L. C., Ledoux, C., Wiersema, K., Thöne, C. C., Sollerman, J., et al. (2006). Probing cosmic chemical evolution with gamma-ray bursts: GRB 060206 at z = 4.048. *Astronomy Astrophysics* 451, L47–L50. doi:10.1051/0004-6361:20065056

Galama, T., Groot, P. J., Vanparadijs, J., Kouveliotou, C., Robinson, C. R., Fishman, G. J., et al. (1997). The decay of optical emission from the γ-ray burst GRB 970228. *Nature* 387, 479–481. doi:10.1038/387479a0

Galama, T. J., Vreeswijk, P. M., Van Paradijs, J., Kouveliotou, C., Augusteijn, T., Böhnhardt, H., et al. (1998). An unusual supernova in the error box of the γ-ray burst of 25 April 1998. *Nature* 395, 670–672. doi:10.1038/27150

Gehrels, N., Chincarini, G., Giommi, P., Mason, K. O., Nousek, J. A., Wells, A. A., et al. (2004). The*Swift*Gamma-ray burst mission. *Astrophysical J.* 611, 1005–1020. doi:10.1086/422091

GLAST Facility Science Team, Gehrels, N., and Michelson, P. (1999). Glast: The next-generation high energy gamma-ray astronomy mission. *Astropart. Phys.* 11, 277–282. doi:10.1016/S0927-6505(99)00066-3

Glazebrook, K., Offer, A. R., and Deeley, K. (1998). Automatic redshift determination by use of principal component analysis. I. Fundamentals. *Astrophysical J.* 492, 98–109. doi:10.1086/305039

Goldstein, A., Hamburg, R., Wood, J., Hui, C. M., Cleveland, W. H., Kocevski, D., et al. (2019). *Updates to the Fermi GBM targeted sub-threshold Search in preparation for the third observing run of LIGO/virgo*. arXiv e-prints , arXiv:1903.12597.

Gorosabel, J., Christensen, L., Hjorth, J., Fynbo, J. U., Pedersen, H., Jensen, B. L., et al. (2003). A multi-colour study of the dark GRB 000210 host galaxy and its environment. *Astronomy Astrophysics* 400, 127–136. doi:10.1051/0004-6361:20021907

Gorosabel, J., Fynbo, J. P. U., Fruchter, A., Levan, A., Hjorth, J., Nugent, P., et al. (2005). A possible bright blue supernova in the afterglow of GRB 020305. *Astronomy Astrophysics* 437, 411–418. doi:10.1051/0004-6361:20052783

Götz, D., Covino, S., Hascoët, R., Fernandez-Soto, A., Daigne, F., Mochkovitch, R., et al. (2011). A detailed spectral study of GRB 041219A and its host galaxy. *Mon. Notices R. Astronomical Soc.* 413, 2173–2183. doi:10.1111/j.1365-2966.2011.18290.x

Gou, L. J., Mészáros, P., Abel, T., and Zhang, B. (2004). Detectability of long gamma-ray burst afterglows from very high redshifts. *Astrophysical J.* 604, 508–520. doi:10.1086/382061

Graham, J. F., Fruchter, A. S., Levan, A. J., Melandri, A., Kewley, L. J., Levesque, E. M., et al. (2009). Grb 070714B—discovery of the highest spectroscopically confirmed short burst redshift. *Astrophysical J.* 698, 1620–1629. doi:10.1088/0004-637X/698/2/1620

Greiner, J., Bornemann, W., Clemens, C., Deuter, M., Hasinger, G., Honsberg, M., et al. (2008). GROND—A 7-channel imager. *Astronomical Soc. Pac.* 120, 405–424. doi:10.1086/587032

Greiner, J., Fox, D. B., Schady, P., Krühler, T., Trenti, M., Cikota, A., et al. (2015). GAMMA-RAY BURSTS TRACE UV METRICS OF STAR FORMATION OVER 3 <*z*< 5. *Astrophysical J.* 809, 76. doi:10.1088/0004-637X/809/1/76

Greiner, J., Krühler, T., Fynbo, J. P. U., Rossi, A., Schwarz, R., Klose, S., et al. (2009a). Grb 080913 at redshift 6.7. *Astrophysical J.* 693, 1610–1620. doi:10.1088/0004-637X/693/2/1610

Greiner, J., Krühler, T., McBreen, S., Ajello, M., Giannios, D., Schwarz, R., et al. (2009b). A strong optical flare before the rising afterglow of GRB 080129. *Astrophysical J.* 693, 1912–1919. doi:10.1088/0004-637X/693/2/1912

Guo, Q. (2020). *Study on the jet structure and correlation of short gamma-ray bursts*. University of Science and Technology of China. doctor. doi:10.27517/d.cnki.gzkju.2020.001139

Gupta, R., Oates, S. R., Pandey, S. B., Castro-Tirado, A. J., Joshi, J. C., Hu, Y. D., et al. (2021). Grb 140102A: Insight into prompt spectral evolution and early optical afterglow emission. *Mon. Notices R. Astronomical Soc.* 505, 4086–4105. doi:10.1093/mnras/stab1573

Gupta, R., Pandey, S. B., Kumar, A., Aryan, A., Ror, A. K., Sharma, S., et al. (2022). Photometric studies on the host galaxies of gamma-ray bursts using 3.6m Devasthal optical telescope. *J. Astrophysics Astronomy* 43, 82. doi:10.1007/s12036-022-09865-0

Guziy, S., Gorosabel, J., Castro-Tirado, A. J., de Ugarte Postigo, A., Jelínek, M., Pérez Ramírez, M. D., et al. (2005). The search for the host galaxy of the gamma-ray burst GRB 000214. *Astronomy Astrophysics* 441, 975–979. doi:10.1051/0004-6361:20042419

Haislip, J. B., Nysewander, M. C., Reichart, D. E., Levan, A., Tanvir, N., Cenko, S. B., et al. (2006). A photometric redshift of z = 6.39 ± 0.12 for GRB 050904. *Nature* 440, 181–183. doi:10.1038/nature04552

Han, B., Zhang, Y., and Zhao, Y. (2015). "Active learning applied for photometric redshift estimation of quasars," in IAU 29th General Assembly in 2015. Available at: https://ui.adsabs.harvard.edu/abs/2015IAUGA.2256851H/abstract.

Heintz, K. E., Bolmer, J., Ledoux, C., Noterdaeme, P., Krogager, J. K., Fynbo, J. P. U., et al. (2019). New constraints on the physical conditions in H_2-bearing GRB-host damped Lyman-α absorbers. *Astronomy Astrophysics* 629, A131. doi:10.1051/0004-6361/201936250

Hjorth, J., Malesani, D., Jakobsson, P., Jaunsen, A. O., Fynbo, J. P. U., Gorosabel, J., et al. (2012). The optically unbiased gamma-ray burst host (TOUGH) survey. I. Survey design and catalogs. *Astrophysical J.* 756, 187. doi:10.1088/0004-637X/756/2/187

Hjorth, J., Møller, P., Gorosabel, J., Fynbo, J. P. U., Toft, S., Jaunsen, A. O., et al. (2003). Very high column density and small reddening toward GRB 020124 at z=3.20. *Astrophysical J.* 597, 699–705. doi:10.1086/378493

Hunt, L., Palazzi, E., Rossi, A., Savaglio, S., Cresci, G., Klose, S., et al. (2011). The extremely red host galaxy of GRB 080207. *Astrophysical J. Lett.* 736, L36. doi:10.1088/2041-8205/736/2/L36

Izzo, L., Muccino, M., Zaninoni, E., Amati, L., and Della Valle, M. (2015). New measurements of Ω_m from gamma-ray bursts. *Astronomy Astrophysics* 582, A115. doi:10.1051/0004-6361/201526461

Jakobsson, P., Fynbo, J. P. U., Ledoux, C., Vreeswijk, P., Kann, D. A., Hjorth, J., et al. (2006a). H I column densities of*z*> 2*Swift*gamma-ray bursts. *Astronomy Astrophysics* 460, L13–L17. doi:10.1051/0004-6361:20066405

Jakobsson, P., Hjorth, J., Malesani, D., Chapman, R., Fynbo, J. P. U., Tanvir, N. R., et al. (2012). The optically unbiased GRB host (TOUGH) survey. III. Redshift distribution. *Astrophysical J.* 752, 62. doi:10.1088/0004-637X/752/1/62

Jakobsson, P., Levan, A., Fynbo, J. P. U., Priddey, R., Hjorth, J., Tanvir, N., et al. (2006b). A mean redshift of 2.8 for**Swift**gamma-ray bursts. *Astronomy Astrophysics* 447, 897–903. doi:10.1051/0004-6361:20054287

Jakobsson, P., Levan, A., Fynbo, J. P. U., Priddey, R., Hjorth, J., Tanvir, N., et al. (2006c). GRB 050814 at z = 5.3 and the redshift distribution of Swift GRBs. *Am. Inst. Phys. Conf. Ser.* 836, 552–557. doi:10.1063/1.2207953

Japelj, J., Covino, S., Gomboc, A., Vergani, S. D., Goldoni, P., Selsing, J., et al. (2015). Spectrophotometric analysis of gamma-ray burst afterglow extinction curves with X-Shooter. *Astronomy Astrophysics* 579, A74. doi:10.1051/0004-6361/201525665

Jaunsen, A. O., Andersen, M. I., Hjorth, J., Fynbo, J. P. U., Holland, S. T., Thomsen, B., et al. (2003). An HST study of three very faint GRB host galaxies. *Astronomy Astrophysics* 402, 125–132. doi:10.1051/0004-6361:20030154

Jaunsen, A. O., Rol, E., Watson, D. J., Malesani, D., Fynbo, J. P. U., Milvang-Jensen, B., et al. (2008). Grb 070306: A highly extinguished afterglow. *Astrophysical J.* 681, 453–461. doi:10.1086/588602

Johnson, B. D., Leja, J., Conroy, C., and Speagle, J. S. (2021). Stellar population inference with prospector. *Astrophysical J. Suppl. Ser.* 254, 22. doi:10.3847/1538-4365/abef67

Kelly, P. L., Kirshner, R. P., and Pahre, M. (2008). Long γ-ray bursts and type ic core-collapse supernovae have similar locations in hosts. *Astrophysical J.* 687, 1201–1207. doi:10.1086/591925

Klose, S., Schmidl, S., Kann, D. A., Nicuesa Guelbenzu, A., Schulze, S., Greiner, J., et al. (2019). Four GRB supernovae at redshifts between 0.4 and 0.8 the bursts GRB 071112C, 111228A, 120714B, and 130831A. *Astronomy Astrophysics* 622, A138. doi:10.1051/0004-6361/201832728

Krühler, T., Greiner, J., Schady, P., Savaglio, S., Afonso, P. M. J., Clemens, C., et al. (2011a). The SEDs and host galaxies of the dustiest GRB afterglows. *Astronomy Astrophysics* 534, A108. doi:10.1051/0004-6361/201117428

Krühler, T., Küpcü Yoldaş, A., Greiner, J., Clemens, C., McBreen, S., Primak, N., et al. (2008). The 2175 Å dust feature in a gamma-ray burst afterglow at redshift 2.45. *Astrophysical J.* 685, 376–383. doi:10.1086/590240

Krühler, T., Malesani, D., Fynbo, J. P. U., Hartoog, O. E., Hjorth, J., Jakobsson, P., et al. (2015). GRB hosts through cosmic time. VLT/X-Shooter emission-line spectroscopy of 96 γ-ray-burst-selected galaxies at 0.1 < z < 3.6. *Astronomy Astrophysics* 581, A125. doi:10.1051/0004-6361/201425561

Krühler, T., Malesani, D., Milvang-Jensen, B., Fynbo, J. P. U., Hjorth, J., Jakobsson, P., et al. (2012). The optically unbiased GRB host (TOUGH) survey. V. VLT/X-shooter emission-line redshifts for Swift GRBs at z ˜2. *Astrophysical J.* 758, 46. doi:10.1088/0004-637X/758/1/46

Krühler, T., Schady, P., Greiner, J., Afonso, P., Bottacini, E., Clemens, C., et al. (2011b). Photometric redshifts for gamma-ray burst afterglows from GROND and Swift/UVOT. *Astronomy Astrophysics* 526, A153. doi:10.1051/0004-6361/201015327

Kuin, N. P. M., Landsman, W., Page, M. J., Schady, P., Still, M., Breeveld, A. A., et al. (2009). Grb 081203A: Swift UVOT captures the earliest ultraviolet spectrum of a gamma-ray burst. *Mon. Notices R. Astronomical Soc.* 395, L21–L24. doi:10.1111/j.1745-3933.2009.00632.x

Kulkarni, S. R., Djorgovski, S. G., Ramaprakash, A. N., Goodrich, R., Bloom, J. S., Adelberger, K. L., et al. (1998). Identification of a host galaxy at redshift z = 3.42 for the γ-ray burst of 14 December 1997. *Nature* 393, 35–39. doi:10.1038/29927

Laskar, T., Berger, E., Tanvir, N., Zauderer, B. A., Margutti, R., Levan, A., et al. (2014). GRB 120521C at z ˜6 and the properties of high-redshift γ-ray bursts. *Astrophysical J.* 781, 1. doi:10.1088/0004-637X/781/1/1

Le Floc'h, E., Duc, P. A., Mirabel, I. F., Sanders, D. B., Bosch, G., Diaz, R. J., et al. (2003). Are the hosts of gamma-ray bursts sub-luminous and blue galaxies? *Astronomy Astrophysics* 400, 499–510. doi:10.1051/0004-6361:20030001

Le Floc'h, E., Duc, P. A., Mirabel, I. F., Sanders, D. B., Bosch, G., Rodrigues, I., et al. (2002). Very large telescope and [ITAL]Hubble space telescope[/ITAL] observations of the host galaxy of GRB 990705. *Astrophysical J.* 581, L81–L84. doi:10.1086/346072

Leibler, C. N., and Berger, E. (2010). The stellar ages and masses of short gamma-ray burst host galaxies: Investigating the progenitor delay time distribution and the role of mass and star formation in the short gamma-ray burst rate. *Astrophysical J.* 725, 1202–1214. doi:10.1088/0004-637X/725/1/1202

Levan, A. J., Tanvir, N. R., Starling, R. L. C., Wiersema, K., Page, K. L., Perley, D. A., et al. (2014). A new population of ultra-long duration gamma-ray bursts. *Astrophysical J.* 781, 13. doi:10.1088/0004-637X/781/1/13

Li, L., Liang, E. W., Tang, Q. W., Chen, J. M., Xi, S. Q., Lü, H. J., et al. (2012). A comprehensive study of gamma-ray burst optical emission. I. Flares and early shallow-decay component. *Astrophysical J.* 758, 27. doi:10.1088/0004-637X/758/1/27

Li, L., Wang, Y., Shao, L., Wu, X. F., Huang, Y. F., Zhang, B., et al. (2018). A large catalog of multiwavelength GRB afterglows. I. Color evolution and its physical implication. *Astrophysical J. Suppl. Ser.* 234, 26. doi:10.3847/1538-4365/aaa02a

Li, L., Wu, X. F., Huang, Y. F., Wang, X. G., Tang, Q. W., Liang, Y. F., et al. (2015). A correlated study of optical and X-ray afterglows of GRBs. *Astrophysical J.* 805, 13. doi:10.1088/0004-637X/805/1/13

Lin, Y. Q. (2009). Research paper: Comparison between Swift and pre-Swift gamma-ray bursts. *Res. Astronomy Astrophysics* 9, 682–686. doi:10.1088/1674-4527/9/6/006

Littlejohns, O. M., Butler, N. R., Cucchiara, A., Watson, A. M., Kutyrev, A. S., Lee, W. H., et al. (2014). Identifying high-redshift gamma-ray bursts with RATIR. *Astronomical J.* 148, 2. doi:10.1088/0004-6256/148/1/2

Liu, R., Qiao, X. J., and Duan, F. Q. (2008). Redshift estimation of galaxy spectra based on similarity measure. *Spectrosc. Spectr. Analysis* 28, 235–238. doi:10.3964/j.issn.1000-0593.2008.01.057

Lloyd-Ronning, N. M., Fryer, C. L., and Ramirez-Ruiz, E. (2002). Cosmological aspects of gamma-ray bursts: Luminosity evolution and an estimate of the star formation rate at high redshifts. *Astrophysical J.* 574, 554–565. doi:10.1086/341059

Luo, A. L., and Zhao, Y. H. (2001). Steps towards a fully automated classification and redshift-measurement pipeline for LAMOST spectra. I. Continuum level and wavelength estimation for galaxies. *Chin. J. Astronomy Astrophysics* 1, 563–572. doi:10.1088/1009-9271/1/6/563

Lyman, J. D., Levan, A. J., Tanvir, N. R., Fynbo, J. P. U., McGuire, J. T. W., Perley, D. A., et al. (2017). The host galaxies and explosion sites of long-duration gamma ray bursts: Hubble Space Telescope near-infrared imaging. *Mon. Notices R. Astronomical Soc.* 467, stx220–1817. doi:10.1093/mnras/stx220

Maraston, C. (2005). Evolutionary population synthesis: Models, analysis of the ingredients and application to high-z galaxies. *Mon. Notices R. Astronomical Soc.* 362, 799–825. doi:10.1111/j.1365-2966.2005.09270.x

Melandri, A., Malesani, D. B., Izzo, L., Japelj, J., Vergani, S. D., Schady, P., et al. (2019). GRB 171010A/SN 2017htp: A GRB-SN at z = 0.33. *Mon. Notices R. Astronomical Soc.* 490, 5366–5374. doi:10.1093/mnras/stz2900

Metzger, M. R., Cohen, J. G., Chaffee, F. H., and Blandford, R. D. (1997a). Grb 970508. *International Astronomical Union Circular*. Available at: http://www.cbat.eps.harvard.edu/iauc/06600/06676.html.

Metzger, M. R., Djorgovski, S. G., Kulkarni, S. R., Steidel, C. C., Adelberger, K. L., Frail, D. A., et al. (1997b). Spectral constraints on the redshift of the optical counterpart to the γ-ray burst of 8 May 1997. *Nature* 387, 878–880. doi:10.1038/43132

Mirabel, I. F., and Rodrigues, I. (2003). Formation of a black hole in the dark. *Science* 300, 1119–1120. doi:10.1126/science.1083451

Modjaz, M., Kewley, L., Kirshner, R. P., Stanek, K. Z., Challis, P., Garnavich, P. M., et al. (2008). Measured metallicities at the sites of nearby broad-lined type ic supernovae and implications for the supernovae gamma-ray burst connection. *Astronomical J.* 135, 1136–1150. doi:10.1088/0004-6256/135/4/1136

Moresco, M., Amati, L., Amendola, L., Birrer, S., Blakeslee, J. P., Cantiello, M., et al. (2022). Unveiling the Universe with emerging cosmological probes. *Living Rev. Relativ.* 25, 6. doi:10.1007/s41114-022-00040-z

Muccino, M., Izzo, L., Luongo, O., Boshkayev, K., Amati, L., Della Valle, M., et al. (2021). Tracing dark energy history with gamma-ray bursts. *Astrophysical J.* 908, 181. doi:10.3847/1538-4357/abd254

Nakagawa, Y. E., Yoshida, A., Sugita, S., Tanaka, K., Ishikawa, N., Tamagawa, T., et al. (2006). An optically dark GRB observed by HETE-2: Grb 051022. *Publ. Astronomical Soc. Jpn.* 58, L35–L39. doi:10.1093/pasj/58.4.L35

Nugent, A. E., Fong, W.-F., Dong, Y., Leja, J., Berger, E., Zevin, M., et al. (2022). Short GRB host galaxies. II. A legacy sample of redshifts, stellar population properties, and implications for their neutron star merger origins. *Astrophysical J.* 940, 57. doi:10.3847/1538-4357/ac91d1

Nugent, A. E., Fong, W., Dong, Y., Palmese, A., Leja, J., Escorial, A. R., et al. (2020). The distant, galaxy cluster environment of the short GRB 161104A at z ~ 0.8 and a comparison to the short GRB host population. *Astrophysical J.* 904, 52. doi:10.3847/1538-4357/abc24a

Oates, S. R., and Cummings, J. R. (2009). GRB090426 Swift/UVOT observations. *GRB Coord. Netw.* 9265, 1.

O'Connor, B., Troja, E., Dichiara, S., Beniamini, P., Cenko, S. B., Kouveliotou, C., et al. (2022). A deep survey of short GRB host galaxies over z 0-2: Implications for offsets, redshifts, and environments. *Mon. Notices R. Astronomical Soc.* 515, 4890–4928. doi:10.1093/mnras/stac1982

Östlin, G., Zackrisson, E., Sollerman, J., Mattila, S., and Hayes, M. (2008). Constraining the mass of the GRB 030329 progenitor. *Mon. Notices R. Astronomical Soc.* 387, 1227–1236. doi:10.1111/j.1365-2966.2008.13319.x

Pan, J. C., Luo, A. L., Wei, P., Jiang, B., Li, Y. B., and Zheng, Q. (2016). A new redshift measuring method for low-quality galaxy spectra based on multi resolution fusion distance. *Spectrosc. Spectr. Analysis* 36, 1521–1525. doi:10.3964/j.issn.1000-0593(2016)05-1521-05

Paterson, K., Fong, W., Nugent, A., Escorial, A. R., Leja, J., Laskar, T., et al. (2020). Discovery of the optical afterglow and host galaxy of short GRB 181123B at z = 1.754: Implications for delay time distributions. *Astrophysical J. Lett.* 898, L32. doi:10.3847/2041-8213/aba4b0

Penprase, B. E., Berger, E., Fox, D. B., Kulkarni, S. R., Kadish, S., Kerber, L., et al. (2006). Spectroscopy of GRB 051111 at z = 1.54948: Kinematics and elemental abundances of the GRB environment and host galaxy. *Astrophysical J.* 646, 358–368. doi:10.1086/504678

Perley, D. A., Bloom, J. S., Butler, N. R., Pollack, L. K., Holtzman, J., Blake, C. H., et al. (2008a). The troublesome broadband evolution of GRB 061126: Does a gray burst imply gray dust? *Astrophysical J.* 672, 449–464. doi:10.1086/523929

Perley, D. A., Bloom, J. S., Klein, C. R., Covino, S., Minezaki, T., Woźniak, P., et al. (2010). Evidence for supernova-synthesized dust from the rising afterglow of GRB 071025 at z ~ 5. *Mon. Notices R. Astronomical Soc.* 406, 2473–2487. doi:10.1111/j.1365-2966.2010.16772.x

Perley, D. A., Cenko, S. B., Bloom, J. S., Chen, H. W., Butler, N. R., Kocevski, D., et al. (2009). The host galaxies of Swift dark gamma-ray bursts: Observational constraints on highly obscured and very high redshift GRBs. *Astronomical J.* 138, 1690–1708. doi:10.1088/0004-6256/138/6/1690

Perley, D. A., Krühler, T., Schady, P., Michałowski, M. J., Thöne, C. C., Petry, D., et al. (2017). A revised host galaxy association for GRB 020819B: A high-redshift dusty starburst, not a low-redshift gas-poor spiral. *Mon. Notices R. Astronomical Soc. Lett.* 465, L89–L93. doi:10.1093/mnrasl/slw221

Perley, D. A., Krühler, T., Schulze, S., de Ugarte Postigo, A., Hjorth, J., Berger, E., et al. (2016a). THE*SWIFT*GAMMA-RAY burst host galaxy legacy survey. I. Sample selection and redshift distribution. *Astrophysical J.* 817, 7. doi:10.3847/0004-637X/817/1/7

Perley, D. A., Levan, A. J., Tanvir, N. R., Cenko, S. B., Bloom, J. S., Hjorth, J., et al. (2013). A population of massive, luminous galaxies hosting heavily dust-obscured gamma-ray bursts: Implications for the use of GRBs as tracers of cosmic star formation. *Astrophysical J.* 778, 128. doi:10.1088/0004-637X/778/2/128

Perley, D. A., Li, W., Chornock, R., Prochaska, J. X., Butler, N. R., Chandra, P., et al. (2008b). Grb 071003: Broadband follow-up observations of a very bright gamma-ray burst in a galactic halo. *Astrophysical J.* 688, 470–490. doi:10.1086/591961

Perley, D. A., Modjaz, M., Morgan, A. N., Cenko, S. B., Bloom, J. S., Butler, N. R., et al. (2012). The luminous infrared host galaxy of short-duration GRB 100206A. *Astrophysical J.* 758, 122. doi:10.1088/0004-637X/758/2/122

Perley, D. A., and Perley, R. A. (2013). Radio constraints on heavily obscured star formation within dark gamma-ray burst host galaxies. *Astrophysical J.* 778, 172. doi:10.1088/0004-637X/778/2/172

Perley, D. A., Tanvir, N. R., Hjorth, J., Laskar, T., Berger, E., Chary, R., et al. (2016b). The Swift GRB host galaxy legacy survey. II. Rest-Frame near-IR luminosity distribution and evidence for a near-solar metallicity threshold. *Astrophysical J.* 817, 8. doi:10.3847/0004-637X/817/1/8

Piro, L., Frail, D. A., Gorosabel, J., Garmire, G., Soffitta, P., Amati, L., et al. (2002). The bright gamma-ray burst of 2000 february 10: A case study of an optically dark gamma-ray burst. *Astrophysical J.* 577, 680–690. doi:10.1086/342226

Pisani, G. B., Ruffini, R., Aimuratov, Y., Bianco, C. L., Karlica, M., Kovacevic, M., et al. (2018). The first ICRANet catalog of binary-driven hypernovae. *Eur. Phys. J. Web Conf.* 168, 04002. doi:10.1051/epjconf/201816804002

Price, P. A., Berger, E., Kulkarni, S. R., Djorgovski, S. G., Fox, D. W., Mahabal, A., et al. (2002). The unusually long duration gamma-ray burst GRB 000911: Discovery of the afterglow and host galaxy. *Astrophysical J.* 573, 85–91. doi:10.1086/340585

Qiu, B., Hu, Z. Y., and Zhao, Y. H. (2002). A pseudo triangle technique for redshift identification of celestial spectra. *Spectrosc. Spectr. Analysis* 22, 695–698. doi:10.3321/j.issn:1000-0593.2002.04.051

Rastinejad, J. C., Gompertz, B. P., Levan, A. J., Fong, W., Nicholl, M., Lamb, G. P., et al. (2022). *A kilonova following a long-duration gamma-ray burst at 350 mc. arXiv e-prints* , arXiv:2204.10864.

Rau, A., Salvato, M., and Greiner, J. (2005). The host of GRB/XRF 030528 - an actively star forming galaxy at z = 0.782. *Astronomy Astrophysics* 444, 425–430. doi:10.1051/0004-6361:20053773

Reichart, D. E. (1998). The redshift of GRB 970508. *Astrophysical J.* 495, L99–L101. doi:10.1086/311222

Rodney, S. A., Riess, A. G., Scolnic, D. M., Jones, D. O., Hemmati, S., Molino, A., et al. (2015). Two SNe ia at redshift ~2: Improved classification and redshift determination with medium-band infrared imaging. *Astronomical J.* 150, 156. doi:10.1088/0004-6256/150/5/156

Roming, P. W. A., Kennedy, T. E., Mason, K. O., Nousek, J. A., Ahr, L., Bingham, R. E., et al. (2005). The Swift ultra-violet/optical telescope. *Space Sci. Rev.* 120, 95–142. doi:10.1007/s11214-005-5095-4

Roming, P. W. A., Schady, P., Fox, D. B., Zhang, B., Liang, E., Mason, K. O., et al. (2006a). Very early optical afterglows of gamma-ray bursts: Evidence for relative paucity of detection. *Astrophysical J.* 652, 1416–1422. doi:10.1086/508481

Roming, P. W. A., Vanden Berk, D., Pal'shin, V., Pagani, C., Norris, J., Kumar, P., et al. (2006b). Grb 060313: A new paradigm for short-hard bursts? *Astrophysical J.* 651, 985–993. doi:10.1086/508054

Rossi, A., de Ugarte Postigo, A., Ferrero, P., Kann, D. A., Klose, S., Schulze, S., et al. (2008). A photometric redshift of $z = 1.8^{+0.4}_{-0.3}$ for the AGILE GRB 080514B. *Astronomy Astrophysics* 491, L29–L32. doi:10.1051/0004-6361:200810736

Rossi, A., Frederiks, D. D., Kann, D. A., De Pasquale, M., Pian, E., Lamb, G., et al. (2022). A blast from the infant Universe: The very high-z GRB 210905A. *Astronomy Astrophysics* 665, A125. doi:10.1051/0004-6361/202243225

Rossi, A., Piranomonte, S., Savaglio, S., Palazzi, E., Michałowski, M. J., Klose, S., et al. (2014). A quiescent galaxy at the position of the long GRB 050219A. *Astronomy Astrophysics* 572, A47. doi:10.1051/0004-6361/201423865

Rowlinson, A., Wiersema, K., Levan, A. J., Tanvir, N. R., O'Brien, P. T., Rol, E., et al. (2010). Discovery of the afterglow and host galaxy of the low-redshift short GRB 080905A. *Mon. Notices R. Astronomical Soc.* 408, 383–391. doi:10.1111/j.1365-2966.2010.17115.x

Saccardi, A., Vergani, S. D., De Cia, A., D'Elia, V., Heintz, K. E., Izzo, L., et al. (2023). Dissecting the interstellar medium of a z = 6.3 galaxy. X-shooter spectroscopy and HST imaging of the afterglow and environment of the Swift GRB 210905A. *Astronomy Astrophysics* 671, A84. doi:10.1051/0004-6361/202244205

Sahu, K. C., Livio, M., Petro, L., Bond, H. E., Macchetto, F. D., Galama, T. J., et al. (1997). Observations of GRB 970228 and GRB 970508 and the neutron star merger model. *Astrophysical J. Lett.* 489, L127–L131. doi:10.1086/316786

Salim, S., Rich, R. M., Charlot, S., Brinchmann, J., Johnson, B. D., Schiminovich, D., et al. (2007). UV star formation rates in the local universe. *Astrophysical J. Suppl. Ser.* 173, 267–292. doi:10.1086/519218

Salvaterra, R., Campana, S., Vergani, S. D., Covino, S., D'Avanzo, P., Fugazza, D., et al. (2012). A complete sample of bright Swift long gamma-ray bursts. I. Sample presentation, luminosity function and evolution. *Astrophysical J.* 749, 68. doi:10.1088/0004-637X/749/1/68

Salvaterra, R., Della Valle, M., Campana, S., Chincarini, G., Covino, S., D'Avanzo, P., et al. (2009). GRB090423 at a redshift of z˜8.1. *Nature* 461, 1258–1260. doi:10.1038/nature08445

Salvaterra, R. (2015). High redshift gamma-ray bursts. *J. High Energy Astrophysics* 7, 35–43. doi:10.1016/j.jheap.2015.03.001

Sánchez-Ramírez, R., Gorosabel, J., de Ugarte Postigo, A., Castro-Tirado, A. J., Thöne, C. C., Fynbo, J. P. U., et al. (2013). "Grb 100316A: An explosion at a high redshift galaxy," in *Highlights of Spanish astrophysics VII*. Editors J. C. Guirado, L. M. Lara, V. Quilis, and J. Gorgas, 463.

Savaglio, S., Glazebrook, K., and Le Borgne, D. (2009). The galaxy population hosting gamma-ray bursts. *Astrophysical J.* 691, 182–211. doi:10.1088/0004-637X/691/1/182

Schulze, S., Chapman, R., Hjorth, J., Levan, A. J., Jakobsson, P., Björnsson, G., et al. (2015). The optically unbiased GRB host (TOUGH) survey. VII. The host galaxy luminosity function: Probing the relationship between GRBs and star formation to redshift ~ 6. *Astrophysical J.* 808, 73. doi:10.1088/0004-637X/808/1/73

Selsing, J., Krühler, T., Malesani, D., D'Avanzo, P., Schulze, S., Vergani, S. D., et al. (2018). The host galaxy of the short GRB 111117A at z = 2.211. Impact on the short GRB redshift distribution and progenitor channels. *Astronomy Astrophysics* 616, A48. doi:10.1051/0004-6361/201731475

Selsing, J., Malesani, D., Goldoni, P., Fynbo, J. P. U., Krühler, T., Antonelli, L. A., et al. (2019). The X-shooter GRB afterglow legacy sample (XS-GRB). *Astronomy Astrophysics* 623, A92. doi:10.1051/0004-6361/201832835

Si, S. K., Qi, Y. Q., Xue, F. X., Liu, Y. J., Wu, X., Yi, S. X., et al. (2018). The three-parameter correlations about the optical plateaus of gamma-ray bursts. *Astrophysical J.* 863, 50. doi:10.3847/1538-4357/aad08a

Skrutskie, M. F., Cutri, R. M., Stiening, R., Weinberg, M. D., Schneider, S., Carpenter, J. M., et al. (2006). The two micron all sky survey (2MASS). *Astronomical J.* 131, 1163–1183. doi:10.1086/498708

Smith, K. L., Ridden-Harper, R., Fausnaugh, M., Daylan, T., Omodei, N., Racusin, J., et al. (2021). Grb 191016A: A long gamma-ray burst detected by tess. *Astrophysical J.* 911, 43. doi:10.3847/1538-4357/abe6a2

Soderberg, A. M., Berger, E., Kasliwal, M., Frail, D. A., Price, P. A., Schmidt, B. P., et al. (2006). The afterglow, energetics, and host galaxy of the short-hard gamma-ray burst 051221a. *Astrophysical J.* 650, 261–271. doi:10.1086/506429

Stanway, E. R., Levan, A. J., Tanvir, N., Wiersema, K., van der Horst, A., Mundell, C. G., et al. (2015). Grb 080517: A local, low-luminosity gamma-ray burst in a dusty galaxy at z = 0.09. *Mon. Notices R. Astronomical Soc.* 446, 3911–3925. doi:10.1093/mnras/stu2286

Stratta, G., Basa, S., Butler, N., Atteia, J. L., Gendre, B., Pélangeon, A., et al. (2007). X-ray flashes or soft gamma-ray bursts?. The case of the likely distant XRF 040912. *Astronomy Astrophysics* 461, 485–492. doi:10.1051/0004-6361:20065831

Svensson, K. M., Levan, A. J., Tanvir, N. R., Fruchter, A. S., and Strolger, L. G. (2010). The host galaxies of core-collapse supernovae and gamma-ray bursts. *Mon. Notices R. Astronomical Soc.* 405, 57–76. doi:10.1111/j.1365-2966.2010.16442.x

Tan, W. W., Cao, X. F., and Yu, Y. W. (2013). Determining the luminosity function of Swift long gamma-ray bursts with pseudo-redshifts. *Astrophysical J. Lett.* 772, L8. doi:10.1088/2041-8205/772/1/L8

Tanvir, N. R., Fox, D. B., Levan, A. J., Berger, E., Wiersema, K., Fynbo, J. P. U., et al. (2009). A γ-ray burst at a redshift of z˜8.2. *Nature* 461, 1254–1257. doi:10.1038/nature08459

Tanvir, N. R., Fynbo, J. P. U., de Ugarte Postigo, A., Japelj, J., Wiersema, K., Malesani, D., et al. (2019). The fraction of ionizing radiation from massive stars that escapes to the intergalactic medium. *Mon. Notices R. Astronomical Soc.* 483, 5380–5408. doi:10.1093/mnras/sty3460

Tanvir, N. R., and Jakobsson, P. (2007). Observations of GRBs at high redshift. *Philosophical Trans. R. Soc. Lond. Ser. A* 365, 1377–1384. doi:10.1098/rsta.2006.1992

Tanvir, N. R., Laskar, T., Levan, A. J., Perley, D. A., Zabl, J., Fynbo, J. P. U., et al. (2018). The properties of GRB 120923A at a spectroscopic redshift of z ≈ 7.8. *Astrophysical J.* 865, 107. doi:10.3847/1538-4357/aadba9

Tanvir, N. R., Levan, A. J., Rol, E., Starling, R. L. C., Gorosabel, J., Priddey, R. S., et al. (2008). The extreme, red afterglow of GRB 060923A: Distance or dust? *Mon. Notices R. Astronomical Soc.* 388, 1743–1750. doi:10.1111/j.1365-2966.2008.13496.x

Thöne, C. C., Greiner, J., Savaglio, S., and Jehin, E. (2007). ISM studies of GRB 030329 with high-resolution spectroscopy. *Astrophysical J.* 671, 628–636. doi:10.1086/522558

Tinney, C., Stathakis, R., Cannon, R., Galama, T., Wieringa, M., Frail, D. A., et al. (1998). Grb 980425. *IAU Circ.* 6896, 3.

Tonry, J., and Davis, M. (1979). A survey of galaxy redshifts. I. Data reduction techniques. *Astronomical J.* 84, 1511–1525. doi:10.1086/112569

Tsvetkova, A., Frederiks, D., Golenetskii, S., Lysenko, A., Oleynik, P., Pal'shin, V., et al. (2017). The konus-wind catalog of gamma-ray bursts with known redshifts. I. Bursts detected in the triggered mode. *Astrophysical J.* 850, 161. doi:10.3847/1538-4357/aa96af

Tu, L. P., Luo, A. L., Jiang, B., Wei, P., Zhao, Y. H., and Liu, R. (2012). A novel method to determine the redshifts of active galaxies based on wavelet transform. *Spectrosc. Spectr. Analysis* 32, 2858–2862. doi:10.3964/j.issn.1000-0593(2012)10-2858-05

Van Paradijs, J., Groot, P. J., Galama, T., Kouveliotou, C., Strom, R. G., Telting, J., et al. (1997). Transient optical emission from the error box of the γ-ray burst of 28 February 1997. *Nature* 386, 686–689. doi:10.1038/386686a0

Volnova, A. A., Pozanenko, A. S., Gorosabel, J., Perley, D. A., Frederiks, D. D., Kann, D. A., et al. (2014). Grb 051008: A long, spectrally hard dust-obscured GRB in a lyman-break galaxy at z ≈ 2.8. *Mon. Notices R. Astronomical Soc.* 442, 2586–2599. doi:10.1093/mnras/stu999

Vreeswijk, P. M., Ledoux, C., Raassen, A. J. J., Smette, A., De Cia, A., Woźniak, P. R., et al. (2013). Time-dependent excitation and ionization modelling of absorption-line variability due to GRB 080310. *Astronomy Astrophysics* 549, A22. doi:10.1051/0004-6361/201219652

Vreeswijk, P. M., Ledoux, C., Smette, A., Ellison, S. L., Jaunsen, A. O., Andersen, M. I., et al. (2007). Rapid-response mode VLT/UVES spectroscopy of GRB 060418: Conclusive evidence for UV pumping from the time evolution of Fe II and Ni II excited- and metastable-level populations. *Astronomy Astrophysics* 468, 83–96. doi:10.1051/0004-6361:20066780

Wadadekar, Y. (2005). Estimating photometric redshifts using Support vector machines. *Astronomical Soc. Pac.* 117, 79–85. doi:10.1086/427710

Wang, D., Zhang, Y. X., Liu, C., and Zhao, Y. H. (2008). Two novel approaches for photometric redshift estimation based on SDSS and 2MASS. *Chin. J. Astronomy Astrophysics* 8, 119–126. doi:10.1088/1009-9271/8/1/13

Wang, J., Qiu, Y.-L., and Wei, J.-Y. (2020). A pilot study of catching high-z GRBs and exploring circumburst environment in the forthcoming SVOM era. *Res. Astronomy Astrophysics* 20, 124. doi:10.1088/1674-4527/20/8/124

Wei, J., Cordier, B., Antier, S., Antilogus, P., Atteia, J. L., Bajat, A., et al. (2016). *The Deep and transient universe in the SVOM era: New challenges and opportunities - scientific prospects of the SVOM mission. arXiv e-prints* , arXiv:1610.06892.

Wei, J. J., Wu, X. F., Wang, F. Y., Liu, Z., Dai, Z. G., and Zhang, B. (2015). "High redshift gamma-ray bursts as a probe of the early universe and first stars," in *The 3rd einstein probe science workshop. SCIENTIA SINICA physica, mechanica & astronomica)*. Editor X. F. Wu, 1–30.

Wijers, R. A. M. J., Rees, M. J., and Meszaros, P. (1997). Shocked by GRB 970228: The afterglow of a cosmological fireball. *Mon. Notices R. Astronomical Soc.* 288, L51–L56. doi:10.1093/mnras/288.4.L51

Woosley, S. E., and Bloom, J. S. (2006). The supernova gamma-ray burst connection. *Annu. Rev. Astronomy Astrophysics* 44, 507–556. doi:10.1146/annurev.astro.43.072103.150558

Xie, L., Wang, X. G., Zheng, W., Filippenko, A. V., Qin, S. M., Li, L., et al. (2020). Early optical observations of GRB 150910A: Bright jet optical afterglow and X-ray dipole radiation from a magnetar central engine. *Astrophysical J.* 896, 4. doi:10.3847/1538-4357/ab8d2a

Xu, X., Luo, A. L., Wu, F. C., and Zhao, Y. H. (2006a). Using neural networks based template matching method to obtain redshifts of normal galaxies. *Spectrosc. Spectr. Analysis* 25, 996–1001. doi:10.3321/j.issn:1000-0593.2005.06.050

Xu, X., Wu, F. C., Hu, Z. Y., and Luo, A. L. (2006b). A novel method for the determination of redshifts of normal galaxies by non-linear dimensionality reduction. *Spectrosc. Spectr. Analysis* 26, 182–186. doi:10.3321/j.issn:1000-0593.2006.01.049

Yuan, H. Y., Lü, H. J., Li, Y., Zhang, B. B., Sun, H., Rice, J., et al. (2022). Probing the progenitor of high-z short-duration GRB 201221D and its possible bulk acceleration in prompt emission. *Res. Astronomy Astrophysics* 22, 075011. doi:10.1088/1674-4527/ac712d

Yue, C., Hu, Q., Zhang, F. W., Liang, Y. F., Jin, Z. P., Zou, Y. C., et al. (2018). How special is GRB 170817A? *Astrophysical J. Lett.* 853, L10. doi:10.3847/2041-8213/aaa66c

Zaninoni, E., Bernardini, M. G., Margutti, R., Oates, S., and Chincarini, G. (2013). Gamma-ray burst optical light-curve zoo: Comparison with X-ray observations. *Astronomy Astrophysics* 557, A12. doi:10.1051/0004-6361/201321221

Zhang, B. (2007). Gamma-ray burst afterglows. *Adv. Space Res.* 40, 1186–1198. doi:10.1016/j.asr.2007.01.016

Zhou, H., Huang, L. Y., and Luo, M. L. (2000). A stratified approach for quasar recognition based on hough transform and neural network. *J. Electron. Inf. Technol.* 2000, 529–535. doi:10.1007/BF02948846

Zhu, G. H., Luo, A. L., and Zhao, Y. H. (2005). An automated measurement of the galaxies spectra of lamost. *Spectrosc. Spectr. Analysis* 25, 1002–1005. doi:10.3321/j.issn:1000-0593.2005.06.051

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 11 July 2023
DOI 10.3389/fspas.2023.1159166

Check for updates

OPEN ACCESS

EDITED BY
Qiang Hu,
University of Alabama in Huntsville,
United States

REVIEWED BY
Tae Kim,
University of Alabama in Huntsville,
United States
Keiji Hayashi,
George Mason University, United States

*CORRESPONDENCE
Ming Xiong,
mxiong@swl.ac.cn
Kairan Ying,
kairanying@ninhm.ac.cn

RECEIVED 05 February 2023
ACCEPTED 20 June 2023
PUBLISHED 11 July 2023

CITATION
Xiong M, Feng X, Li B, He J, Wang W, Gao Y, Zhang M, Yang L, Huang Z, Cheng J, Su C, Yan Y and Ying K (2023), Interplanetary scintillation observation and space weather modelling. *Front. Astron. Space Sci.* 10:1159166. doi: 10.3389/fspas.2023.1159166

COPYRIGHT
© 2023 Xiong, Feng, Li, He, Wang, Gao, Zhang, Yang, Huang, Cheng, Su, Yan and Ying. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Interplanetary scintillation observation and space weather modelling

Ming Xiong[1,2,3]*, Xueshang Feng[1,3], Bo Li[4], Jiansen He[5], Wei Wang[1], Yanchen Gao[1,2], Man Zhang[1], Liping Yang[1], Zhenghua Huang[4], Jun Cheng[1], Cang Su[1], Yihua Yan[1,2] and Kairan Ying[6]*

[1]SIGMA Weather Group, State Key Laboratory of Space Weather, National Space Science Center, Chinese Academy of Sciences, Beijing, China, [2]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, China, [3]HIT Institute of Space Science and Applied Technology, Shenzhen, China, [4]Shandong Provincial Key Laboratory of Optical Astronomy and Solar-Terrestrial Environment, Institute of Space Sciences, Shandong University, Weihai, China, [5]School of Earth and Space Sciences, Peking University, Beijing, China, [6]National Institute of Natural Hazards, Ministry of Emergency Management of the People's Republic of China, Beijing, China

Interplanetary scintillation (IPS) refers to random fluctuations in radio intensity of distant small-diameter celestial object, over time periods of the order of 1 s. The scattering and scintillation of emergent radio waves are ascribed to turbulent density irregularities transported by the ubiquitous solar wind streams. The spatial correlation length of density irregularities and the Fresnel radius of radio diffraction are two key parameters in determining the scintillation pattern. Such a scintillation pattern can be measured and correlated between multi-station radio telescopes on the Earth. Using the "phase-changing screen" scenario based on the Born approximation, the bulk-flow speed and turbulent spectrum of the solar wind streams can be extracted from the single-station power spectra fitting and the multi-station cross-correlation analysis. Moreover, a numerical computer-assisted tomography (CAT) model, iteratively fit to a large number of IPS measurements over one Carrington rotation, can be used to reconstruct the global velocity and density structures in the inner heliosphere for the purpose of space weather modelling and prediction. In this review, we interpret the underlying physics governing the IPS phenomenon caused by the solar wind turbulence, describe the power spectrum and cross correlation of IPS signals, highlight the space weather application of IPS-CAT models, and emphasize the significant benefits from international cooperation within the Worldwide IPS Stations (WIPSS) network.

KEYWORDS

interplanetary scintillation, Fresnel diffraction, solar wind turbulence, computer-assisted tomography, space weather, coronal mass ejection, Sun, radio astronomy

1 Introduction

The solar wind escapes from the Sun, travels at a supersonic and super-Alfvénic speed, permeates interplanetary space, and ultimately fills the whole heliosphere. The ion species of the solar wind plasma include protons, alpha particles, and minor heavy ions. The solar wind was intuitively conceived as a result of hot coronal expansion. The Alfvén surface is the locus where the radial motion of the accelerating solar wind passes the radial Alfvén

speed. The sub-Alfvénic solar wind in the inner heliosphere was firstly measured by Parker Solar Probe (*PSP*) during its 8th and 9th solar encounters, at a distance of ≈ 16 solar radii ($R_\odot$) from the Sun. The first *in situ* detections enabled by PSP reveal the distinct nature of turbulence, anisotropy, intermittency, and directional switchback properties of these sub-Alfvénic winds (Kasper et al., 2019). According to PSP observation during encounter 8, the magnetic field and flow velocity vectors were highly aligned in the sub-Alfvénic solar wind (Zank et al., 2022). The solar wind in the heliosphere was locally sampled by Ulysses along its out-of-ecliptic orbit. According to the Ulysses' fast-latitudinal scan between ±80°, the distribution of the solar wind within the heliosphere is clearly bimodal, with differing compositions, temperatures, temperature anisotropies, speeds, small scale fluctuations (McComas et al., 2002). At solar maxima, this bimodal configuration gives way to a more complex mixture of slow and fast streams at all heliospheric latitudes, depending on the distribution of open and closed magnetic regions and the highly tilted magnetic polarity inversion line. The background solar wind flow is frequently disturbed by coronal mass ejections (CMEs), large-scale expulsions of plasma and magnetic field from the solar atmosphere. CMEs cause phenomena on the Earth such as geomagnetic storms and solar energetic particles that can result in major space weather effects (Forbes et al., 2006). As CMEs are the main driver of interplanetary and geomagnetic disturbances, interplanetary space can be considered as a transmission channel of hazardous space weather effects connecting the Sun and the Earth. The space weather chain between the CME initiation at the Sun and its ensuing disturbance at 1 AU can be quantitatively studied via global three-dimensional (3D) magnetohydrodynamic (MHD) simulations of the heliosphere such as Space Weather Modeling Framework (SWMF) (Tóth et al., 2005), Conservation Element and Solution Element (CESE)-MHD model (Feng et al., 2010; Feng, 2020), CORona-HELiosphere (CORHEL) (Riley et al., 2012), and EUropean Heliospheric FORecasting Information Asset (EUHFORIA) (Pomoell and Poedts, 2018), ENLIL-Cone model (Odstrcil et al., 2004), Multi-Scale FLUid-Kinetic Simulation Suite (MS-FLUKSS) flux-rope model (Singh et al., 2019), Grid Agnostic MHD for Extended Research Applications (GAMERA) (Merkin et al., 2016). Presently, and in the foreseeable future, numerical models based on the 3D MHD equations are the only self-consistent mathematical descriptions that can span the enormous distances associated with large-scale space weather phenomena (De Zeeuw et al., 2000).

The currently observable signatures of interplanetary solar wind and CMEs include local particle/field properties, heliospheric white-light emission, interplanetary radio bursts, and interplanetary scintillation (IPS). The former three are observed aboard interplanetary spacecraft, and the last is received by ground-based radio telescopes. Using both particle and wave detectors, the solar wind can be comprehensively diagnosed for its properties such as the speed, temperature, mass flux, composition, magnetic field, charge states, and waves/turbulence. *In-situ* measurements of solar wind are inherently confined to interplanetary Sun-orbiting trajectories of host spacecraft. Heliospheric white-light imaging fills the observation gap between near-Sun coronagraph imaging and *in situ* measurements. At large elongations from the Sun, the white-light brightness depends on both the local electron density and the efficiency of the Thomson-scattering process (Xiong et al., 2013). The CME substructures can be identified from their white-light emission observed by Heliospheric Imager (HI) instrument onboard Solar TErrestrial RElations Observatory (STEREO), such as the leading-edge pileup, interior void, filamentary structure and rear cusp (DeForest et al., 2011). Across the interplanetary CME-driven shock front, non-thermal electrons are accelerated to produce kilometric type II bursts (Cane et al., 1987). The accelerated electron beams generate plasma waves, which get converted into electromagnetic radiation at the fundamental and harmonic of the local plasma frequency. As the interplanetary shock moves anti-sunward through the expanding solar wind, the slow-drifting type II radio emissions are generated at lower frequencies. By triangulating the type II radio source at several times and frequencies, the CME-driven shock can be tracked through interplanetary space (Reiner et al., 1998). In addition, faint IPS radio signals, transmitted through the turbulent solar wind, can be received by large-aperture radio telescopes on the Earth. IPS is the variation in the apparent strength of extragalactic radio sources introduced by density irregularites in the solar wind (Hewish et al., 1964).

IPS telescopes can be used to measure both the solar wind speed and the micro-turbulence spectrum of the interplanetary plasma. The above-mentioned four techniques of *in situ* measurements, white-light imaging, radio burst detecting, and IPS recording complement each other for the common purpose of interplanetary space weather monitoring. The single-point *in situ* measurements by spacecraft cannot provide a global view of interplanetary 3D large-scale structures. Either white-light imager or type II radio burst detector can be used to track one CME throughout the interplanetary space. The interplanetary CME is generally identified as a moving brightness front in white-light images and a drifting brightness pattern in radio spectrogram. Both white-light and radio signatures can be related to coronagraph images and *in situ* measurements. Moreover, IPS data set during one Carrington rotation can be fed into a IPS-based computer-assisted tomography (CAT) model to reconstruct 3D spatial distribution of large-scale solar wind structures within 1 AU (e.g., Jackson et al., 1998; Kojima et al., 1998; Hayashi et al., 2003). Particularly, several radio telescopes, such as LOw Frequency ARray (LOFAR) (van Haarlem et al., 2013), have been commissioned to observe the solar corona during PSP perihelion passages. By combining the imaging and beam-forming observational modes, LOFAR can be used to detect solar type III radio bursts and other plasma processes associated with energetic electrons in the corona. The low-frequency bands available with LOFAR can be used to monitor IPS signals of the ubiquitous solar wind from the Mercury orbit out to beyond 1 AU. Both the imaging of the corona and the observation of IPS can contribute to the study of large-scale solar wind flows along the PSP orbit.

Radio scattering and scintillation observations are well suited for probing the corona and the inner heliosphere. The various radio methods include intensity scintillation, phase scintillation, angular broadening, and spectral broadening. Angular and spectral broadening are used to diagnose the high-frequency end of the solar wind turbulence spectrum, where kinetic damping and other wave-particle interactions are important (Harmon and Coles, 2005). Intensity scintillation observations using spaced antennas have proven to be particularly effective in extracting the solar wind speed.

In this review paper, we focus on the topic of interplanetary intensity scintillation, introduce the historical and current IPS-related radio telescopes around the world in Section 2, present the magnetic fluctuation spectrum and density fluctuation spectrum of solar wind turbulence in Section 3, interpret Fresnel diffraction caused by turbulent density irregularities in Section 4, analyze spatial and temporal correlation of IPS time-series signals in Section 5, describe the IPS-CAT modelling for space weather prediction in Section 6, emphasize the significance of international cooperation within the worldwide IPS Stations (WIPSS) network in Section 7, and present a summary and discussion in Section 8.

2 Historical and current IPS-related radio telescopes around the world

The IPS phenomenon refers to random fluctuations in radio intensity of distant small-diameter celestial object, over time periods of the order of 1 s. Such random radio fluctuations result from the passage of radio rays through the irregular interplanetary plasmas. Density irregularities are present in the solar wind, and are manifested through fluctuations in the refractive index. The first IPS discovery at Cambridge gives the observation evidence that such a scintillation is found to occur at any angular distance from the Sun (Hewish et al., 1964). Using the Cambridge IPS array during its early IPS sky survey, the serendipitous discovery of pulsars (pulsating radio stars) was found by Jocelyn Bell Burnell and published by Hewish et al. (1968). The Cambridge IPS survey at 81.5 MHz covers the area of sky between declinations −10° and +83° at all values of right ascension. Those compact radio sources, such as pulsars and some unusual extragalactic sources, or those in which energy is being released from active beams in the outer lobes of intrinsically powerful radio galaxies and quasars, are summarized in a catalog of 1,789 radio sources which exhibit IPS at 81.5 MHz (Purvis et al., 1987). The sensitivity of the IPS survey is not uniform over the sky, being determined largely by the galactic background emission.

Historical and current IPS-related radio telescopes are symbolized in Figure 1 to display geographic distribution in a worldwide Mercator map. These telescopes can be classified as three types: (1) decommissioned IPS-related stations, including Cambridge IPS array in United Kingdom (Dennison and Hewish, 1967), University of California at San Diego (UCSD) IPS array in United States (Armstrong and Coles, 1972; Coles and Kaufman, 1978), Miyun Synthesis Radio Telescope (MSRT) in China (Qiu, 1996), and Arecibo radio telescope in Puerto Rico (Abe Pacini, 2020); (2) non-IPS-dedicated stations in commission, including Five-hundred-meter Aperture Spherical radio Telescope (FAST) in China (Nan et al., 2011), Murchison Widefield Array (MWA) in Australia (Oberoi and Benkevitch, 2010), LOw Frequency ARray (LOFAR) in western Europe (van Haarlem et al., 2013; Fallows et al., 2023), and European Incoherent SCATter (EISCAT) Radar in northern Europe (Bourgois et al., 1985); (3) IPS-dedicated station in commission, including Institute for Space-Earth Environmental research (ISEE) IPS array in Japan (Tokumaru et al., 2011), Ooty Radio Telescope (ORT) in India (Manoharan and Ananthakrishnan, 1990), MEXican Array Radio Telescope (MEXART) in Mexico (Mejia-Ambriz et al., 2010), Big Scanning Array of Lebedev Physical Institute (BSA LPI) in Russia (Shishov et al., 2010), and 327-MHz radio array in Korea. As shown in Figure 2, multiple radio telescopes at different geographic locations can be combined to form a multi-station IPS array, which allows an accurate determination of the spatial properties of the drifting pattern of IPS (Little and Ekers, 1971; Coles and Kaufman, 1978; Bourgois et al., 1985). The early such work of multi-station IPS observations was done at United Kingdom (Dennison and Hewish, 1967), Russia (Vitkevich and Vlasov, 1970), United States (Armstrong and Coles, 1972), and Japan (Watanabe and Kakinuma, 1972). Multi-station IPS observations had been made regularly at frequency of 74 MHz (Armstrong and Coles, 1972) and at 69.3 MHz (Kakinuma et al., 1973) for which the probed coronal regions were beyond 60 $R_\odot$. The first observations of IPS with EISCAT facilities were made at frequency of 933.5 MHz to measure the solar wind velocity at the heliocentric distances between 15 and 70 $R_\odot$ (Bourgois et al., 1985). Now, the ISEE IPS array consisting of Toyokawa, Kiso, and Fuji stations is the only operational multi-station IPS facility. In addition, a brand-new three-site IPS-dedicated system in northern China is expected to be constructed in the late 2023 and calibrated afterwards, supported by the mega-framework of the Meridian Space Weather Monitoring Project in China (Yan et al., 2018).

The current IPS-dedicated radio telescopes include the single-station systems [such as ORT (Manoharan et al., 1995), MEXART (Romero-Hernandez et al., 2015), and BSA LPI (Shishov et al., 2010)] and the multi-site ISEE system (Tokumaru et al., 2011). The ORT has a parabolic cylinder 530 m long in the north-south direction and 30 m wide in the east-west direction (Swarup et al., 1971), which is equipped with a 12−consecutive simultaneous beam system with a separation of ~ 3 arc-minute between adjacent beams. The BSA LPI telescope at Pushchino has two independent 16-beam systems at frequency of 111 MHz, which is the world's largest (70,000 m^2) currently operating radio array (Shishov et al., 2010). The MEXART consists of 64 × 64 (4,096) full-wavelength dipole antenna array, operating at 140 MHz, with a bandwidth of 2 MHz, occupying about 9,660 square meters (69 m × 140 m) (Mejia-Ambriz et al., 2010). The current ISEE IPS system in commission consists of three antennas at Toyokawa, Fuji, and Kiso. The radio telescope at Toyokawa, called the Solar Wind Imaging Facility Telescope (SWIFT), has the largest aperture among the ISEE IPS system. The SWIFT consists of a pair of asymmetric cylinderical parabolic reflector antennas with an aperture size of 108 m (north-south) by 19 m (east-west) (Tokumaru et al., 2011). A new Chinese IPS-dedicated system under construction consists of one main station and two sub-stations: (1) a parabolic dish antenna with its aperture of 30 m at either sub-station; (2) three cyclindrical parabolic reflector antennas placed side by side at the main station at Ming'antu, each of which is 140 m in north-south direction and 40 m in east-west direction (Wang et al., 2021).

IPS observations have been extensively used to study the solar wind since the first discovery of IPS signals in 1964 by the Cambridge IPS array (Hewish et al., 1964). A comparison with spacecraft observations in the ecliptic is used as a calibration for the IPS observations. According to correlation of IPS and spacecraft plasma density measurements, heliospheric structures can be classified as either corotating or detached from the Sun (Houminer, 1971). The IPS observations have proved to be remarkably sensitive to long lived structures such as corotating interaction region (CIR) (Burnell, 1969; Coles, 1978). In 1970s, the recurrent solar wind

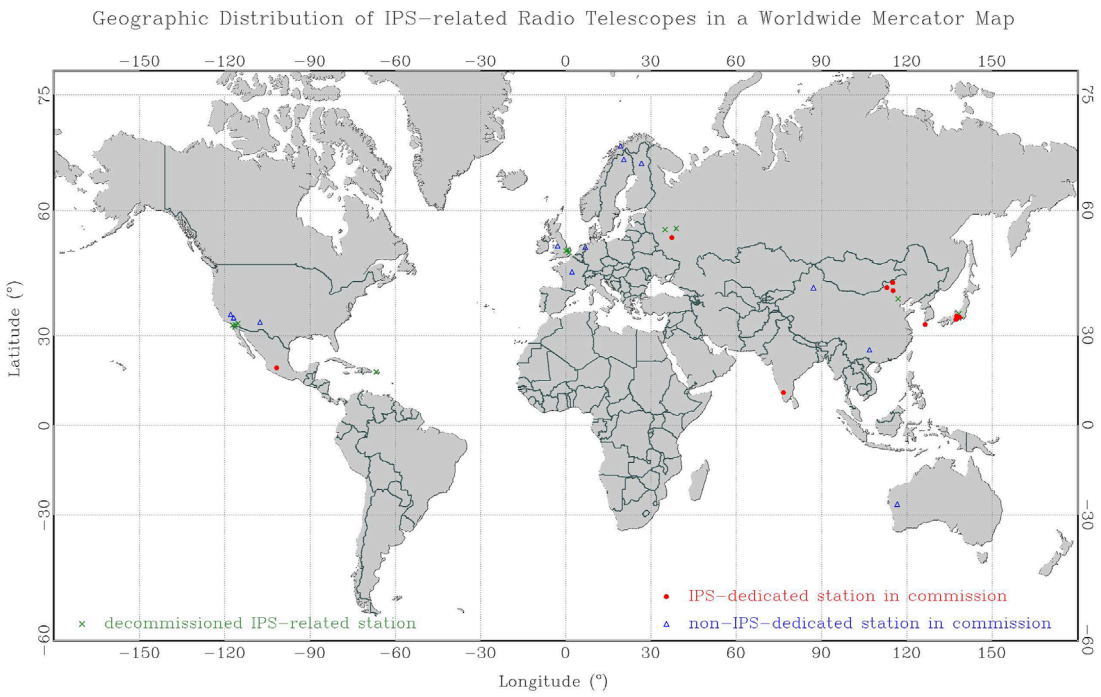


FIGURE 1
Geographic distribution of IPS-related radio telescopes in a worldwide Mercator map.

streams discovered by IPS observations are statistically correlated with the brightness of the EUV corona in Fe XV (Watanabe et al., 1974). The inverse correlation between the solar wind speed and the magnetic flux expansion factor, proposed by Wang and Sheeley (1990), was substantiated by IPS observations of solar wind speed at different latitudes of the heliosphere during one full solar cycle (Sheeley et al., 1991). The stream interface could be a shear and sliding layer between a fast stream at the high latitude and a slow stream at the ecliptic plane, or a compression layer within the CIR, as the IPS observations detect the velocity gradient and normal scintillation level for the sliding layer, and an intermediate velocity and enhanced scintillation level for the compression layer (Bisi et al., 2010). As revealed from IPS observations, interplanetary CME-driven shocks tend to propagate toward the low-latitude region near the solar equator and the fastest propagation directions tend toward the heliospheric current sheet near 1 AU (Wei and Dryer, 1991; Wei et al., 2005). Using a large-aperture radio telescope, both recurrent CIRs and transient CMEs are readily detected in IPS time-series signals.

3 Solar wind as a natural turbulence laboratory

The solar wind plasma is collisionless, and waves/turbulence are ubiquitous. Using the solar wind magnetic field data, an effective magnetic Reynolds number is estimated to be about 260,000 ± 20,000 (Weygand et al., 2007). The very large magnetic Reynolds number suggests that large-scale turbulent irregularities should develop, even if the nascent wind were originally smooth and laminar. Because of solar wind density gradients, turbulence in the sub-Alfvénic solar wind is driven by reflection of low-frequency Alfvén waves (Verdini et al., 2009). As the first-ever mission to "touch" the Sun, PSP offers an unprecedented opportunity to characterize the solar wind and the solar energetic particles near their origin. According to PSP's *in situ* observation within 0.3 AU, the dominant composition of solar wind turbulence is ascribed to the outward Alfvén mode, and the minor composition is found to be the inward Alfvén mode and the outward fast mode (Zhu et al., 2020). MHD turbulence is always locally imbalanced in creating patches of positive and negative cross-helicities. The conservation of cross-helicity results in a hierarchical structure of MHD turbulence. The slow solar wind is fully mixed by sub-diffusive eddies on time scales corresponding to a 1–2 h crossing time on the Earth; and that solar wind variability on shorter timescales is dominated by turbulent processing, rather than by remnants of variability in the source process at the Sun (DeForest et al., 2015).

The solar wind turbulence is an ensemble of fluctuations with random phases and a broad range of wave vectors. Such fluctuations are manifested in, say, the magnetic field $\delta\mathbf{B}$, electron number density δn, plasma bulk flow velocity $\delta\mathbf{v}$. The Elsässer variables for the fluctuating velocity $\delta\mathbf{v}$ and magnetic $\delta\mathbf{B}$ fields are defined as $\mathbf{z}^{\pm} = \delta\mathbf{v} \pm \delta\mathbf{B} / \sqrt{\mu_0 \rho}$, where ρ is the mean plasma density. The fluctuation energy is usually transferred through a forward cascade. Namely, the energy is injected from a preexisting population of MHD waves at very low angular frequencies ω and wave numbers $\mathbf{k}$, nonlinearly transported through successively shorter wavelengths via wave-wave interaction, and eventually dissipated at ion kinetic scales via collisionless wave-particle interaction. For Alfvén waves with isotropic kinetic pressure, the density spectrum of wave energy $E(\mathbf{k})$ is related to the magnetic fluctuation spectrum $S_B(\mathbf{k})$ in the moving plasma frame. Such a mathematical relation is governed by

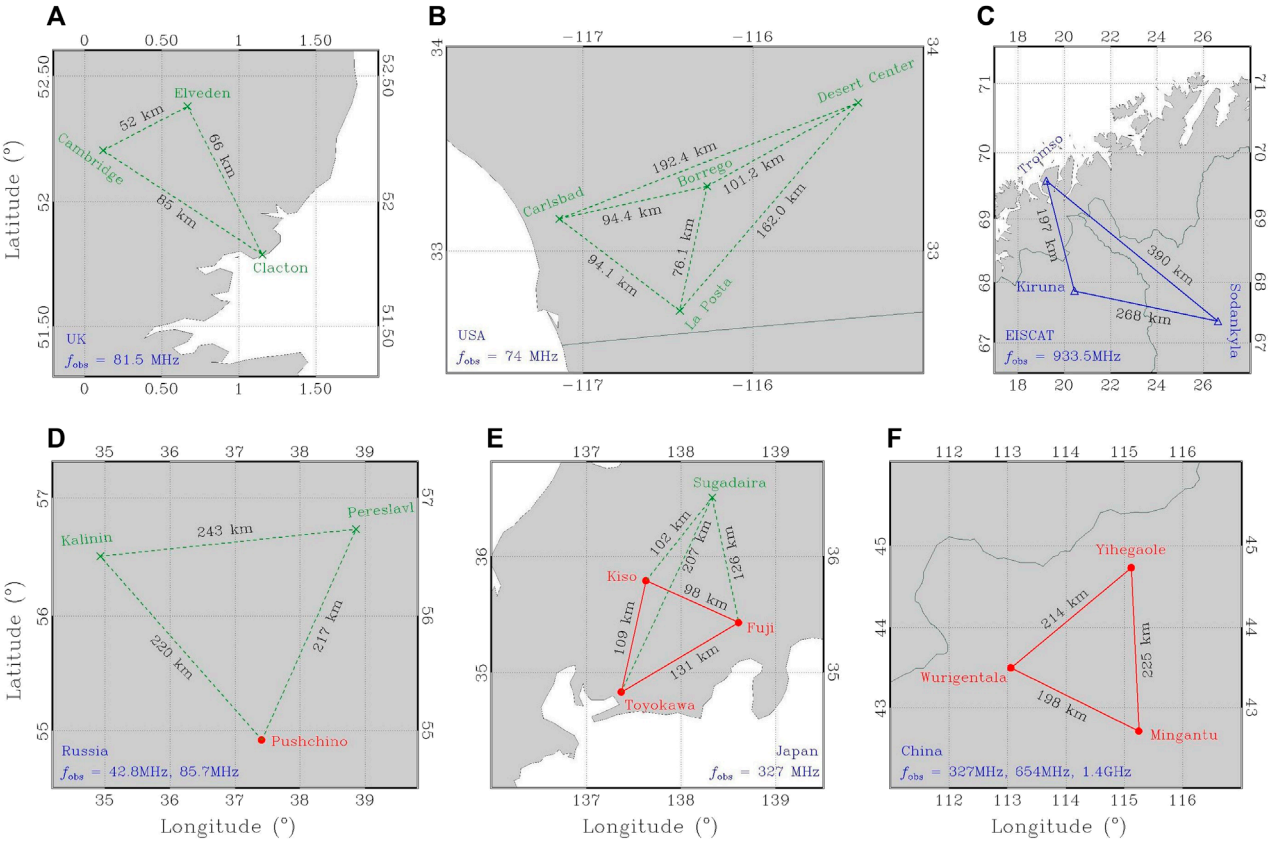


FIGURE 2
Multi-station IPS observations in **(A)** United Kingdom, **(B)** United States, **(C)** northern Europe, **(D)** Russia, **(E)** Japan, and **(F)** China, with the annotated geographic name of each station, baseline length between either two of stations, and observing frequency f_{obs}. The de-commissioned stations and corresponding baselines are annotated in green in each panel.

the following equation (Barnes, 1979):

$$S_B(\mathbf{k}) = \mu_0 \left(1 + \mathbf{e}_k \cdot \mathbf{v}/v_p\right)^{-1} E(\mathbf{k}) \tag{1}$$

Here $\mathbf{e}_k$ is the unit wave vector of $\mathbf{k}$, $\mathbf{v}$ the solar wind speed vector, and $v_p = \omega_0/|\mathbf{k}|$ the wave phase velocity, where ω_0 is the intrinsic wave frequency in the plasma frame. From $S_B(\mathbf{k})$, the density fluctuation spectrum $S_n(\mathbf{k})$ can be derived on basis of wave compressibility $C(\mathbf{k})$

$$S_n(\mathbf{k}) = \left(\frac{n^2}{B^2}\right) C(\mathbf{k})\, S_B(\mathbf{k}) \tag{2}$$

$$C(\mathbf{k}) \equiv \frac{(\delta n/n)^2}{(\delta B/B)^2} = \left(\frac{\mu_0\, \Omega}{\omega_p^2}\right) \frac{|\mathbf{k}\cdot\boldsymbol{\sigma}\cdot\delta\mathbf{E}|^2}{|\mathbf{k}\times\delta\mathbf{E}|^2} \tag{3}$$

Here n is the mean plasma number density, B magnetic flux density, δn density fluctuation, δB magnetic fluctuation, $\delta\mathbf{E}$ electric fluctuation, Ω proton cycloron frequency, ω_p plasma frequency, $\boldsymbol{\sigma}$ conductivity tensor (Gary, 1986; Harmon and Coles, 2005). The δn, δB, and $\delta\mathbf{E}$ refer to temporal fluctuations at the kinetic scale of Ω, which are estimated to be 32 Hz at 10 $R_\odot$, 1 Hz at 55 $R_\odot$, and 0.1 Hz at 1 AU, respectively (Bale et al., 2016). Whereas, n and B are statistically averaged over a much longer macroscopic scale to be slowly-changing background values. The turbulence amplitude in the corona (Yamauchi et al., 1996) and the heliosphere (Coles and Harmon, 1978) can be measured by IPS observations of electron density spectra. A "turbulent" density spectrum is a direct signature of non-Alfvénic fluctuations such as convected pressure-balance structures (Burlaga et al., 1990) or fast magnetoacoustic waves (Tu and Marsch, 1995; Bruno and Carbone, 2013), because low-frequency Alfvén waves are essentially noncompressive. The near-Sun broadband density spectrum begins with a roughly Kolmogorov power law at large spatial scales of $>10^3$ km, then breaks to a local flattening at ~10^2 km, and finally cuts off at an inner scale of ~10 km (Coles and Harmon, 1989). The flattening and cutting-off parts in the density spectrum can be respectively ascribed to enhanced compressibility and cyclotron damping of highly oblique Alfvén/ion cyclotron waves. Harmon and Coles (2005) proposed that the density fluctuation spectrum observed in radio scattering is a composite of a power law component mimicking the underlying magnetic spectrum and a flatter component associated with the enhanced compressibility of oblique Alfvén waves at high wavenumber. As shown in Figure 3, the density spectrum in the sub-Alfvénic solar wind observed by PSP during Encounter 8 is a power law that resembles neither the $\mathbf{z}^{\pm}$ spectra nor the compressible magnetic field spectrum, suggesting that these are advected entropic rather than magnetosonic modes and not due to the parametric decay instability (Zank et al., 2022).

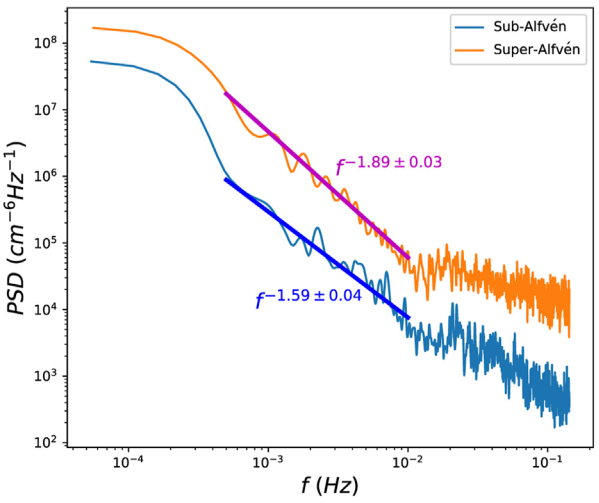


FIGURE 3
Power spectrums for the fluctuating density variance δn in the sub-Alfvénic and super-Alfvénic regions, observed by PSP during its encounter 8 (Zank et al., 2022).

The three-dimensional magnetic fluctuation spectrum $S_B(\mathbf{k})$, advected at the solar wind speed $\mathbf{v}$, is measured at a relatively stationary spacecraft as temporal turbulence data $S_{Bsc}(f)$. The one-dimensional frequency spectrum measured by spacecraft is termed "spacecraft magnetic spectrum" (Harmon and Coles, 2005). Using Taylor's hypothesis, the measured magnetic frequency spectrum can be interpreted as mapping of the turbulence wave-number into the stationary spacecraft frame.

$$S_{Bsc}(f) = \frac{2\pi}{v}\iint S_B\left(k_x, k_y, k_z\right) dk_y\, dk_z, \quad \text{where } k_x = 2\pi f/v \tag{4}$$

Here the x-direction is along the solar wind velocity vector. Taylor's hypothesis assumes that when the turbulent fluctuations are advected with a speed v that is much higher than the typical fluctuation speed δv ($v \gg \delta v$), the time (τ) and the spatial ($\mathbf{r}$) lags of the measured structures are connected as $\mathbf{r} = \mathbf{v}\tau$. The frozen-in approximation is not strictly valid inside the Alfvén surface, where PSP is periodically scheduled to explore.

In-situ measurements of magnetic field fluctuation of interplanetary solar wind have substantiated a power law spectrum as $|\delta\mathbf{B}(f)|^2 \approx f^{-\alpha}$. The solar wind turbulence in an inertial range is similar to the classical Kolmogorov picture of fluid turbulence, and the spectral index α is approximately 5/3 (Leamon et al., 1998). Then, the Kolmogorov dissipation rate Q_{kol} can be evaluated from velocity fluctuation $\langle \delta v^2 \rangle$, correlation length of the fluctuations L_c, and solar wind density ρ (Hollweg, 1986).

$$Q_{\text{kol}} = \rho \langle \delta v^2 \rangle^{3/2} / L_c \tag{5}$$

Where $\langle \delta v^2 \rangle$ is related to the wave pressure of solar wind turbulence p_w and cross-section area of solar wind flow tube by

$$\langle \delta v^2 \rangle = \frac{\langle \delta B^2 \rangle}{\mu_0 \rho} = \frac{2 p_w}{\rho} \tag{6}$$

Further, Q_{kol} can be quantitatively incorporated as an observation-constrained empirical term in numerical MHD models of Alfvén wave turbulence-driven solar wind to generate the global distribution of solar wind MHD parameters from the inner corona to 1 AU (e.g., Li et al., 2004).

4 Fresnel diffraction from the phase-changing screen in the turbulent solar wind

Consider a planar layer of turbulent plasma with its inherent density fluctuations $\delta n(\mathbf{r})$, a planar radio wave propagating through the turbulent layer would experience accompanying phase fluctuations $\delta\phi(\mathbf{r}) = r_e \lambda\, \delta n(\mathbf{r})\, \Delta z$ (Bastian, 1994). Here, r_e is the classical electron radius, λ the radiation wavelength, Δz the depth of turbulent plasma layer.

The planar turbulent layer can be consider as an infinitely thin "two-dimensional phase-changing screen" $\phi(x,y)$ at $z = 0$, whose statistical properties are independent of x, y, and time t. A mean-squared phase ϕ_0^2 is statistically defined to be a constant of $\phi_0^2 \equiv \langle \phi^2(x,y) \rangle$. For arbitrary distance $\mathbf{r} = (r_x, r_y)$, the real phase function ϕ at two points of (x,y) and $(x + r_x, y + r_y)$ is statistically correlated as an auto-correlation function (ACF) $\varrho(\mathbf{r})$.

$$\phi_0^2\, \varrho\left(r_x, r_y\right) \equiv \left\langle \phi(x,y)\, \phi\left(x + r_x,\, y + r_y\right) \right\rangle \tag{7}$$

With $\varrho(0,0) = 1$ and $\langle \phi(x,y) \rangle = 0$. As denoted by angular brackets, the statistical averaging over space (x, y) and over time (t) is equivalent. The Fourier transform of ϱ and its second moment are given as.

$$\phi_0^2\, \varrho\left(r_x, r_y\right) \equiv \iint dq_x\, dq_y\, \phi^2\left(q_x, q_y\right) \exp\left[i\left(q_x x + q_y y\right)\right] \tag{8}$$

$$\begin{aligned} q_{0x}^2 &\equiv \left[-d^2 \varrho\left(r_x, r_y\right)/dx^2\right]_{x=0,y=0} \\ &= \iint dq_x\, dq_y\, q_x^2\, \phi^2\left(q_x, q_y\right) / \phi_0^2 \end{aligned} \tag{9}$$

$$\begin{aligned} q_{0y}^2 &\equiv \left[-d^2 \varrho\left(r_x, r_y\right)/dy^2\right]_{x=0,y=0} \\ &= \iint dq_x\, dq_y\, q_y^2\, \phi^2\left(q_x, q_y\right) / \phi_0^2 \end{aligned} \tag{10}$$

The first derivative of $\varrho(r)$ vanishes at the origin, under the reasonable assumptions that q_{0x} and q_{0y} are finite and that $\phi^2(q_x, q_y)$ is an even function. The two-dimensional (2D) ACF may be anisotropic and have different correlation lengths along the x– and y– directions. Here, the 2D "correlation lengths" are defined by $a_x \equiv q_{0x}^{-1}$ and $a_y \equiv q_{0y}^{-1}$, respectively. The smaller correlation length of $a = \min \cdot (a_x, a_y)$ is estimated to be larger than 100 km in the interplanetary solar wind (Cohen et al., 1967). $\varrho(\mathbf{r})$ decreases rapidly for the distance of $|\mathbf{r}| > a$. The correlation length a is used to delimit the size of "turbules" of fluctuating plasma irregularities. For a plane radio wave of wavenumber $k = 2\pi/\lambda$, $ka \gg 1$ in all cases. The phase fluctuation $\phi(\mathbf{r})$ is contributed by many turbules of solar wind flows along the line of sight (LOS). Since the solar wind density decreases rapidly with distance from the Sun, the majority of the phase fluctuations are introduced close to the pierce point (p-point), which is defined to be the position closest to the Sun on the LOS. The depth Δz of scintillation region along the LOS is fairly small, compared to its distance z from the Earth. The fluctuating interplanetary plasma

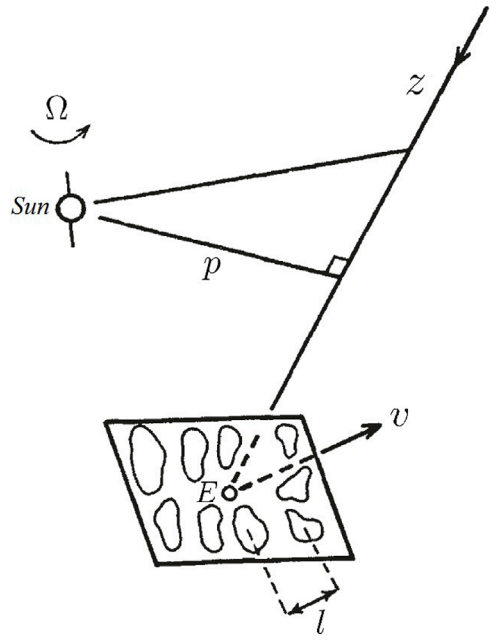


FIGURE 4
Geometry of the IPS diffraction pattern observed at the Earth due to turbulent solar wind irregularities (Hewish et al., 1964).

is approximated by a plane surface at distance z along which the phase of the radio wave fluctuates but the amplitude is constant. Such an approximation is formulated into a "thin phase-changing screen" theory (Salpeter, 1967). The correlation length a within the phase screen is generally Gaussian (Cohen et al., 1967). The incident plane wavefront, emitted from a point-like radio source such as pulsars, would be randomly distorted after its traversal through the phase screen. Interference between the different emergent plane waves makes up the angular spectrum. The density inhomogeneities in the screen can result in diffraction patterns of intensity variations observed on the Earth (Hewish et al., 1964), as shown in Figure 4. The observed time-series fluctuations of radio flux intensity depend on the relative velocity of the Earth through the pattern, and is consequently related to the orbital velocity of the Earth (30 km s^{-1}) and the velocity of the interplanetary solar wind. The diffraction scale $(\lambda z)^{1/2}$, also termed the radius of the first Fresnel zone, is about 122 km for $\lambda = 10$ cm and $z = 1$ AU. The width of the scattered cone of radiation is determined by the ACF across the emergent wavefront. As interpreted in a diffraction theory of IPS (Little, 1951), the ratio of diffraction scale $(\lambda z)^{1/2}$ to the correlation length a is an important criterion determining the type of scintillation phenomena observed.

A length $z_0 \equiv k\,a^2$ is defined as the "Fresnel distance" for diffraction from turbules of size a. Then, another length $l_0 \equiv z_0/\phi_0 \equiv k\,a^2/\phi_0$ is defined as the "typical focal length". The "thin phase screen" theory can be extended to the regime where the rms phase fluctuations are large and distances to the screen z are comparable with typical focal lengths l_0 (Salpeter, 1967). Supposing that a plasma slab of thickness Δz depicted in Figure 4 is comparable with the distance of closest approach to the Sun, the triple inequality of $\lambda \ll a \ll \Delta z \ll z$ generally holds. The phase along one radio ray path is delayed due to the fluctuating refractive index. Also, the ray path is lengthened due to refraction. The "thin phase-screen" approximation is still justified when the two inequilities of $z_0 \gg \Delta z$ and $l_0 \gg \Delta z$ hold, for any ϕ_0 and z (Salpeter, 1967).

For an extended radio source with its angular size θ_0, the brightness is usually supposed to have a Gaussian distribution $I(\varepsilon) = I_0 \exp(-\varepsilon^2/\theta_0^2)$ with respect to solar elongation ε. Both the source size $\varepsilon_0 z$ and the diffraction scale $(\lambda z)^{1/2}$ are measured at a distance z, the normal distance of the receiver from the "phase screen". For small ε_0, the turbulence at distance z contributes proportionally to $z^{5/6}$, as for a point source. However, when the source size approaches the diffraction scale $(\varepsilon_0 z \approx (\lambda\, z)^{1/2})$, high wavenumbers are attenuated by the averaging over the source rather than by diffraction; the relative contribution from z reaches a maximum and then decreases, being proportional to $z^{1/3}$ for large z. This decrease is so gradual that the major contribution to the integral over z still comes from near the Sun, for small ε. Hence, the scintillation from an extended source varies with ε in about the same way as that from a point source (Young, 1971).

As δI is the fluctuation in intensity I about its mean $\langle I\rangle$, scintillation index m is defined as the ratio of the rms intensity to the mean $(m^2 = \langle \delta I^2\rangle \,/\, \langle I\rangle^2)$. In the weak scattering case of an ideal source, the first Born approximation is valid and the index is much less than 1 ($m \ll 1$). Both finite bandwidth and finite source diameter have the effect of reducing the source coherence and thus reducing the intensity fluctuation. Numerical simulation of wave propagation in 3D random medium demonstrates the following consequences of gradual scattering enhancement (Martin and Flatte, 1988): (1) Under weak scattering, the small-scale Fresnel length structure of the medium dominates the intensity scattering pattern; (2) As the strength of scattering increases, caustics and interference fringes around focal regions begin to form; (3) In stronger scattering, the clustering of bright regions begins to reflect the large-scale structure of the medium. As shown in Figure 5, the scintillation index m for each compact source increases in the weak scattering region, reaches a peak where the index for an ideal source would saturate, then begins to decrease as the spatial intensity spectrum starts to widen (Coles, 1978). m peaks at a certain distance, where transition from weak to strong scattering occurs. The transition distance from weak to strong scattering, also called the scattering turnover distance, depends on the observing frequency, e.g., 4 $R_\odot$ for 8,085 MHz, 12 $R_\odot$ for 1,410 MHz, 34 $R_\odot$ for 327 MHz, and 50 $R_\odot$ for 195 MHz. The normalized scintillation level called the g–level, or g factor, is defined by $g = m/\langle m\rangle$. Here, $\langle m\rangle$ is the mean level of scintillation for the source at elongation ε at the time of observation. The intensity variation of signal strength from a compact radio source can be measured as g–level (Bisi et al., 2010). Such g–level is a proxy for the level of interplanetary disturbance along the LOS.

5 Spatial and temporal correlation of IPS time-series signals measured on the Earth

The space and time correlation of intensity fluctuations is evaluated by equation $\varrho(\mathbf{s},\tau) = \langle \delta I(\mathbf{r_0}, t_0)\, \delta I(\mathbf{r_0}+\mathbf{s}, t_0+\tau)\rangle$. The IPS power spectrum can be derived from Fourier Transformation of ACF, $P_I = \frac{1}{2\pi}\int \varrho(0,t)\exp(-i\,2\pi f t)\,dt$. To yield g–levels in real time,

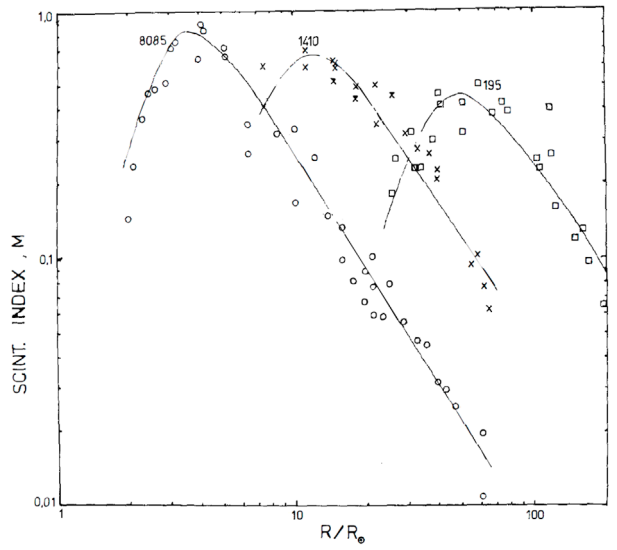


FIGURE 5
Heliocentric variances of radio scintillation indices at three different frequencies of 8,085, 1,410, and 195 MHz (Coles, 1978).

the white noise P_{WN} is subtracted from the scintillation signal spectrum P_f, and then system gain corrections are determined by automatically calibrating with the white noise level at the high-frequency end of the power spectrum (Jackson and Hick, 2004). For each radio source, IPS data are automatically edited to remove any obvious interference discerned in the daily observations, and white noise is subtracted from the signal to obtain scintillation index, i.e., $m = \int (P(f) - P_{WN})/P_{WN} df$.

The IPS model for the local intensity modulation of radio waves makes use of a Born approximation to the general weak scattering theory (Tatarskii et al., 1993). The spatial fluctuation of the local density irregularities is conveyed by the ambient solar wind flow, and consequently introduces a scintillation pattern in the (x, y) plane perpendicular to the IPS ray-path along the z direction. Here (x, y, z) is a Cartesian coordinate system centred on the Earth. The intensity $P_{\delta I}$ in a total spectrum of the spatial wave vector $\mathbf{q}$ is merely a linear superposition of all contributions from every thin scattering layer along the IPS ray-path connecting the Earth to an extragalactic radio source, as described by $P_{\delta I} = \int \delta P_{\delta I}(\mathbf{q}, z) dz$. The two-dimensional spatial wave vector $\mathbf{q}$ can be resolved into (q_x, q_y) components in the reference of IPS ray-path or $(q_\parallel, q_\perp)$ components in the reference of local magnetic field. The absolute value of spatial wavenumber $\mathbf{q}$ is evaluated as $q = |\mathbf{q}| = (q_x^2 + q_y^2)^{1/2} = (q_\parallel^2 + q_\perp^2)^{1/2}$. For each scattering layer with its depth dz at distance z, a mathematical relation between the radio intensity scintillation $\delta P_{\delta I}(\mathbf{q}, z)$ and the electron density irregularities $\delta n(z)$ is given by the following equation set.

$$\delta P_{\delta I}(\mathbf{q}, z) = w(\mathbf{q})\ \delta n^2(z) \tag{11}$$

$$w(\mathbf{q}) = 2\pi (r_e \lambda)^2 F_{\text{diff}}(\mathbf{q})\ F_{\text{source}}(\mathbf{q}, z)\ S_n(\mathbf{q}) \tag{12}$$

$$F_{\text{diff}}(\mathbf{q}) = 4 \sin^2\left(\frac{q^2 r_F^2}{4\pi}\right) \tag{13}$$

$$F_{\text{source}}(\mathbf{q}, z) = \exp\left(-q^2 z^2 \theta_0^2/2\right) \tag{14}$$

$$S_n(\mathbf{q}) = \left(q_\parallel^2 + q_\perp^2/A_R^2\right)^{-\alpha/2} \tag{15}$$

Here, $w(\mathbf{q})$ is a weighting factor at spatial wave vector $\mathbf{q}$, the plasma number density fluctuation δn, the classic electron radius $r_e = 2.82 \times 10^{-5}$ m, the Fresnel radius $r_F = (\lambda z)^{1/2}$, the spectral exponent α, and the axial ratio of anisotropy degree A_R (Coles and Harmon, 1989; Klinglesmith, 1997; Xiong et al., 2011). When $\lambda = 1$ m, and $z = 1$ AU (1.5×10^{11} m), the Fresnel radius is about 387 km. The cylindrically-symmetric coordinate system $(q_\parallel, q_\perp)$ with its axis along the local magnetic field is transferred from the coordinate system (x, y, z) defined with respect to the IPS ray-path. As the variation of δn along the IPS ray-path z is not known *a priori*, it is a common practice in IPS studies to assume some empirical δn variances such as $\delta n \propto n$ (Houminer and Hewish, 1972) and $\delta n \propto r^{PWR} n^{PWN}$ (Asai et al., 1998). Here, r is the heliocentric distance, PWR a power of radial falloff, and PWN the power of electron density. The parameters of PWR and PWN are determined via best fit of IPS data over one interval chosen. Particularly, the term F_{diff} is a high-pass Fresnel filter with its Fresnel radius r_F. Such a radius r_F determines the maximum scale of the irregularities, at which the amplitude fluctuation can be received at the Earth.

Using solar wind velocity to relate the temporal and spatial fluctuations, the one-dimensional (1D) temporal power spectrum P_f recorded at the Earth can be calculated as an integral of multiple 2D "phase screen" $\delta P_{\delta I}(q_x, q_y, z)$ along the z–axis. The spatial spectrum of intensity $\delta P_{\delta I}(q_x, q_y, z)$ carried by an anti-sunward solar wind speed $\mathbf{v}$, which is drifted across an IPS ray-path with its intersection angle $90° - \theta$. Only the speed component perpendicular to the IPS ray-path $V_{\text{drift}} = |\mathbf{v}| \cdot \cos\theta$ is detectable by a terrestrial radio antenna. V_{drift} is the solar wind velocity projected on to the "phase screen" perpendicular to the z–axis. As depicted in Figure 6, the geometry factor $\cos\theta$ varies along the entire IPS ray-path. Supposing V_{drift} in the "phase screen" is along x–axis, the temporal frequency of IPS signal f is derived as $f = q_x V_{\text{drift}}/(2\pi)$, and the temporal IPS spectrum $P(f)$ is derived from the following equations:

$$P(f) = (2\pi r_e \lambda)^2 \frac{1}{|\mathbf{v}|} \delta z \int F_{\text{diff}} F_{\text{source}} S_n\, dq_y \tag{16}$$

for an infinitely thin "phase screen" δz at point p.

$$P(f) = (2\pi r_e \lambda)^2 \int \frac{1}{|\mathbf{v}| \cdot \cos\theta} \int F_{\text{diff}} F_{\text{source}} S_n\, dq_y\, dz \tag{17}$$

for an finite-depth "phase screen" astride point p.

Accordingly, the scintillation index m is related to small-scale density variations along the z–axis by the equation $m^2 \propto \int_0^{+\infty} P(f)\, df$. Because the IPS power spectrum reflects radial speed v and other parameters of the solar wind, a spectral-fitting method such as the Levenberg-Marquardt method for nonlinear least-squares data fitting (Dennis et al., 1981) can be used to extract those parameters from IPS time-series signals recorded at a single station (e.g., Manoharan and Ananthakrishnan, 1990; Tokumaru et al., 2011; Chang et al., 2019). According to Eqs **14–16**, the four fitting parameters include solar wind velocity $\mathbf{v}$, anisotropic axial ratio A_R, power-law spectral index of density fluctuation α, angular size of radio source θ_0. For instance, as shown in Figure 7, a typical IPS power spectrum of radio source 3C48 observed at Toyokawa, Japan in 2009, which is characterized by a flat level at low frequencies and a steep fall at high frequencies, with a knee-like transition point at Fresnel frequency, is well fitted

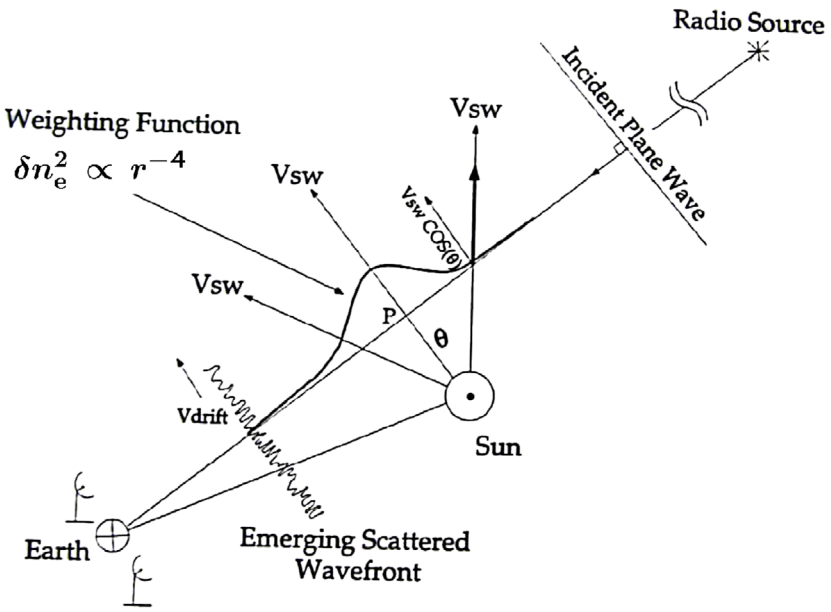


FIGURE 6
Geometry factor cosθ varies along the entire IPS ray-path (Grall, 1995).

to infer $v = 499$ km s^{-1} and $\alpha = 4.34$ (Tokumaru et al., 2011). As shown in Figure 8, a change in velocity scales the spectrum in frequency and the whole spectrum contracts or expands depending on whether the velocity decreases or increases (Manoharan and Ananthakrishnan, 1990). As the signal-to-noise (S/N) ratio of each observed spectrum should be high enough to show a well formed Fresnel knee, Chang et al. (2019) suggested that the minimum S/N ratio for a reliable power spectral fitting is ≥13.5 dB. If the signal-to-noise ratio has become too small, the Fresnel knee in this IPS spectrum is not clearly identified above the instrumental noise floor at high frequencies. Uncertainties of IPS data is generally ascribed to the degenerated S/N ratio of recorded IPS signals, which is probably caused by the unstable receiver temperature and/or the occasional human-made electromagnetic interference. Under conditions of weak scintillation, IPS signals recorded at different frequency channels are usually highly correlated. The co-spectrum of intensity scintillations at either two frequencies can be normalized by the geometric mean of the corresponding two power spectra. Using a single antenna to receive simultaneous three-frequency IPS signals, Scott et al. (1983b) found a clear two-frequency Fresnel filter from the normalized co-spectrum, and estimate the solar wind velocity at elongations 16°–32°. The solar wind velocity inferred from the spectrum-fitting method may have linear change and minute-scale variability (Liu et al., 2022).

The IPS signals can be simultaneously recorded at multiple stations, as shown in Figure 2. The separation of two receiving antennas at the order of 200 km provides an appropriate range of baselines and allows an accurate determination of the spatial properties of the drifting pattern. Because of the Earth's rotation, the baseline projected onto the celestial sphere describes an ellipse and a radio source can be observed with a baseline heavily fore-shortened (Bourgois et al., 1985). For instance, a triangular network of observing sites at Serpukhov, Pereslavl, and Kalinin in Russia was built to record IPS signals from the same radio source (Figure 9). The multi-station scintillation method of measuring the solar wind velocity has been very accurate, particularly near the Sun and at high heliographic latitudes (Kojima et al., 2013). The nearly frozen-in diffraction pattern $\delta P_{\delta I}(q_x, q_y, z)$ can be sequentially received by two radio stations separated by a baseline $\mathbf{s}$. If the baseline $\mathbf{s}$ is approximately parallel to the drifting direction $\mathbf{V}_{\rm drift}$ of interplanetary density irregularities, the scintillation patterns at the two telescopes will be correlated with some time lag τ. Specifically, the spatial-to-temporal conversion is merely a cut in the spatial correlation function along the direction of $(\mathbf{s} - \mathbf{V}_{\rm drift} \cdot \tau)$. When the baseline $\mathbf{s}$ is zero, the cross-correlation function (CCF) between two stations is degenerated to the ACF, which always has a component of uncorrelated white noise at zero lag. For a single scattering layer, the CCF is simply derived by shifting the ACF at a time lag $|\mathbf{s}|/|\mathbf{V}_{\rm drift}|$. Inversely, the drifting speed $|\mathbf{V}_{\rm drift}|$ can be inferred from the time lag τ, which is the manifestation of flow speed of local density irregularities. Thus, the vector velocity $\mathbf{V}_{\rm drift}$ is calculated from three time delays identified along each baseline of joint three-station measurements. However, the calculated flow speed $\mathbf{V}_{\rm drift}$ is a weighted integral along one IPS ray-path. The apparent drifting speed $\mathbf{V}_{\rm drift}$ could be biased from the local solar wind velocity $\mathbf{v}(z)$.

In the Born approximation, the CCF along a baseline $\mathbf{s}$ can be described by the following equation.

$$C(\mathbf{s}, \tau) = 2\pi r_e^2 \lambda^2 \iint F_{\rm diff} F_{\rm source} S_n \exp\left(-i\mathbf{q} \cdot (\mathbf{s} - \mathbf{v}(z)\ \tau)\right) \times d^2\mathbf{q}\, dz \tag{18}$$

The velocity $\mathbf{v}(z)$ probably has two random components: (1) one parallel component expected from the radial variations of solar wind speed in the scattering region; (2) one perpendicular component

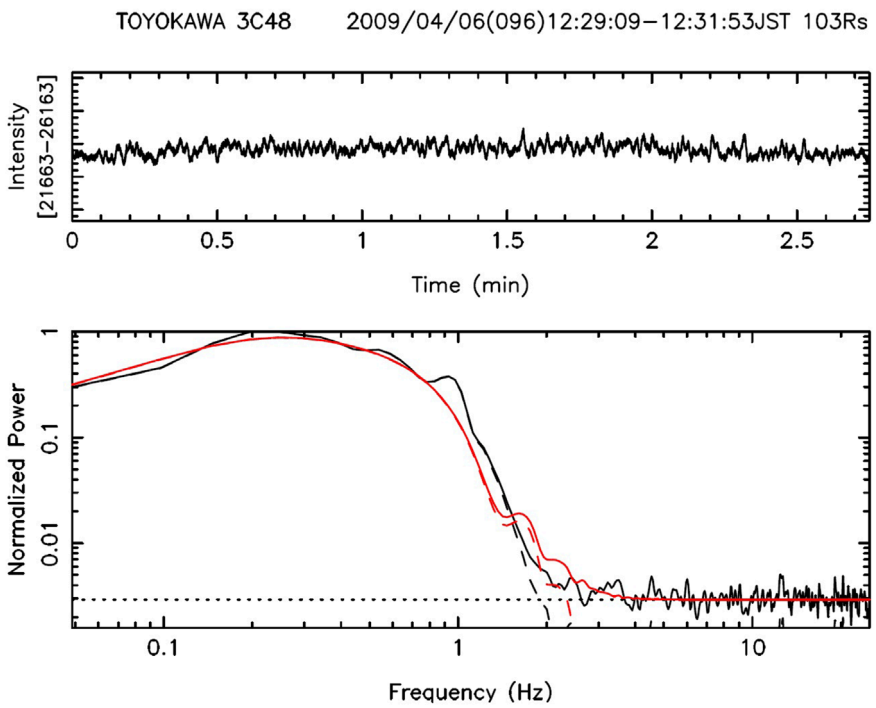


FIGURE 7
A typical time-series of IPS signals from radio source 3C48 recorded by SWIFT at Toyokawa, with its normalized power spectral fitting denoted as red lines (Tokumaru et al., 2011).

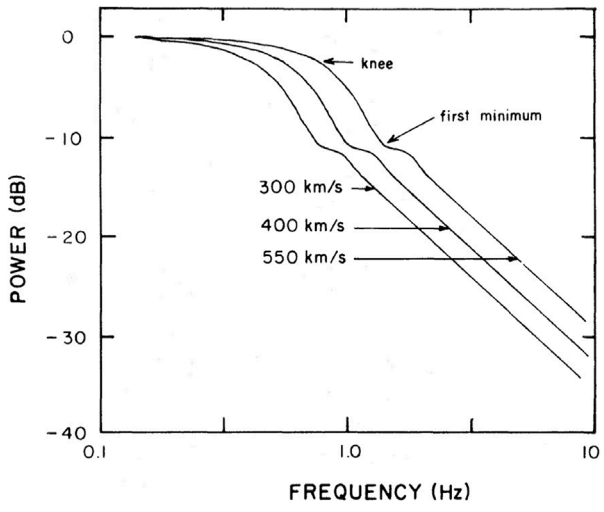


FIGURE 8
An example of power spectral fitting of ORT IPS data for various solar wind speeds (Manoharan and Ananthakrishnan, 1990).

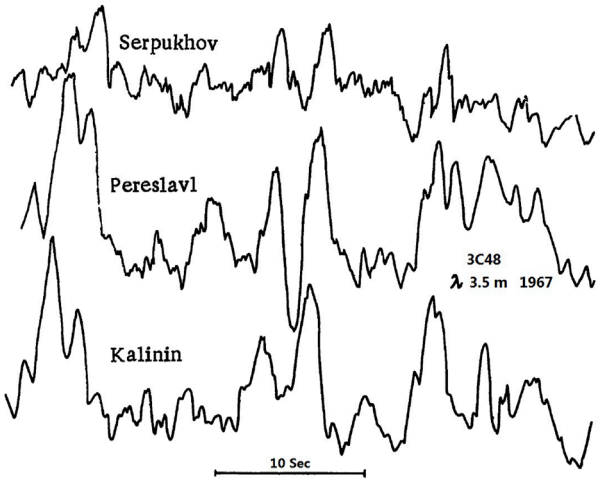


FIGURE 9
An example of simultaneous three-station recording of IPS signals from radio source 3C48 in 1967 (Vitkevich and Vlasov, 1970).

anticipated from radially propagating Alfvén waves. These random components result in an apparent spread of pattern-drifting speed around some mean large-scale bulk-flow speed. Little and Ekers (1971) proposed that the drifting and changing scintillation pattern can be modelled as an ensemble of frozen-in patterns, each having a spatial ACF $C(\mathbf{s},0)$ and a pattern speed $\mathbf{v}$. Then, IPS CCF can be integrated over a probability density function (PDF) $p(\mathbf{v})$ of the form

$$C(\mathbf{s},\tau) = \iint C(\mathbf{s} - \mathbf{v}\,\tau, 0)\; p(\mathbf{v})\; d^2\mathbf{v} \tag{19}$$

A prescription of $p(\mathbf{v})$ can be uniform (Grall et al., 1996) or Gaussian (Kojima et al., 2013). The observed CCF $C(\mathbf{s},\tau)$ are usually normalized by the square root of the product of the two

ACFs at zero lag $\tau = 0$. IPS velocities are computed by making model fits to the shape of the normalized CCF. A long baseline can improve the ability to resolve streams of different velocities (Grall et al., 1996; Breen et al., 2006; 2008; Bisi et al., 2010), provided that the time interval is long enough to derive well-defined scintillation spectra, when the observing geometry is suitable for CCF analysis. Herein, though the baseline between Multi-Element Radio Linked Interferometer Network (MERLIN) and EISCAT network is extremely long to be 2,000 km, correlation between the scintillation patterns observed on 15 May 2002 was still significant (Breen et al., 2006). When an IPS ray-path traverses through both fast and slow solar wind streams, the delay–CCF profile measured by widely separated stations such as EISCAT radar has two clearly separated peaks: one narrow peak at a short delay, the other broad peak at a large delay. If measured by closely spaced stations such as ISEE radio array in Japan, only one peak exists in the delay–CCF profile. As the baseline between two stations is gradually shortened, two distinct peaks would overlap and merge into one compound peak. Multi-streams along an IPS ray-path produce a skewed CCF, but the degree of this skewness depends on the baseline length: a longer baseline produce a stronger skewness (Kojima et al., 2013). The CCF analyses between more closely spaced stations is used to estimate an intermediate velocity perpendicular to one IPS ray-path, which is averaged between multiple solar wind streams cut by the IPS ray-path.

Using a long baseline to de-convolve the LOS integration of CCF measurements, multiple solar wind streams are discernible from the measured delay–CCF profile. As shown in Figure 10, double-peaked CCFs are found from EISCAT observation of south polar solar wind, whose baseline was 240 km radial and 3 km tangential. The fast and slow wind speeds are inferred to be 500 and 777 km s^{-1} on basis of one model fit of the delay–CCF profile (Grall et al., 1996). Various IPS observations using EISCAT, MERLIN, Very Long Baseline Array (VLBA) antennas substantiate that the acceleration of polar solar wind is almost completed within a heliocentric distance of 10 $R_\odot$ (Grall et al., 1996; Klinglesmith, 1997). However, the velocities of the near-Sun solar wind derived from IPS measurements have a large velocity spread. This large velocity spread, mostly aligned along the radial direction within 20 $R_\odot$, is considered to be caused by the Alfvén-wave motion (Harmon and Coles, 2005; Kojima et al., 2013). The dominant contributor to the velocity spread is LOS projection effects rather than changes in wave dispersion near the ion cyclotron frequency (Harmon and Coles, 2005). As the amplitude of Alfvén turbulence is much smaller in the interplanetary space compared with the inner solar corona, IPS speed is likely to be close to the bulk plasma flow speed, at least in the slow solar wind (Klinglesmith, 1997).

The basic solar wind parameters include the mean velocity, random velocity, spatial anisotropy, and micro-turbulence spectrum. Near the Sun, spatial structure is quite anisotropic (Dennison and Blesing, 1972) and the solar wind speed has an important random component (Little and Ekers, 1971). The random velocity is insignificant at distances greater than 40 $R_\odot$, but increases rapidly with decreasing distance, approaching the mean velocity by 20 $R_\odot$ (Scott et al., 1983a). The solar wind velocity can be recursively estimated by fitting the temporal power spectra observed with a single antenna, or straightforwardly calculated by using the time delays between three pairs of radio antennas. A

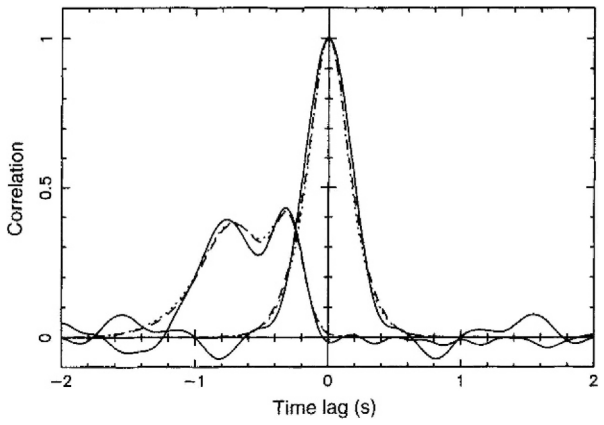


FIGURE 10
Intensity correlation functions from EISCAT observation of radio source 0323 + 055 at −52° helio-latitude on 28 April 1994 (Grall et al., 1996).

comparison between single-station and three-station solar wind measurements (Manoharan and Ananthakrishnan, 1990; Mejia-Ambriz et al., 2015; Chashei et al., 2021) generally showed good agreement within the error estimates. For instance, using nearly simultaneous IPS observations by both BSA LPI and ISEE radio telescopes, Chashei et al. (2021) found the correlation between the daily speed estimates from the compact source 3C48 is 50% during 6 years from 2014 to 2019. All the IPS records studied by Shishov et al. (2010) were selected by the range of radio source elongations between 25° and 60°, which corresponds to heliocentric distances of LOS p–points within the range of 0.4 and 0.8 AU. However, the single-station model-fitting process was found to be difficult for solar distance less than 40 $R_\odot$, because in this range neither random velocity nor spatial anisotropy can be neglected (Scott et al., 1983a). The multi-station joint observation can separate the similar effects of random velocity and anisotropy. The models developed with the multi-station data can be extrapolated to the entire class of single station observations (Scott et al., 1983a). The solar wind speed vector derived from the multi-station cross-correlation analyses is valid for any strength of scattering and is independent of the existence of turbulence in the velocity field (Bourgois et al., 1985; Chashei et al., 2021). A comparison between the single-station and multi-station methods of velocity estimation is essential to examine a possible systematic bias between them and thus to improve the accuracy of IPS observation (Tokumaru et al., 2011), which is routinely performed for the ISEE IPS observations.

6 IPS-CAT modelling for space weather prediction

The 3D solar wind flows are remotely sampled at multiple discrete IPS ray-paths over a large portion of Sun-centered sky map. To optimize the use of IPS observations and produce 3D global heliospheric representations, two iterative CAT models were separately developed in late 1990s (Jackson et al., 1998; Kojima et al., 1998) that fit IPS observations of drifting speed

$V_{\rm drift}$ and scintillation factor g to a kinematic solar wind model to provide 3D velocities and densities. The general flowchart for numerical IPS-CAT models is plotted in Figure 11, with the same notations as used in the text of Jackson et al. (2010c). The subscripts j and k are used when physical quantities such as density n and speed $\mathbf{v}$ refer to discrete $N_\theta \times N_\varphi$ grids on a 2D reference sphere. The spatial resolution of current CAT models along latitude and longitude directions is typically at 5°–10° (Hayashi et al., 2003; Jackson and Hick, 2004; Hayashi et al., 2016). The reference sphere, also called source surface, is usually set at the super-Alfvénic region, i.e., $r_{\rm ref} = 30R_s$. Because the realistic solar wind flows are already accelerated at the inner boundary of $r_{\rm ref}$, the complicated mechanisms governing tran-Alfvénic acceleration and heating of solar wind flows are absent within the simulation domain of IPS-CAT models. An initial prescription of the inner boundary condition can be arbitrary, as a unique solution of solar wind speed distribution on the inner boundary is gradually converged as a result of repeated perturbation of radial speed $\delta v_{r,j,k}$ at the boundary mesh.

The essence of IPS-based CAT models is an automatically iterative procedure modifying the solar wind velocity distribution on an inner boundary sphere. The boundary-driven 3D CAT models can be ballistic, kinematic, or magneto-hydrodynamical according to physical description of solar wind flows. For a ballistic model, the solar wind plasma velocity is assumed constant along the streamlines (Kojima et al., 1998). For a kinematic model, the momentum and mass are conserved to describe the collision and merge of different solar wind plasma parcels emanating from the inner boundary (Jackson et al., 1998). For an MHD model, nonlinear MHD processes are self-consistently simulated to address very complicated interplanetary dynamics from near the Sun to well beyond 1 AU, such as shock formation. The mathematical treatment of the solar wind flows in IPS-CAT models is increasingly sophisticated from the ballistic model to the kinematic model, and then to the MHD model. Using any of these IPS-CAT models, 3D volumetric data of speed $\mathbf{v}_{i,j,k}$ is numerically reconstructed.

The simulation of IPS velocity measurement $V^{\rm mdl}_{{\rm drift},\,\mu}$ in numerical solar wind is performed for multiple IPS ray-paths ($\mu = 1,\ldots,N_{\rm obs}$). The IPS simulation is made with the integration (Kojima et al., 1998; Hayashi et al., 2003)

$$V^{\rm mdl}_{{\rm drift},\,\mu} = \int \bar{\omega}(z)\ \delta n_{\rm e}^2(z)\ |\mathbf{v}(z)\times d\mathbf{z}| \Big/ \int \bar{\omega}(z)\ \delta n_{\rm e}^2(z)\, dz$$
$$\equiv \int W(z)\ \delta n_{\rm e}^2(z)\ |\mathbf{v}(z)\times d\mathbf{z}| \qquad (20)$$

where $\mathbf{z}$ is the position vector of a point on the ray-path at the distance z from Earth. The weighting factor $\bar{\omega}(z)$ is calculated as an integral over spatial wave vector $\mathbf{q}$, i.e., $\bar{\omega}(z) = \iint w(\mathbf{q})\, d^2\mathbf{q} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(q_x, q_y)\, dq_x\, dq_y$ with $w(\mathbf{q})$ term given in Eq. **13**. The weight $\bar{\omega}(z)$ is normalized to be $W(z)$. The normalized weighting factor of segments along each ray-path $W_{\nu,\mu}, \nu = 1,\ldots,N_{\rm los}$ is pre-calculated. The observation-modelling comparisons of drifting speed and scintillation factor are evaluated as $\Delta V_{{\rm drift},\,\mu} = V^{\rm mdl}_{{\rm drift},\,\mu} - V^{\rm obs}_{{\rm drift},\,\mu}$ and $\Delta(g^2)_\mu = (g^2)^{\rm mdl}_\mu - (g^2)^{\rm obs}_\mu$ for all IPS ray-paths $\mu = 1,\ldots,N_{\rm obs}$. The discrepancy between the actual and simulated IPS velocities $(\Delta V_{{\rm drift},\mu}, \mu = 1,\ldots,N_{\rm obs})$ is examined for its standard deviation $\sigma(\Delta V)$. A provisional analysis with several iterations is carried out to discard the outliers of bad IPS velocity data, whose discrepancy is larger than a threshold of $|\Delta V_{{\rm drift},\mu}| \geq 1.5\sigma(\Delta V)$ (Hayashi et al., 2003).

With the 3D distribution of $\mathbf{v}_{i,j,k}$ within the simulation domain, IPS data made at various heliocentric distances and times are traced back along the streamlines to the source positions on the inner boundary. In Figure 11, subscript μ is used to identify a LOS while ν refers to a segment at a certain distance from the observer along the LOS. On the footpoints backwards mapped from each IPS ray-path, radial speed is perturbed to be $\delta v_{r,\mu,\nu}$ for $\nu = 1,2,\ldots,N_{\rm los}$ and $\mu = 1,2,\ldots,N^*_{\rm obs}$. The $\delta v_{r,\mu,\nu}$ and $\delta n_{\mu,\nu}$ at the streamline footpoints are used to modify surrounding speed and density on the $j-k$ mesh. The conversion from $(\delta v_{r,\mu,\nu}, \delta n_{\mu,\nu})$ to $(\delta v_{r,j,k}, \delta n_{j,k})$ is in a Gaussian spread function of equal-solar-surface areas to improve the stability of the iteration process (Jackson et al., 2010c). The differences between actually observed and calculated model values are gradually minimized via the iterative reconstruction. As a rate of convergence, $\Delta V_{\rm drift}$ rapidly decreases to 10% or better level of the average IPS velocity $V^{\rm obs}_{\rm drift}$ (Hayashi et al., 2003). The LOS-integrated IPS observations of the solar wind speed and density turbulence level are de-convolved by the CAT procedure to reconstruct the 3D solar wind speed (and density) structures.

IPS data-constrained numerical CAT 3D models can reconstruct the transient CME structures and their ambient large-scale solar wind flows. The Carrington maps of velocity and density at the inner boundary are smoothed each iteration using a 2D Gaussian spatial filter and a 1D Gaussian temporal filter (Jackson et al., 2010c). Such a temporal filter on the inner boundary is not used to reconstruct the global structures of corotating solar wind. For a corotating IPS-CAT model, the global solar wind is assumed to be slowly evolving in time during one Carrington rotation. The absolute density scale of the solution cannot be determined from the tomographic reconstruction itself, because g–level is a proxy for the solar wind density variance. The averaged scale of reconstructed density is externally calibrated according to *in situ* measurements near the Earth. For the UCSD kinematic IPS-CAT model, g–levels can be replaced or augmented by heliospheric Thomson-scattering intensity as measured by the Solar Mass Ejection Imager in a polar orbit around the Earth (Jackson et al., 2010c). As shown in Figure 12A, IPS scintillation indices calculated from the 3D density distribution of numerical MHD simulations are useful in denoting CME propagation direction on the plane of sky (Iwai et al., 2021). As shown in Figure 12B, the 3D numerical solution of MHD IPS-CAT model has been demonstrated to be able to simulate the Archimedean spiral structure of the solar wind on the equatorial plane, and match the *in situ* near-Earth and/or Ulysses observation data during one Carrington rotation period (Hayashi et al., 2016). The 3D global structures of interplanetary disturbances, daily reconstructed from the UCSD kinematic IPS-CAT model, are shown in the website of https://ips.ucsd.edu/ for the purpose of real-time space weather forecast. One of the data products generated from the tomographic reconstruction is the large-scale distribution of 3D density structures shown in Figure 12C (Jackson and Hick, 2004).

Regarding multiple iteration steps, there is a major difference between the UCSD time-dependent kinematic IPS-CAT model and other MHD IPS-CAT models. The former goes through a Carrington rotation or any time period in a single "time-dependent" run (or iteration). In other words, each iteration is a time-dependent simulation by itself. On the other hand, the latter performs

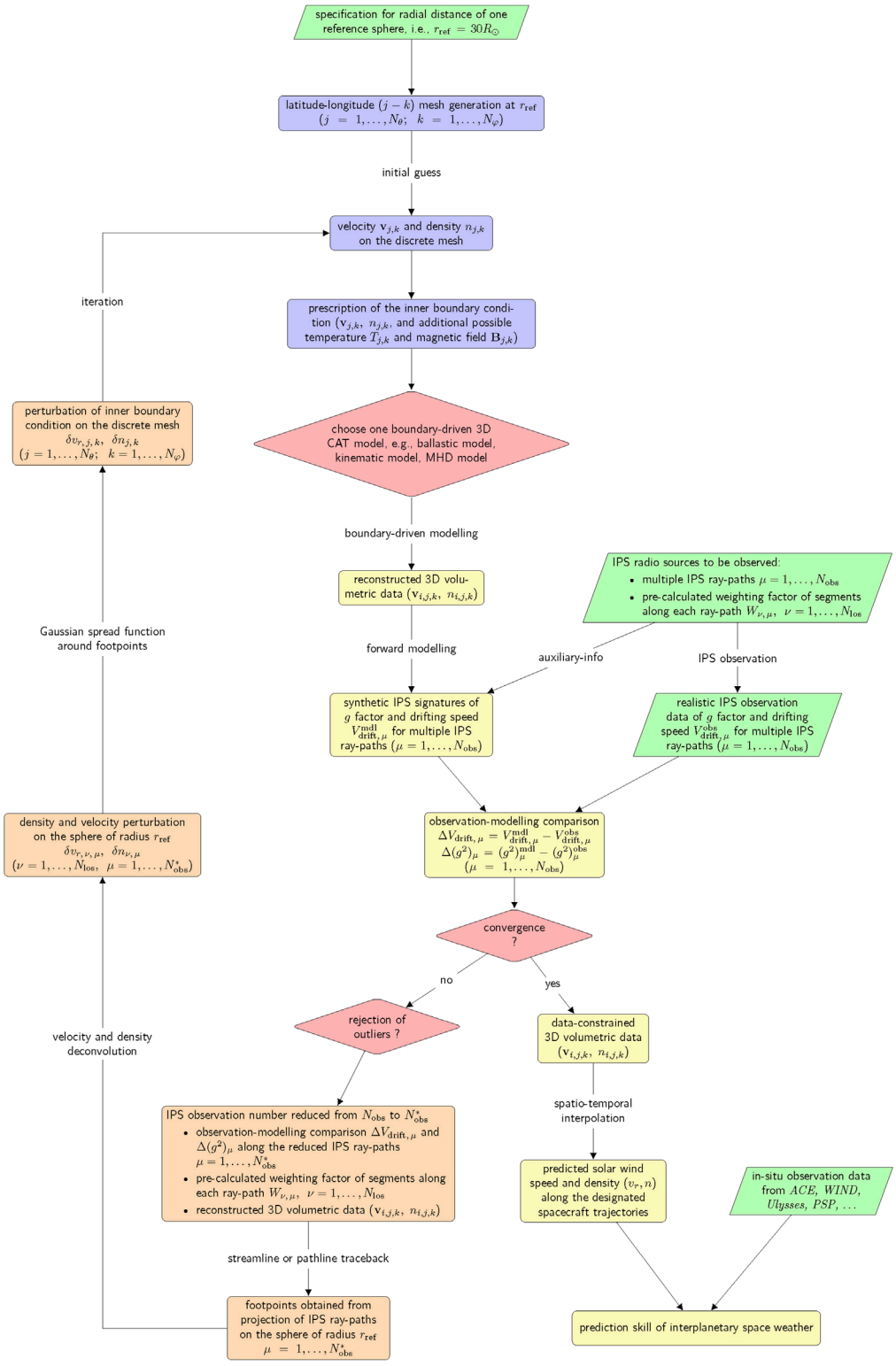


FIGURE 11
The general flowchart for numerical CAT models of the inner heliosphere constrained by IPS data.

tomographic iteration for a steady state simulation at a certain time cadence (typically 1 day), and then a time-dependent MHD simulation is performed afterwards using the iteratively constrained boundary conditions. Streamlines are imaginary lines that represent the direction of the flowing fluid at a certain point in time. Only in a steady flow are streamlines identical to pathlines, which are defined to be flow paths that the fluid particles take while flowing. The above-mentioned difference between different tomography

models is essentially ascribed to the inconsistency between pathlines (also called trajectories) and streamlines in an unstable flow.

The UCSD kinematic IPS-CAT model can accurately reproduce the general large-scale 3D structure of the solar wind density and speed in the inner heliosphere, and even smaller-scale fluctuations at 6-h cadence at Earth (Jackson et al., 2015; Jackson et al., 2020). Obviously, all the IPS ray-paths converge at Earth, so near-Earth space in the IPS-CAT model is always very well constrained by IPS data. The *in situ* velocity measurements at the first Sun-Earth Lagrangian point (L1), such as near-real time Advanced Composition Explorer (ACE) data, are further included into the UCSD time-dependent tomography to constrain the iterated results (Jackson et al., 2010b). As a result, the current "L1-data constrained" kinematic IPS tomography performs remarkably at Earth, with Pearson correlations typically above 0.9. In addition, the 3D solar wind velocity reconstructed from the UCSD kinematic modelling can be used to carry and extend the solar magnetic field lines, described by the current sheet source surface model (Zhao and Hoeksema, 1995), out to the edge of the global boundary considered by the IPS analysis (Jackson et al., 2012). Such IPS-aided information about disturbed solar wind structures in three dimensions is valuable in interpreting *in situ* observations from interplanetary spacecraft.

The inner boundary conditions provided by the UCSD time-dependent tomography have been used to drive multiple numerical MHD models, such as MS-FLUKSS model (Kim et al., 2014), ENLIL model (Jackson et al., 2010a), Alfvén Wave Solar atmosphere Model (AWSoM) (Sachdeva et al., 2019). An *ad hoc* adjustment of the inner boundary values in the MS-FLUKSS MHD model could deliver more comparable results at Earth, but diverges considerably from the UCSD kinematic IPS-CAT model below 100 $R_\odot$ (Kim et al., 2014). Sachdeva et al. (2019) compared the AWSoM MHD model with the UCSD kinematic IPS tomography at 20, 100, and 215 $R_\odot$, and found reasonable agreements at 100 and 215 $R_\odot$ and a 20–30% discrepancy at 20 $R_\odot$; Jian et al. (2015) compared all the coronal and heliospheric model combinations (including MHD models) installed at the NASA Community Coordinated Modeling Center (CCMC) and found that the kinematic IPS tomography correlated best with the solar wind speed and density measured at Earth. On the other hand, when the same CCMC models were compared with Ulysses data during the fast latitudinal scan in 2007, the kinematic IPS tomography deviated significantly when Ulysses was away from Earth and the ecliptic plane (Jian et al., 2016), confirming what Kim et al. (2014) showed earlier. The performance drawback of the UCSD kinematic IPS-CAT model at Ulysses orbit is owing to two reasons that "the numbers of IPS signals are limited out of the ecliptic plane at present" and "IPS observations over the north and south ecliptic poles are generally obtained close to the solar surface" (Jian et al., 2016). The outstanding mystery of the solar wind acceleration nearer the Sun is insufficiently addressed in the mathematical descriptions of current IPS-CAT models, which results in the degenerated performance of kinematic IPS-CAT model at Ulysses orbit (Jian et al., 2016) and the noticeable prediction discrepancy between different IPS-CAT models at 20 $R_\odot$ (Sachdeva et al., 2019).

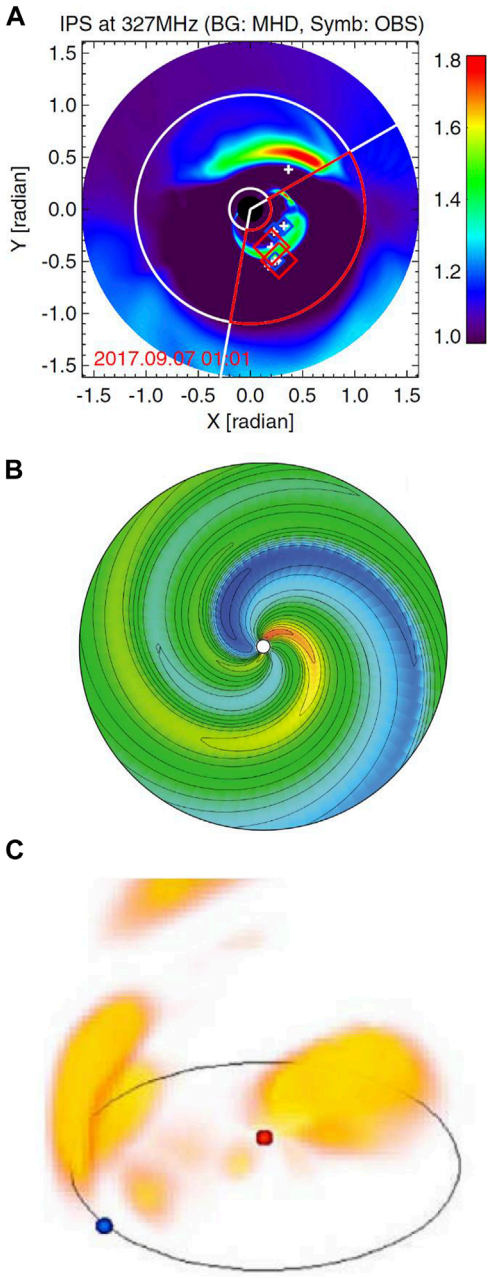


FIGURE 12
Some IPS-related numerical 3D models for space weather modelling: **(A)** multi-source observations of IPS *g*–value and their comparison with the SUSANOO-CME MHD simulation (Iwai et al., 2021), **(B)** radial speed of solar wind on the equatorial plane derived by the MHD-IPS Hayashi model (Hayashi et al., 2016), **(C)** 3D distribution of the corotating component of plasma density in the inner heliosphere during Carrington rotation 1884 reconstructed by the UCSD IPS-CAT model (Jackson and Hick, 2004). In panel **(A)**, the cross symbols represent all observed radio sources, and the diamond ones indicate radio sources with the *g*-values of $g > 2$ (red), $1.5 < g < 2.0$ (green), and $1.2 < g < 1.5$ (blue). In panel **(B)**, blue, yellow, and orange colors approximately represent radial speed of 700, 500, and 300 km s^{-1}, respectively. In panel **(C)**, the density is normalized by the removal of its r^{-2} dependence on heliocentric distance *r*.

The differences between the IPS kinematic-model data-fitting procedure and the current 3D MHD modeling techniques provide interesting insights into the physical principles governing the expulsion of CMEs as well as expectations and limitations of various IPS-CAT models (e.g., Jackson et al., 2015; Yu et al., 2015).

7 International cooperation within the worldwide IPS stations network

The intensity variance of IPS signals on the time-scale of 0.1–10 s for the micro-scale density irregularities, with a characteristic scale of tens to a few hundred of kilometers. The sensitivity of radio telescope is characterized as system noise temperature T_{sys}, effective aperture area A_e, and minimum detectable flux density ΔS_{min}. A high-sensitivity radio system to observe celestial compact radio sources is desired, because the flux of IPS signals received at Earth are generally faint to be less than 1 Jansky. Due to increasing of radio telescope sensitivity, much more fainter radio sources are found to have the IPS phenomena. However, man-made radio emissions in all frequency bands are increasingly causing serious interference signals for most of radio astronomical observatories in the world. Under the guidance of International Telecommunication Union, radio frequency spectra are painstakingly parcelled out between radio-based applications such as personal communications, satellite broadcasting, GPS and amateur radio, and the sciences of radio astronomy, Earth exploration and deep-space research. The key bands of VHF and UHF radio spectra widely used in IPS observations should be carefully protected by the local administrative regulations. As shown in Figure 2, the observing frequencies for the current IPS-dedicated radio telescopes are 327 MHz in India and Japan, 111 MHz in Russia, 140 MHz in Mexico.

The numerical IPS-CAT models have been proved to be capable of reconstructing 3D large-scale velocity and density structures in the inner heliosphere. Such a 3D reconstruction of interplanetary structures is of great interest for the space weather prediction community. Since the routine IPS observations at the ISEE in 1999, a near real-time prediction analysis system on basis of UCSD kinematic IPS-CAT model have been operated and developed for over two decades to provide heliospheric solar wind speed and density at Earth (Jackson et al., 2010b) and at Mars (Jackson et al., 2007). The construction of SWIFT at Toyokawa in the late 2000s significantly upgraded the overall sensitivity of the ISEE IPS system, and achieved a finer resolution in the CAT reconstruction from those IPS data (Tokumaru et al., 2011; Hayashi et al., 2016). The time cadence interval used by the UCSD IPS-CAT modelling is relatively short in comparison with that of a solar rotation, i.e., about 1 day for ISEE IPS data (Jackson et al., 2015). The spatial and temporal resolutions of IPS-CAT models are essentially constrained by the number of available IPS ray-paths.

The insufficient sampling of the inner heliosphere provided by the existing observations of IPS can be ascribed to several reasons including limited sky coverage in declination and hour-angle, insufficient sensitivity, limitations posed by weather and geographic locations, and slow slew rate of antenna beams (Oberoi and Benkevitch, 2010). For instance, the SWIFT antenna at Toyokawa forms a single beam in the local meridian, steerable between 60°S and 30°N with respect to the zenith (Tokumaru et al., 2011), which results in sparse polar coverage, especially in the southern hemisphere. The solar wind speeds and scintillation levels have been basically determined on a daily basis between April and December from ISEE IPS observations. However, the ISEE IPS data are unavailable in winter owing to snow at the observatories (Tokumaru et al., 2021), and has noticeable coverage gaps in certain parts of the sky because of the unevenly scattered IPS sources. An extensive gap in the ISEE IPS tomography map data due to the above-mentioned reasons had better be filled by an analytic Legendre-polynomial formula for the latitudinal profiles of the yearly filtered solar wind speed (Porowski et al., 2022). As IPS observations are usually taken around the time of meridian transit for each radio source, daily meridian scan observations of the sky map around the Sun can be sequentially made at multiple IPS stations located at different geo-longitudes. The coordinated observation of multiple IPS-dedicated stations distributed at different time zones on the Earth is highly desired to improve the routine prediction performance of IPS-based tomographic reconstruction.

The Worldwide IPS Stations (WIPSS) Network aims to bring together the worldwide real-time capable IPS observatories with well-developed and tested analysis techniques (Bisi et al., 2019). Now, the real-time internet access to the original IPS observation data is only provided by the ISEE IPS system among these IPS-dedicated radio telescopes. Other IPS-dedicated radio telescopes at Ooty and Pushchino have accumulated a huge volume of IPS data since their operations in the 1970s. The Ooty radio telescope is unique for its long axis of parabolic cylinder reflector being parallel to the Earth's rotation axis. Using such an equatorial mount, a celestial radio source can be tracked for about 10 h by the mechanical rotation of the cylinder in the east-west direction (Manoharan, 2010). Some non IPS-dedicated radio telescopes such as EISCAT (Bourgois et al., 1985), LOFAR (Fallows et al., 2016), and MWA (Morgan et al., 2018) have been intermittently used to study IPS phenomena. The southern-biased FOV of MWA provides a good complement to the northern-biased FOV of other telescopes located at the northern hemisphere of the Earth. The LOFAR4SW project, implemented under the European Horizon 2020 programme as a design study, was to develop an upgrade to the existing LOFAR infrastructure so that it would be possible to use the instrument's capabilities for space weather purposes (Bisi et al., 2020; Carley et al., 2020). In addition, the new Chinese IPS-dedicated radio telescopes under construction in northern China will join the WIPSS network for the international space weather service. The daily observed radio sources are estimated to be nearly 100 by SWIFT (Iwai et al., 2021), 500 by MEXART (Mejia-Ambriz et al., 2010), and several thousand by ORT (Manoharan et al., 2017) and BSA LPI (Chashei et al., 2021). Different IPS techniques (single-site versus multi-site) and analysis procedures followed at different observatories need to be cross-calibrated against one another to yield consistent results (Oberoi and Benkevitch, 2010). Jackson et al. (2023) demonstrated that both spatial and temporal coverage of UCSD kinematic CAT model can be significantly increased with more IPS data input from ISEE, LOFAR, and BSA LPI within the WIPSS network groups.

8 Summary and discussion

IPS is the phenomenon of random fluctuations in the intensity of distant compact object, smaller than an arc-second, induced by the ubiquitous inhomogeneous structures in the turbulent solar wind. The radio signal from a distant compact radio source is scattered by the density inhomogeneities of the solar wind on the order of 100 km. A diffraction pattern of radio intensity fluctuations arises from the interference between any pair of incident plane waves. The spatial correlation length of density inhomogeneities and the Fresnel radius of radio diffraction are two key parameters in determining the scintillation pattern. In weak scattering, a "2D phase-chaning screen" scenario is justified on the basis of the Born approximation. The diffractive contribution of the scattered signal is quantified as a Fresnel filter in weak-scattering theory. In strong scattering, scintillation index decreases because intensity variation interferes or cancels each other within the observation frequency band. A typical power spectra of IPS signals recorded at a single station is flat at low frequencies, then has a sudden drop over the Fresnel knee. Using a nonlinear spectral-fitting method, the IPS power spectra around the knee is analyzed to infer solar wind velocity, radio source size, and turbulence spectral index. The intensity diffraction pattern that transported away from the Sun by the solar wind can be measured and correlated between multiple radio stations on the Earth. Two simultaneous IPS time series from separate stations can be cross-correlated to determine the solar wind speed from time delay. Using three stations for IPS allows the solar wind velocity vector to be determined to a high degree of accuracy. Therefore, IPS serves as an effective tool for remote sensing of the inner heliosphere, determining the flow speed and turbulence amplitude associated with either undisturbed solar wind or transients, such as CMEs.

The global 3D distribution of solar wind speed and density can be determined by current IPS-based numerical CAT models. The simulation domain of those IPS-CAT models are generally prescribed beyond the Alfvén surface of solar wind flows. Tokumaru et al. (2021) compared solar wind speeds derived from the IPS-CAT analysis with *in situ* observations conducted by the near-Earth and Ulysses spacecraft, and found that the discrepancy between the IPS and *in situ* observations can be improved by changing the power index of the empirical relation between the solar wind speed $\mathbf{v}$ and density fluctuations δn_e. The real-time IPS predictions often suffer reduced performance since they are limited to the data available only up to the time when the prediction is made (Jackson et al., 2015; Jackson et al., 2020; Jackson et al., 2023). With a large number of IPS radio sources to be observed by the WIPSS network on a regular daily basis, IPS-CAT models can be fully exploited to map and predict the structure and motion of major disturbances in the inner heliosphere. The potential synergy between the above-mentioned IPS-capable radio telescopes on the Earth, complemented with other ground-based and spaceborne instruments, would enable the simultaneous exploration and scientific understanding of solar wind dynamics and structures in three dimensions, thereby facilitating major breakthroughs in the interplanetary space weather prediction.

Author contributions

MX was responsible for the organization of this article, contributed to all sections, and wrote the initial manuscript version. KY provided the important insights on how to design figures 2 and 11, and made major contributions to the revision of sections 2. All authors contributed to the initial concept, read and revised the paper, approved the final version.

Funding

This work was jointly supported by the National Natural Science Foundation of China (41874205, 42074208, 42241118, 42030204, 41974200, 42174201, U2031134, 12003049, 12203056), National Key R&D Program of China (2021YFA1600500, 2021YFA1600503), the Specialized Research Fund for State Key Laboratories of China, the Meridian Space Weather Monitoring Project in China.

Acknowledgments

We sincerely thank both reviewers for their constructive comments and suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abe Pacini, A. (2020). "The Arecibo NOON (coroNa tO sOlar wiNd) project," in *AGU fall meeting abstracts*, 2020, SH041–09.

Armstrong, J. W., and Coles, W. A. (1972). Analysis of three-station interplanetary scintillation. *J. Geophys. Res. Space Phys.* 77, 4602–4610. doi:10.1029/JA077i025p04602

Asai, K., Kojima, M., Tokumaru, M., Yokobe, A., Jackson, B. V., Hick, P. L., et al. (1998). Heliospheric tomography using interplanetary scintillation observations 3. Correlation between speed and electron density fluctuations in the solar wind. *J. Geophys. Res. Space Phys.* 103, 1991–2001. doi:10.1029/97JA02750

Bale, S. D., Goetz, K., Harvey, P. R., Turin, P., Bonnell, J. W., Dudok de Wit, T., et al. (2016). The FIELDS instrument suite for solar Probe plus. Measuring the coronal plasma and magnetic field, plasma waves and turbulence, and radio signatures of solar transients. *Space Sci. Rev.* 204, 49–82. doi:10.1007/s11214-016-0244-5

Barnes, A. (1979). "Hydromagnetic waves and turbulence in the solar wind,". *Solar system plasma physics*. E. N. Parker, C. F. Kennel, and L. J. Lanzerotti, 1, 249–319.

Bastian, T. S. (1994). Angular scattering of solar radio emission by coronal turbulence. *Astrophys. J.* 426, 774. doi:10.1086/174114

Bisi, M. M., Fallows, R. A., Breen, A. R., and O'Neill, I. J. (2010). Interplanetary scintillation observations of stream interaction regions in the solar wind. *Sol. Phys.* 261, 149–172. doi:10.1007/s11207-009-9471-1

Bisi, M. M., Jackson, B. V., Gonzalez-Esparza, A., Aguilar-Rodriguez, E., Tokumaru, M., Fallows, R. A., et al. (2019). "The worldwide interplanetary scintillation (IPS) stations (WIPSS) network: A ground-based heliospheric observatory for science and space weather," in *AGU fall meeting abstracts*, 2019. SA11C–3233.

Bisi, M. M., Ruiter, M., Fallows, R. A., Vermeulen, R., Robertson, S. C., Vilmer, N., et al. (2020). "LOFAR4SpaceWeather (LOFAR4SW) - increasing European space-weather capability with Europe's largest radio telescope: Beyond the detailed design review (DDR)," in *EGU general assembly conference abstracts* (Vienna: EGU General Assembly Conference Abstracts), 14948. doi:10.5194/egusphere-egu2020-14948

Bourgois, G., Daigne, G., Coles, W. A., Silen, J., Turunen, T., and Williams, P. J. (1985). Measurements of the solar wind velocity with EISCAT. *Astron. & Astrophys.* 144, 452–462.

Breen, A. R., Fallows, R. A., Bisi, M. M., Thomasson, P., Jordan, C. A., Wannberg, G., et al. (2006). Extremely long baseline interplanetary scintillation measurements of solar wind velocity. *J. Geophys. Res. Space Phys.* 111, A08104. doi:10.1029/2005JA011485

Breen, A. R., Fallows, R. A., Bisi, M. M., Jones, R. A., Jackson, B. V., Kojima, M., et al. (2008). The solar eruption of 2005 May 13 and its effects: Long-baseline interplanetary scintillation observations of the Earth-directed coronal mass ejection. *Astrophys. J. Lett.* 683, L79–L82. doi:10.1086/591520

Bruno, R., and Carbone, V. (2013). The solar wind as a turbulence laboratory. *Living Rev. Sol. Phys.* 10. doi:10.12942/lrsp-2013-2

Burlaga, L. F., Scudder, J. D., Klein, L. W., and Isenberg, P. A. (1990). Pressure-balanced structures between 1 AU and 24 AU and their implications for solar wind electrons and interstellar pickup ions. *J. Geophys. Res. Space Phys.* 95, 2229–2239. doi:10.1029/JA095iA03p02229

Burnell, J. (1969). Enhancements of interplanetary scintillation, corotating streams and Forbush decreases. *Nature* 224, 356–357. doi:10.1038/224356a0

Cane, H. V., Sheeley, N. R., and Howard, R. A. (1987). Energetic interplanetary shocks, radio emission, and coronal mass ejections. *J. Geophys. Res. Space Phys.* 92, 9869–9874. doi:10.1029/JA092iA09p09869

Carley, E. P., Baldovin, C., Benthem, P., Bisi, M. M., Fallows, R. A., Gallagher, P. T., et al. (2020). Radio observatories and instrumentation used in space weather science and operations. *J. Space Weather Space Clim.* 10, 7. doi:10.1051/swsc/2020007

Chang, O., Bisi, M. M., Aguilar-Rodriguez, E., Fallows, R. A., Gonzalez-Esparza, J. A., Chashei, I., et al. (2019). Single-site IPS power spectra analysis for space weather products using cross-correlation function results from EISCAT and MERLIN IPS data. *Space weather.* 17, 1114–1130. doi:10.1029/2018SW002142

Chashei, I. V., Lukmanov, V. R., Tyul'bashev, S. A., and Tokumaru, M. (2021). Comparison of solar wind speed estimates from nearly simultaneous IPS observations at 327 and 111 MHz. *Sol. Phys.* 296, 63. doi:10.1007/s11207-021-01804-6

Cohen, M. H., Gundermann, E. J., Hardebeck, H. E., and Sharp, L. E. (1967). Interplanetary scintillations. II observations. *Astrophys. J.* 147, 449. doi:10.1086/149028

Coles, W. A., and Harmon, J. K. (1978). Interplanetary scintillation measurements of the electron density power spectrum in the solar wind. *J. Geophys. Res. Space Phys.* 83, 1413–1420. doi:10.1029/JA083iA04p01413

Coles, W. A., and Harmon, J. K. (1989). Propagation observations of the solar wind near the Sun. *Astrophys. J.* 337, 1023. doi:10.1086/167173

Coles, W. A., and Kaufman, J. J. (1978). Solar wind velocity estimation from multi-station IPS. *Radio Sci.* 13, 591–597. doi:10.1029/RS013i003p00591

Coles, W. A. (1978). Interplanetary scintillation. *Space Sci. Rev.* 21, 411–425. doi:10.1007/BF00173067

De Zeeuw, D. L., Gombosi, T. I., Groth, C. P. T., Powell, K. G., and Stout, Q. F. (2000). An adaptive MHD method for global SpaceWeather simulations. *IEEE Trans. Plasma Sci.* 28, 1956–1965. doi:10.1109/27.902224

DeForest, C. E., Howard, T. A., and Tappin, S. J. (2011). Observations of detailed structure in the solar wind at 1 AU with STEREO/HI-2. *Astrophys. J.* 738, 103. doi:10.1088/0004-637X/738/1/103

DeForest, C. E., Matthaeus, W. H., Howard, T. A., and Rice, D. R. (2015). Turbulence in the solar wind measured with comet tail test particles. *Astrophys. J.* 812, 108. doi:10.1088/0004-637X/812/2/108

Dennis, J., Gay, D., and Welsch, R. (1981). Algorithm 573: An adaptive nonlinear least-squares algorithm. *ACM Trans. Math. Softw.* 7, 367–383. doi:10.1145/355958.355966

Dennison, P. A., and Blesing, R. G. (1972). Coronal broadening of the crab nebula 1969-71: *Publications of the Astronomical Society ofAustralia* 2, 86–88. doi:10.1017/S1323358000012959

Dennison, P. A., and Hewish, A. (1967). The solar wind outside the plane of the ecliptic. *Nature* 213, 343–346. doi:10.1038/213343a0

Fallows, R. A., Bisi, M. M., Forte, B., Ulich, T., Konovalenko, A. A., Mann, G., et al. (2016). Separating nightside interplanetary and ionospheric scintillation with LOFAR. *Astrophys. J.* 828, L7. doi:10.3847/2041-8205/828/1/L7

Fallows, R. A., Iwai, K., Jackson, B. V., Zhang, P., Bisi, M. M., and Zucca, P. (2023). Application of novel interplanetary scintillation visualisations using LOFAR: A case study of merged CMEs from september 2017. *Adv. Space Res.* in press. doi:10.1016/j.asr.2022.08.076

Feng, X., Yang, L., Xiang, C., Wu, S. T., Zhou, Y., and Zhong, D. (2010). Three-dimensional solar wind modeling from the Sun to Earth by a SIP-CESE MHD model with a six-component grid. *Astrophys. J.* 723, 300–319. doi:10.1088/0004-637X/723/1/300

Feng, X. (2020). *Magnetohydrodynamic modeling of the solar corona and heliosphere*. Singapore: Springer. doi:10.1007/978-981-13-9081-4

Forbes, T. G., Linker, J. A., Chen, J., Cid, C., Kóta, J., Lee, M. A., et al. (2006). CME theory and models. *Space Sci. Rev.* 123, 251–302. doi:10.1007/s11214-006-9019-8

Gary, S. P. (1986). Low-frequency waves in a high-beta collisionless plasma: Polarization, compressibility and helicity. *J. Plasma Phys.* 35, 431–447. doi:10.1017/S0022377800011442

Grall, R. R., Coles, W. A., Klinglesmith, M. T., Breen, A. R., Williams, P. J. S., Markkanen, J., et al. (1996). Rapid acceleration of the polar solar wind. *Nature* 379, 429–432. doi:10.1038/379429a0

Grall, R. R. (1995). *Remote sensing observations of the solar wind near the Sun*. San Diego: Ph.D. thesis, University of California.

Harmon, J. K., and Coles, W. A. (2005). Modeling radio scattering and scintillation observations of the inner solar wind using oblique Alfvén/ion cyclotron waves. *J. Geophys. Res. Space Phys.* 110, A03101. doi:10.1029/2004JA010834

Hayashi, K., Kojima, M., Tokumaru, M., and Fujiki, K. (2003). MHD tomography using interplanetary scintillation measurement. *J. Geophys. Res. Space Phys.* 108, 1102. doi:10.1029/2002JA009567

Hayashi, K., Tokumaru, M., and Fujiki, K. (2016). MHD-IPS analysis of relationship among solar wind density, temperature, and flow speed. *J. Geophys. Res. Space Phys.* 121, 7367–7384. doi:10.1002/2016JA022750

Hewish, A., Scott, P. F., and Wills, D. (1964). Interplanetary scintillation of small diameter radio sources. *Nature* 203, 1214–1217. doi:10.1038/2031214a0

Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F., and Collins, R. A. (1968). Observation of a rapidly pulsating radio source. *Nature* 217, 709–713. doi:10.1038/217709a0

Hollweg, J. V. (1986). Transition region, corona, and solar wind in coronal holes. *J. Geophys. Res. Space Phys.* 91, 4111–4125. doi:10.1029/JA091iA04p04111

Houminer, Z., and Hewish, A. (1972). Long-lived sectors of enhanced density irregularities in the solar wind. *Planet. Space Sci.* 20, 1703–1716. doi:10.1016/0032-0633(72)90192-4

Houminer, Z. (1971). Corotating plasma streams revealed by interplanetary scintillation. *Nat. Phys. Sci.* 231, 165–167. doi:10.1038/physci231165a0

Iwai, K., Shiota, D., Tokumaru, M., Fujiki, K., Den, M., and Kubo, Y. (2021). Validation of coronal mass ejection arrival-time forecasts by magnetohydrodynamic simulations based on interplanetary scintillation observations. *Earth, Planets Space* 73, 9. doi:10.1186/s40623-020-01345-5

Jackson, B. V., and Hick, P. P. (2004). Three-dimensional tomography of interplanetary disturbances. In *Astrophysics and space science library*. D. E. Gary, and C. U. Keller, 314. Dordrecht: Astrophysics and Space Science Library, 355. doi:10.1007/1-4020-2814-8_17

Jackson, B. V., Hick, P. L., Kojima, M., and Yokobe, A. (1998). Heliospheric tomography using interplanetary scintillation observations 1. Combined Nagoya and Cambridge data. *J. Geophys. Res. Space Phys.* 103, 12049–12067. doi:10.1029/97JA02528

Jackson, B. V., Boyer, J. A., Hick, P. P., Buffington, A., Bisi, M. M., and Crider, D. H. (2007). Analysis of solar wind events using interplanetary scintillation remote sensing 3D reconstructions and their comparison at Mars. *Sol. Phys.* 241, 385–396. doi:10.1007/s11207-007-0276-9

Jackson, B. V., Buffington, A., Hick, P. P., Clover, J. M., Bisi, M. M., and Webb, D. F. (2010a). SMEI 3D reconstruction of a coronal mass ejection interacting with a corotating solar wind density enhancement: The 2008 April 26 CME. *Astrophy. J.* 724, 829–834. doi:10.1088/0004-637X/724/2/829

Jackson, B. V., Hick, P. P., Bisi, M. M., Clover, J. M., and Buffington, A. (2010b). Inclusion of *in-situ* velocity measurements into the UCSD time-dependent tomography to constrain and better-forecast remote-sensing observations. *Sol. Phys.* 265, 245–256. doi:10.1007/s11207-010-9529-0

Jackson, B. V., Hick, P. P., Buffington, A., Bisi, M. M., Clover, J. M., and Tokumaru, M. (2010c). Solar mass ejection imager (SMEI) and interplanetary scintillation (IPS) 3D-reconstructions of the inner heliosphere. In *Advances in Geosciences*. Solar Terrestrial (ST). 21, 339–366. doi:10.1142/9789812838209_0025

Jackson, B. V., Hick, P. P., Buffington, A., Clover, J. M., and Tokumaru, M. (2012). Forecasting transient heliospheric solar wind parameters at the locations of the inner planets. *Adv. Geosciences* 30, 93–115. doi:10.1142/9789814405744_0007

Jackson, B. V., Odstrcil, D., Yu, H.-S., Hick, P. P., Buffington, A., Mejia-Ambriz, J. C., et al. (2015). The UCSD kinematic IPS solar wind boundary and its use in the ENLIL 3-D MHD prediction model. *Space Weather* 13, 104–115. doi:10.1002/2014SW001130

Jackson, B. V., Buffington, A., Cota, L., Odstrcil, D., Bisi, M. M., Fallows, R., et al. (2020). Iterative tomography: A key to providing time-dependent 3-d reconstructions of the inner heliosphere and the unification of space weather forecasting techniques. *Front. Astronomy Space Sci.* 7, 76. doi:10.3389/fspas.2020.568429

Jackson, B., Tokumaru, M., Fallows, R., Bisi, M., Fujiki, K., Chashei, I., et al. (2023). Interplanetary scintillation (IPS) analyses during LOFAR campaign mode periods that include the first three Parker Solar Probe close passes of the Sun. *Adv. Space Res.* in press. doi:10.1016/j.asr.2022.06.029

Jian, L. K., MacNeice, P. J., Taktakishvili, A., Odstrcil, D., Jackson, B., Yu, H.-S., et al. (2015). Validation for solar wind prediction at Earth: Comparison of coronal and heliospheric models installed at the ccmc. *Space Weather* 13, 316–338. doi:10.1002/2015SW001174

Jian, L. K., MacNeice, P. J., Mays, M. L., Taktakishvili, A., Odstrcil, D., Jackson, B., et al. (2016). Validation for global solar wind prediction using ulysses comparison: Multiple coronal and heliospheric models installed at the community coordinated modeling center. *Space Weather* 14, 592–611. doi:10.1002/2016SW001435

Kakinuma, T., Washimi, H., and Kojima, M. (1973). On the analysis of the observations of interplanetary scintillations obtained with three spaced receivers. *Publ. Astron. Soc. Jpn.* 25, 271–280.

Kasper, J. C., Bale, S. D., Belcher, J. W., Berthomier, M., Case, A. W., Chandran, B. D. G., et al. (2019). Alfvénic velocity spikes and rotational flows in the near-Sun solar wind. *Nature* 576, 228–231. doi:10.1038/s41586-019-1813-z

Kim, T. K., Pogorelov, N. V., Borovikov, S. N., Jackson, B. V., Yu, H. S., and Tokumaru, M. (2014). MHD heliosphere with boundary conditions from a tomographic reconstruction using interplanetary scintillation data. *J. Geophys. Res. Space Phys.* 119, 7981–7997. doi:10.1002/2013JA019755

Klinglesmith, M. (1997). *The polar solar wind from 2.5 to 40 solar radii: Results of intensity scintillation measurements*. San Diego: Ph.D. thesis, University of California.

Kojima, M., Tokumaru, M., Watanabe, H., Yokobe, A., Asai, K., Jackson, B. V., et al. (1998). Heliospheric tomography using interplanetary scintillation observations 2. Latitude and heliocentric distance dependence of solar wind structure at 0.1-1 AU. *J. Geophys. Res. Space Phys.* 103, 1981–1989. doi:10.1029/97JA02162

Kojima, M., Coles, W. A., Tokumaru, M., and Fujiki, K. (2013). Scintillation measurements of the solar wind velocity in strong scattering near the Sun. *Sol. Phys.* 283, 519–540. doi:10.1007/s11207-012-0207-2

Leamon, R. J., Smith, C. W., Ness, N. F., Matthaeus, W. H., and Wong, H. K. (1998). Observational constraints on the dynamics of the interplanetary magnetic field dissipation range. *J. Geophys. Res. Space Phys.* 103, 4775–4787. doi:10.1029/97JA03394

Li, B., Li, X., Hu, Y.-Q., and Habbal, S. R. (2004). A two-dimensional Alfvén wave-driven solar wind model with proton temperature anisotropy. *J. Geophys. Res. Space Phys.* 109, A07103. doi:10.1029/2003JA010313

Little, L. T., and Ekers, R. D. (1971). A method for analysing drifting random patterns in astronomy and geophysics. *Astron. Astrophys.* 10, 306.

Little, C. G. (1951). A diffraction theory of the scintillation of stars on optical and radio wave-lengths. *Mon. Notices Royal Astron. Soc.* 111, 289–302. doi:10.1093/mnras/111.3.289

Liu, L.-J., Peng, B., Yu, L., Liu, B., Lu, J.-G., Yu, Y.-Z., et al. (2022). Linear change and minutes variability of solar wind velocity revealed by FAST. *Mon. Notices Royal Astron. Soc.* 515, 3346–3351. doi:10.1093/mnras/stac2059

Manoharan, P. K., and Ananthakrishnan, S. (1990). Determination of solar-wind velocities using single-station measurements of interplanetary scintillation. *Mon. Notices Royal Astron. Soc.* 244, 691.

Manoharan, P. K., Ananthakrishnan, S., Dryer, M., Detman, T. R., Leinbach, H., Kojima, M., et al. (1995). Solar wind velocity and normalized scintillation index from single-station IPS observations. *Sol. Phys.* 156, 377–393. doi:10.1007/BF00670233

Manoharan, P. K., Subrahmanya, C. R., and Chengalur, J. N. (2017). Space weather and solar wind studies with OWFA. *J. Astrophysics Astronomy* 38, 16. doi:10.1007/s12036-017-9435-z

Manoharan, P. K. (2010). Ooty interplanetary scintillation - remote-sensing observations and analysis of coronal mass ejections in the heliosphere. *Sol. Phys.* 265, 137–157. doi:10.1007/s11207-010-9593-5

Martin, J. M., and Flatte, S. M. (1988). Intensity images and statistics from numerical simulation of wave propagation in 3-D random media. *Appl. Opt.* 27, 2111–2126. doi:10.1364/AO.27.002111

McComas, D. J., Elliott, H. A., Gosling, J. T., Reisenfeld, D. B., Skoug, R. M., Goldstein, B. E., et al. (2002). Ulysses' second fast-latitude scan: Complexity near solar maximum and the reformation of polar coronal holes. *Geophys. Res. Lett.* 29, 4-1–4-4. doi:10.1029/2001GL014164

Mejia-Ambriz, J. C., Villanueva-Hernandez, P., Gonzalez-Esparza, J. A., Aguilar-Rodriguez, E., and Jeyakumar, S. (2010). Observations of interplanetary scintillation (IPS) using the Mexican array radio telescope (MEXART). *Sol. Phys.* 265, 309–320. doi:10.1007/s11207-010-9562-z

Mejia-Ambriz, J. C., Jackson, B. V., Gonzalez-Esparza, J. A., Buffington, A., Tokumaru, M., and Aguilar-Rodriguez, E. (2015). Remote-sensing of solar wind speeds from IPS observations at 140 and 327 MHz using MEXART and STEL. *Sol. Phys.* 290, 2539–2552. doi:10.1007/s11207-015-0694-z

Merkin, V. G., Lyon, J. G., Lario, D., Arge, C. N., and Henney, C. J. (2016). Time-dependent magnetohydrodynamic simulations of the inner heliosphere. *J. Geophys. Res. Space Phys.* 121, 2866–2890. doi:10.1002/2015JA022200

Morgan, J. S., Macquart, J. P., Ekers, R., Chhetri, R., Tokumaru, M., Manoharan, P. K., et al. (2018). Interplanetary scintillation with the Murchison Widefield array I: A sub-arcsecond survey over 900 deg^2 at 79 and 158 MHz. *Mon. Notices Royal Astron. Soc.* 473, 2965–2983. doi:10.1093/mnras/stx2284

Nan, R., Li, D., Jin, C., Wang, Q., Zhu, L., Zhu, W., et al. (2011). The Five-hundred Aperture Spherical radio Telescope (FAST) project. *Int. J. Mod. Phys. D* 20, 989–1024. doi:10.1142/S0218271811019335

Oberoi, D., and Benkevitch, L. (2010). Remote sensing of the heliosphere with the Murchison Widefield Array. *Sol. Phys.* 265, 293–307. doi:10.1007/s11207-010-9580-x

Odstrcil, D., Riley, P., and Zhao, X. P. (2004). Numerical simulation of the 12 May 1997 interplanetary CME event. *J. Geophys. Res. Space Phys.* 109, A02116. doi:10.1029/2003JA010135

Pomoell, J., and Poedts, S. (2018). EUHFORIA: European heliospheric forecasting information asset. *J. Space Weather Space Clim.* 8, A35. doi:10.1051/swsc/2018020

Porowski, C., Bzowski, M., and Tokumaru, M. (2022). A new 3D solar wind speed and density model based on interplanetary scintillation. *Astrophys. J. Supp.* 259, 2. doi:10.3847/1538-4365/ac35d7

Purvis, A., Tappin, S. J., Rees, W. G., Hewish, A., and Duffett-Smith, P. J. (1987). The Cambridge IPS survey at 81.5 MHz. *Mon. Notices Royal Astron. Soc.* 229, 589–619. doi:10.1093/mnras/229.4.589

Qiu, Y. (1996). The IPS work in Beijing astronomical observatory. *Astrophysics Space Scence* 243, 255–258. doi:10.1007/BF00644065

Reiner, M. J., Kaiser, M. L., Fainberg, J., and Stone, R. G. (1998). A new method for studying remote type II radio emissions from coronal mass ejection-driven shocks. *J. Geophys. Res. Space Phys.* 103, 29651–29664. doi:10.1029/98JA02614

Riley, P., Linker, J. A., Lionello, R., and Mikic, Z. (2012). Corotating interaction regions during the recent solar minimum: The power and limitations of global MHD modeling. *J. Atmos. Solar-Terrestrial Phys.* 83, 1–10. doi:10.1016/j.jastp.2011.12.013

Romero-Hernandez, E., Gonzalez-Esparza, J. A., Aguilar-Rodriguez, E., Ontiveros-Hernandez, V., and Villanueva-Hernandez, P. (2015). Detection of solar wind disturbances: Mexican array radio telescope IPS observations at 140 MHz. *Sol. Phys.* 290, 2553–2566. doi:10.1007/s11207-015-0690-3

Sachdeva, N., van der Holst, B., Manchester, W. B., Tóth, G., Chen, Y., Lloveras, D. G., et al. (2019). Validation of the Alfvén Wave Solar Atmosphere Model (AWSoM) with observations from the low corona to 1 AU. *Astrophys. J.* 887, 83. doi:10.3847/1538-4357/ab4f5e

Salpeter, E. E. (1967). Interplanetary scintillations I theory. *Astrophys. J.* 147, 433. doi:10.1086/149027

Scott, S. L., Coles, W. A., and Bourgois, G. (1983a). Solar wind observations near the Sun using interplanetary scintillation. *Astron. Astrophys.* 123, 207–215.

Scott, S. L., Rickett, B. J., and Armstrong, J. W. (1983b). The velocity and the density spectrum of the solar wind from simultaneous three-frequency IPS observations. *Astron. Astrophys.* 123, 191–206.

Sheeley, J., N. R., Swanson, E. T., and Wang, Y. M. (1991). Out-of-ecliptic tests of the inverse correlation between solar wind speed and coronal expansion factor. *J. Geophys. Res. Space Phys.* 96, 13861–13868. doi:10.1029/91JA01168

Shishov, V. I., Tyul'bashev, S. A., Chashei, I. V., Subaev, I. A., and Lapaev, K. A. (2010). Interplanetary and ionosphere scintillation monitoring of radio sources ensemble at the solar activity minimum. *Sol. Phys.* 265, 277–291. doi:10.1007/s11207-010-9565-9

Singh, T., Yalim, M. S., Pogorelov, N. V., and Gopalswamy, N. (2019). Simulating solar coronal mass ejections constrained by observations of their speed and poloidal flux. *Astrophysical J. Lett.* 875, L17. doi:10.3847/2041-8213/ab14e9

Swarup, G., Sarma, N. V. G., Joshi, M. N., Kapahi, V. K., Bagri, D. S., Damle, S. H., et al. (1971). Large steerable radio telescope at Ootacamund, India. *Nat. Phys. Sci.* 230, 185–188. doi:10.1038/physci230185a0

Tatarskii, V. I., Ishimaru, A., and Zavorotny, V. U. (1993). *Wave propagation in random media (Scintillation)*. Bellingham: SPIE Press.

Tokumaru, M., Kojima, M., Fujiki, K., Maruyama, K., Maruyama, Y., Ito, H., et al. (2011). A newly developed UHF radiotelescope for interplanetary scintillation observations: Solar wind imaging facility. *Radio Sci.* 46, RS0F02. doi:10.1029/2011RS004694

Tokumaru, M., Fujiki, K., Kojima, M., and Iwai, K. (2021). Global distribution of the solar wind speed reconstructed from improved tomographic analysis of interplanetary scintillation observations between 1985 and 2019. *Astrophys. J.* 922, 73. doi:10.3847/1538-4357/ac1862

Tóth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., de Zeeuw, D. L., et al. (2005). Space weather modeling framework: A new tool for the space science community. *J. Geophys. Res. Space Phys.* 110, A12226. doi:10.1029/2005JA011126

Tu, C. Y., and Marsch, E. (1995). MHD structures, waves and turbulence in the solar wind: Observations and theories. *Space Sci. Rev.* 73, 1–210. doi:10.1007/BF00748891

van Haarlem, M. P., Wise, M. W., Gunst, A. W., Heald, G., McKean, J. P., Hessels, J. W. T., et al. (2013). LOFAR: The LOw-Frequency ARray. *Astron. Astrophys.* 556, A2. doi:10.1051/0004-6361/201220873

Verdini, A., Velli, M., and Buchlin, E. (2009). Turbulence in the sub-Alfvénic solar wind driven by reflection of low-frequency Alfvén Waves. *Astrophys. J. Lett.* 700, L39–L42. doi:10.1088/0004-637X/700/1/L39

Vitkevich, V. V., and Vlasov, V. I. (1970). Radioastronomical investigations of the drift of the inhomogeneous interplanetary plasma. *Sov. Astron.* 13, 669.

Wang, Y. M., and Sheeley, N. R. (1990). Solar wind speed and coronal flux-tube expansion. *Astrophys. J.* 355, 726. doi:10.1086/168805

Wang, W., Yan, Y., and Xiong, M. (2021). Antenna design for a new IPS array for space weather in China. Proceeding of the 43rd COSPAR Scientific Assembly. 28 January - 4 February 2021 43, 2371.

Watanabe, T., and Kakinuma, T. (1972). The recurrent solar wind streams observed by interplanetary scintillation of 3C 48. *Publ. Astron. Soc. Jpn.* 24, 459–467.

Watanabe, T., Shibasaki, K., and Kakinuma, T. (1974). Latitudinal distribution of solar wind velocity and its relation to solar EUV corona. *J. Geophys. Res. Space Phys.* 79, 3841–3843. doi:10.1029/JA079i025p03841

Wei, F., and Dryer, M. (1991). Propagation of solar flare-associated interplanetary shock waves in the heliospheric meridional plane. *Sol. Phys.* 132, 373–394. doi:10.1007/BF00152294

Wei, F., Feng, X., Ya, X., and Fan, Q. (2005). Prediction tests by using ISF method for the geomagnetic disturbances. *Adv. Space Res.* 36, 2363–2367. doi:10.1016/j.asr.2004.04.019

Weygand, J. M., Matthaeus, W. H., Dasso, S., Kivelson, M. G., and Walker, R. J. (2007). Taylor scale and effective magnetic Reynolds number determination from plasma sheet and solar wind magnetic field fluctuations. *J. Geophys. Res. Space Phys.* 112. doi:10.1029/2007JA012486

Xiong, M., Breen, A. R., Bisi, M. M., Owens, M. J., Fallows, R. A., Dorrian, G. D., et al. (2011). Forward modelling to determine the observational signatures of white-light imaging and interplanetary scintillation for the propagation of an interplanetary shock in the ecliptic plane. *J. Atmos. Solar-Terrestrial Phys.* 73, 1270–1280. doi:10.1016/j.jastp.2010.09.007

Xiong, M., Davies, J. A., Bisi, M. M., Owens, M. J., Fallows, R. A., and Dorrian, G. D. (2013). Effects of Thomson-scattering geometry on white-light imaging of an interplanetary shock: Synthetic observations from forward magnetohydrodynamic modelling. *Sol. Phys.* 285, 369–389. doi:10.1007/s11207-012-0047-0

Yamauchi, Y., Kojima, M., Tokumaru, M., Misawa, H., Morl, H., Tanaka, T., et al. (1996). Micro-turbulence in the solar wind at 5 – 76 Rs observed with interplanetary scintillation. *J. geomagnetism Geoelectr.* 48, 1201–1217. doi:10.5636/jgg.48.1201

Yan, Y., Wang, W., Chen, L., Liu, F., Geng, L., and Chen, Z. (2018). New interplanetary scintillation telescope array in China for space weather. *Sun Geosph.* 13, 153–155. doi:10.31401/SunGeo.2018.02.05

Young, A. T. (1971). Interpretation of interplanetary scintillations. *Astrophys. J.* 168, 543. doi:10.1086/151108

Yu, H. S., Jackson, B. V., Hick, P. P., Buffington, A., Odstrcil, D., Wu, C. C., et al. (2015). 3D reconstruction of interplanetary scintillation (IPS) remote-sensing data: Global solar wind boundaries for driving 3D-MHD models. *Sol. Phys.* 290, 2519–2538. doi:10.1007/s11207-015-0685-0

Zank, G. P., Zhao, L.-L., Adhikari, L., Telloni, D., Kasper, J. C., Stevens, M., et al. (2022). Turbulence in the sub-alfvenic solar wind. *Astrophysical J. Lett.* 926, L16. doi:10.3847/2041-8213/ac51da

Zhao, X., and Hoeksema, J. T. (1995). Prediction of the interplanetary magnetic field strength. *J. Geophys. Res.* 100, 19–33. doi:10.1029/94JA02266

Zhu, X., He, J., Verscharen, D., Duan, D., and Bale, S. D. (2020). Wave composition, propagation, and polarization of magnetohydrodynamic turbulence within 0.3 AU as observed by Parker Solar Probe. *Astrophys. J. Lett.* 901, L3. doi:10.3847/2041-8213/abb23e

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 21 July 2023
DOI 10.3389/fspas.2023.1223642

Check for updates

OPEN ACCESS

EDITED BY
Josep M. Trigo-Rodríguez,
Spanish National Research Council (CSIC), Spain

REVIEWED BY
Akos Kereszturi,
Hungarian Academy of Sciences (MTA), Hungary
Jianguo Yan,
Wuhan University, China

*CORRESPONDENCE
Jiao Wang,
jiaowang@cugb.edu.cn

RECEIVED 16 May 2023
ACCEPTED 30 June 2023
PUBLISHED 21 July 2023

CITATION
Li K, Wang J and Shi Y (2023), A review of studies on mass-movements on the Moon.
Front. Astron. Space Sci. 10:1223642.
doi: 10.3389/fspas.2023.1223642

COPYRIGHT
© 2023 Li, Wang and Shi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A review of studies on mass-movements on the Moon

Kaixuan Li, Jiao Wang* and Yijian Shi

School of Information Engineering, China University of Geosciences, Beijing, China

The study of the lunar surface is a significant field in exploring the Moon. As a non-negligible geological process shaping the lunar surface, however, the mass movements on the Moon have not been deeply studied regarding all available datasets. In this paper, we report the results of the literature review and quantitative analysis of 965 articles retrieved from the Scopus, WoS, and Astrophysics Data System databases using keyword search terms between the years 1872 and 2022. The main findings show that the lunar mass movements are a lasting but recent prosperous research topic (since 2009). The top three influential journals in this area are Icarus, JGR, and P&SS. Robinson, Lucchitta, and Carr are the most three productive authors. California Institute of Technology, Arizona State University, and Johns Hopkins University are active institutions leading the lunar mass movements research, and the related institutions are mainly located in the United States, Russia, and China. Articles are primarily published in journals specializing in the fields of astronomy and physics. The index analysis of citation indicates the growth of the academic impact of articles about the lunar mass movements. This article summarizes the datasets, methods, and relevant parameters of lunar mass movements to facilitate future research, as well as discussing the limitations and trends in the field. In addition, four unsolved issues are proposed, including: 1) the lack of a comprehensive global database that records all types of landslides incidents; 2) the need for definite classification indicators to quantify the classification criteria for lunar mass movements; 3) the absence of a mathematical model to explain the triggering mechanism of mass movements on the Moon; and 4) the lack of quantitative indexes to evaluate the modification effect of mass movements on the lunar surface morphology. With the support of big data, the combination of artificial intelligence and traditional GIS methods is expected to become the main approach for addressing these issues such a bibliometric analysis can inspire future researchers by exploring the overall characteristics of the published literature on lunar mass movements.

KEYWORDS

mass-movements, landslide, Moon, literature review, surface erosion

1 Introduction

Mass wasting is a material downslope movement driven by gravity. It tends to erase surface topographic differences and form gentle reliefs (Xiao et al., 2013). The processes occur widely on Earth, mainly on steep slopes, both subaerially and underwater (e.g., Wolman and Gerson, 1978; Moore et al., 1994; Deplus et al., 2001; Corsini et al., 2009; Parker et al., 2011). Different types of mass wasting processes have also been observed on several planetary and minor bodies, such as Mars (McEwen, 1989; Shaller and Komatsu, 1994; Brunetti et al., 2014), the Moon (Pike, 1971; Lindsay, 1976; Xiao et al., 2013), Mercury (Xiao and Komatsu, 2013), Venus (Waltham et al., 2008) and some icy satellites and

asteroids like Asteroid Eros, Itokawa, Lutetia, Vesta, Mathilde, Gaspra, Ida (Kereszturi, 2014; Murdoch et al., 2015) in the outer Solar System (Schenk and Bulmer, 1998).

In a terrestrial context, mass wasting represents the result of all the interactions between various geomorphological agents and processes, which act with different intensities on all types of slopes to modify the topography (Corominas et al., 2014; Pradhan and Siddique, 2019; Bhardwaj and Sam, 2022). In general, gravity is a major requirement for driving mass movements, but it is difficult to analyze the independent effects on the scales and dynamics of mass wasting by investigating terrestrial landslides alone, because of the presence of various other environmental elements on Earth, like a dense atmosphere, precipitation, and seismicity. Mass movements on other Solar System bodies share similarities. Thus, planetary studies with variable environmental conditions might help to reveal the effects of the atmosphere, subsurface fluids, precipitation, and seismicity on mass wasting, and could foster an understanding of triggering and propagation mechanisms (Sam and Bhardwaj, 2022).

As the sole natural satellite of the planet, the Moon is an important member of the inventory of planetary bodies that host mass movements. The mass movements of the Moon could bury ice and store it (Kereszturi, 2023), change the geomorphology of the lunar surface at different scales, impact the formation rate and process of lunar soil, affect the longevity of small craters, influence the dating of moderately young lunar surfaces, and record time clues about the lunar topography's evolution. Therefore, it is an important step to study the degree and mechanism of modification of the lunar surface topography by lunar mass movements to fully clarify the evolution of the lunar surface topography, which helps to understand the ancient state of the Earth, the origin, and evolution of the Solar System. Mass movements have been distributed on various terrains on the Moon, including impact craters, volcanic domes, tectonic scarps, rilles, valleys, and wrinkle ridges (Xiao et al., 2013), but most of the landslides occur on the walls of impact craters; including Fall, Slide, Slump, Creep, and Flow, five kinds of gravity erosion processes (Bickle et al., 2021), such as the creep produced so called "girlands" identified recently in deGerlache crater (Kereszturi et al., 2022). The surface gravity of the Moon is small, just one-sixth that of those on the Earth, which means that the Moon can support proportionally higher topography than the Earth for the same material strength (Melosh, 2011). Therefore, the lunar mass movements are on a grander scale usually. On the other hand, the small gravity causes less dynamic fragmentation, so the typical debris of the mass movements on the Moon is many times bigger than similar-sized terrestrial mass movements (Xiao, 2021).

Multiresolution and multisensory remote sensing datasets from lunar exploration missions are constantly improving our understanding of the Moon. Since Pike (1971) published the first paper about mass movements on the Moon using images from the Apollo 10 Mission, a variety of mass-movement features have been reported by researchers (e.g., Lindsay, 1976; Bart, 2007; Bray et al., 2010; Perumal, 2012; Xiao et al., 2013; Otto et al., 2014; Li et al., 2018). These studies provide a qualitative description of the general morphological traits of the lunar mass movements, and based on the preliminary investigation, the causes and effects of the lunar mass movements are examined, and certain classification criteria are formulated. Several approaches have been used to study the lunar mass movements, as corroborated by publications. These approaches include image processing and remote sensing, which are employed to map the shape of mass movements. Mapping tools are utilized to generate inventory maps of mass movements, while methods for classifying lunar mass movements are also explored. Additionally, comparative planetology is employed to examine physical and mineralogical parameters that may impact mass movements. Terrain parameters are utilized to evaluate the effects of morphology modification. Furthermore, model simulation is utilized to replicate the movement process and investigate the factors that influence mass movements.

Xiao. (2021) exhaustively summarized the patterns, causes, and effects of lunar mass movements based on the geological background of the Moon. Despite this, there is a lack of systematic analysis of mass movement studies on the Moon. To evaluate the research status of lunar mass movements, this review examines available literature on mass movements on the Moon. We conducted a quantitative analysis of the publications, authorships, disciplines, and quotations of 965 articles to explore trends, leading scholars, research fields, and academic impact in the study of the lunar mass movements. And then we carried out discussions: 1) compiled a list of the data catalogue used for lunar mass movements studies; 2) summarized the research techniques for studying mass movements; 3) highlighted the parameters used in terrain analysis and cause-effect analysis; 4) discussed the research gaps and future research trends. The review allowed to identify data, methods, and opportunities in the research of geomorphic processes, and geological evolution of the Moon.

2 Search strategy

Our search was conducted in Scopus, Web of Science (WoS), and Astrophysics Data System (ADS), a digital library for astronomy and astrophysics researchers. The created database spans 151 years (from 1 January 1872 to 18 May 2022) and provides an overview of contemporary research trends. Ten key phrases were used to search articles about lunar mass movements: lunar landslide, lunar mass movements, lunar mass wasting, lunar granular flow, lunar debris avalanche, lunar rock slide, lunar debris slide, lunar rock avalanche, lunar rockfalls, lunar boulder falls. We separated the keywords with quote marks and joined them with the OR logic operator. We examined each database (Scopus, WoS, and ADS) to find duplicate articles, and articles identified in two datasets are deleted, leaving just one version of the duplicated article. The results are primarily from ADS, complemented with Scopus and Wos search results. Abstracts, patents, and news releases did not contribute to the content of this review, since they did not provide a detailed description of the lunar mass movements study.

3 Statistical analysis

In this section, we aim to employ bibliometric analysis to uncover the diverse research aspects within the domain of lunar mass movements research. Based on the literature database, we investigate the overall research pattern by analyzing the annual publication count, and highlight publications that are usually available to acquire articles related to the lunar mass movements

through the distribution analysis of affiliated publications. The analysis of the author network and affiliated institutions can help researchers find the prominent scholars and reliable institutions in the field of lunar mass movements, and the research topics and latest research advances of their respective research groups. The objective of discipline analysis is to find the research trends of lunar mass movements across different disciplines, to identify the current hot research fields, and to comprehend the inter-disciplinary application. Lastly, the citation analysis unveils the trend of influence in lunar mass movements research, while co-citation analysis aids in identifying additional points of intersection for mass movements research.

3.1 Publications

There were 1,742 literature retrieved from the 10 keywords, of which 237 were in Scopus, 425 were in WoS, and 1,080 were in ADS. Using the deduplication rule, the final articles database contained 965 items (Figure 1) including 867 unduplicated articles (those found in only one database) and 98 duplicated articles (those found in several databases).

The number of published papers has grown exponentially since the first article about lunar mass movements appeared in 1872. ($R^2 = 0.7064$) (Figure 2) Taking 1973 and 2009 as key growth nodes, it can be divided into three development periods, with an average number of articles of 0.44, 14, and 29. Taking 10 years as the interval, the total number of articles in each period is calculated, which highlights the boom years of lunar mass movements research (See Figure 2). Researchers have only begun to do a lot of research on the mass movements of the Moon in the last 60 years, and 98 percent of the publications have been published in that time. 53.4 percent of all papers were published in the last 20 years (2003–2022). The high production years (2012, 2016, 2017, 2018, 2019, and 2020) were concentrated over this period, with values that were higher than the production mean (21 articles) reported for the last 30 years (1993–2022).

The articles were published in 175 publications and classified into three categories: Journal article, Conference contributions (talks and posters), and Book chapter, while the majority of them (55.6 percent) were concentrated in 32 publications. The journals with the highest number of publications were Icarus with 96 (9.9%), Journal of Geophysical Research (JGR) with 78 (8.1 percent), Planetary and Space Science (P&SS) with 30 (3.1 percent), Acta Astronautica (AcAau) with 24 (2.5 percent), Advances in Space Research (AdSpR) with 23 (2.4 percent). In addition, some Ph.D. T (Ph.D. thesis) is of great reference value with 27 (2.8 percent) (Figure 3A). The remaining papers (44.4%) were published in 143 different publications. From the data above, we could figure out which journals are usually available to acquire articles related to lunar mass movements.

3.2 Authorship

The 965 articles examined included a total of 2,622 authors. We create an author network by taking the top 200 most often appearing writers in the database and measuring the frequency of collaboration between all authors (Figure 4). According to the frequency of collaboration, the authors were split into eight groups, and the lines indicated some collaboration with authors in other groups. The analysis shows that authors in group 1 focused on the erosion of the surface by ejecta on different planets. The second group studied the characteristics, distribution, and age of valleys and landslides on Mars, and the topic of group 3 is geological structures and process on different planets. Group 4 studied the lunar regolith based on image processing. The fifth group investigated the evolution of the planetary surface. Group 6 focuses on the evolution of craters on the Moon, and group 7 is about the thermodynamics of planets. The rest were grouped into the "other" group.

Over time, the research activeness of each group varies. The first group persisted in their research related to lunar mass movements and produced publications on the subject throughout time. Furthermore, group 4 has continued to conduct relevant research in recent years, and added fresh research findings to the database this year. For the past 4 years, groups 2, 3, and 5 have been without a new article about the subject, and group 7 has not published a new research output for approximately 15 years. Finally, the researchers of the sixth group have not updated their articles anymore since 1997.

The affiliations of the authors of the articles were dispersed among 450 institutions, including universities, research institutes, and centers. The country distribution of the articles was counted based on the affiliated institutions of the first author, as shown in Figure 5, which are mainly distributed in the United States (33.49%), Russia (14.10%), China (13.13%), Germany (6.99%), and Italy (6.27%). Coincidentally, among these countries, the United States has successfully launched 30 lunar exploration missions, Russia has successfully launched 27, China has launched 7, and the European Space Agency has launched 1. More than half of the published articles (54.2 percent) were concentrated in 15 institutions, California Institute of Technology (6.9 percent), Arizona State University (4.9 percent), Johns Hopkins University, and Russian Academy Of Sciences (4.6 percent each), NASA Goddard Space Flight Center and United States Geological Survey (4.1 percent each), Planetary Science Institute (3.7 percent), Max Planck Society (3.4 percent), Brown University (3.3 percent), National Aeronautics and Space Administration (3.3 percent) (Figure 3B).

3.3 Interdisciplinarity

The papers in the database involve 35 disciplines, with Astronomy accounting for 686 (57 percent), Physics for 221 (18 percent), Engineering for 92 (8 percent), Earth and Planetary Sciences for 77 (6 percent), Geochemistry Geophysics for 32 (3 percent), and Geology for 29 (2 percent). (Figure 6).

The changing trend of several disciplines (including a large proportion of articles) from the time dimension was further analyzed. In the direction of physics from 1915 to 1970, a total of 209 publications (94.6 percent) were published, with an average of four papers a year. Papers in Astronomy surged beginning in 1966 and have remained at a high level, averaging 12 publications annually, whereas previously relevant articles were published occasionally. Engineering studies focused mostly on the design and application of spacecraft; the first relevant article we searched was published in

10.3389/fspas.2023.1223642

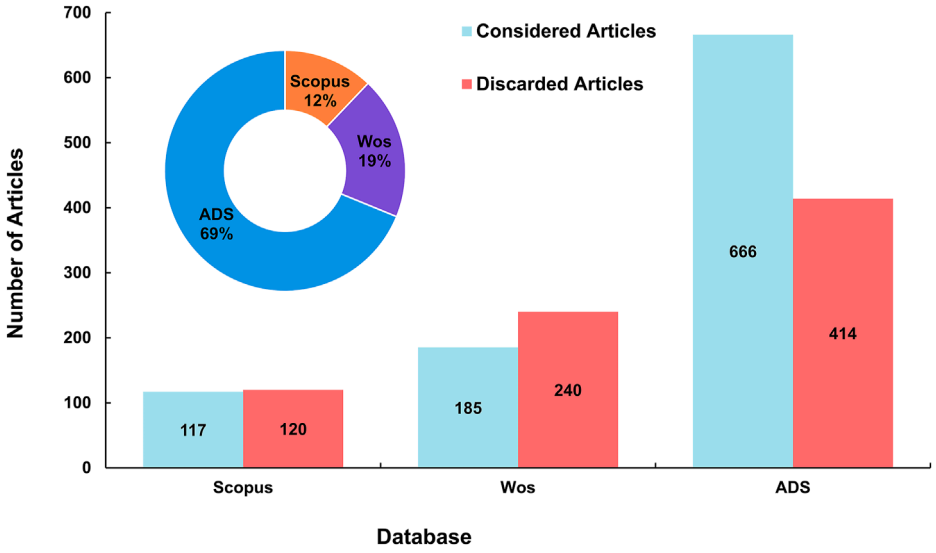


FIGURE 1
Related articles are distributed in three online databases including original search results and deduplicated results.

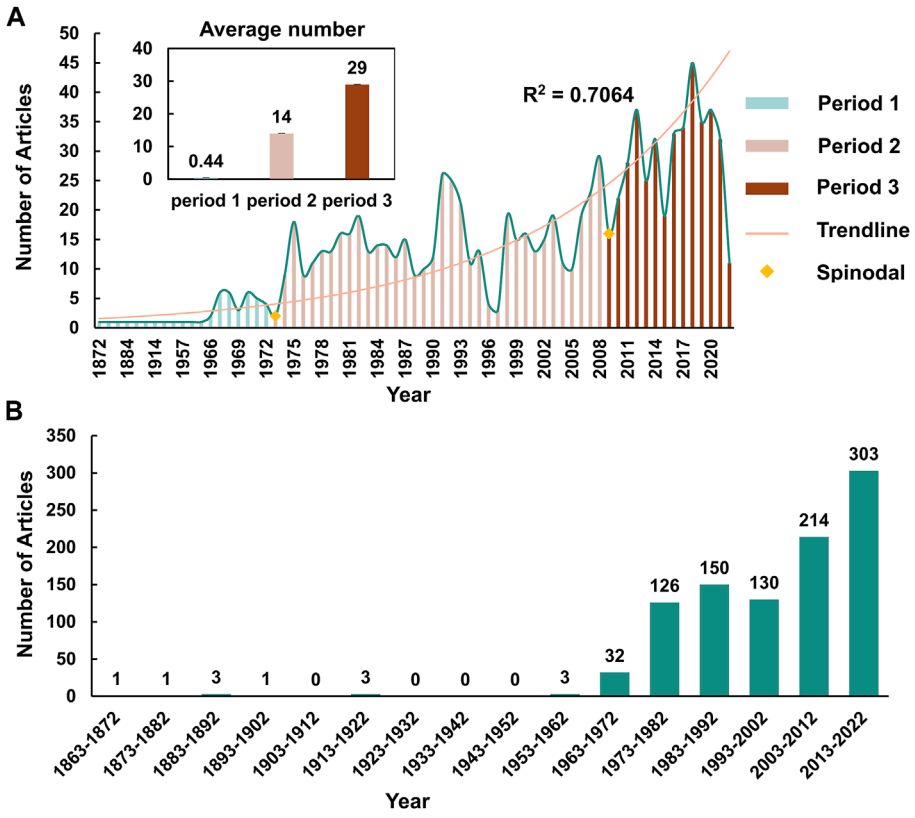


FIGURE 2
Characteristics of annual lunar mass movements literature: **(A)** Number of publications from 1872–2022 increases exponentially (R^2 = 0.7064, remove the years with no publication). The number of articles over years indicates that the evolution of lunar mass movements research could be divided into three stages (1973 and 2009 are the turning points). **(B)** Numbers of publications each decade. Most publications (98%) related to the lunar mass movements have been published in the recent 60 years.

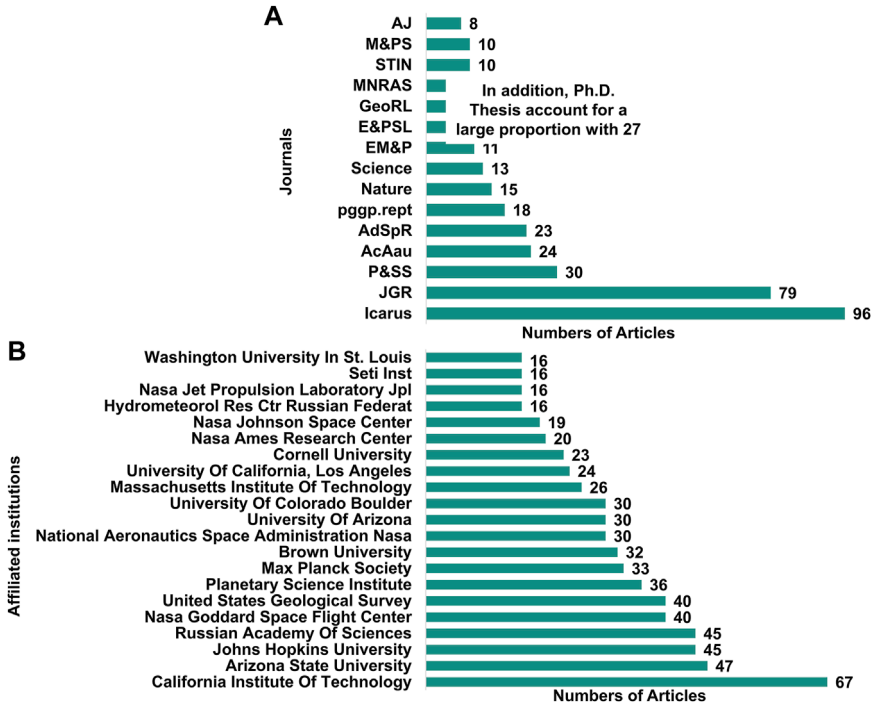


FIGURE 3
Statistics on the sources of literature on lunar mass movements: **(A)** The 15 productive journals in lunar mass movements research. **(B)** The 20 productive institutions in lunar mass movements research.

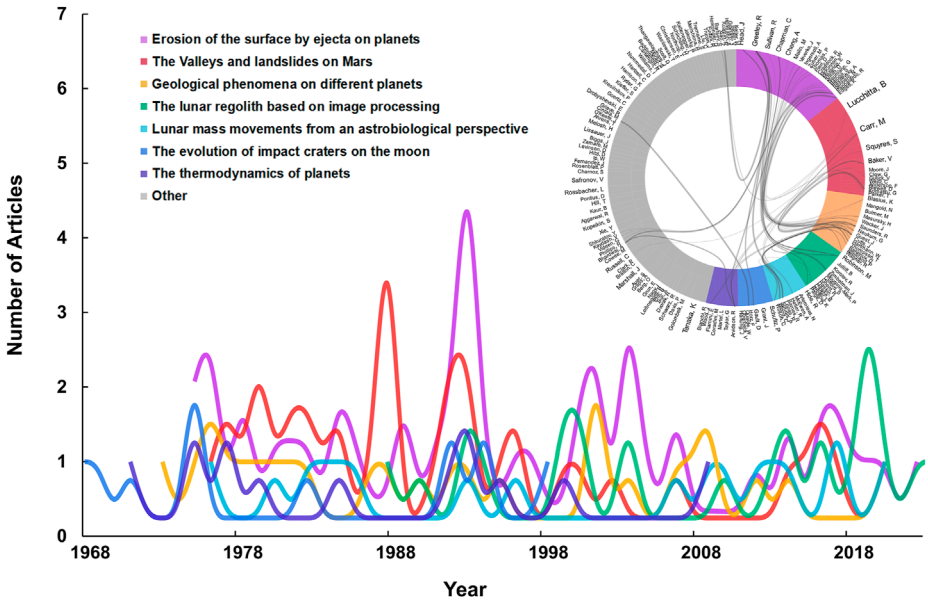


FIGURE 4
Cooperative relationship network of authors. The pie chart displays the authors grouped by the cooperation frequency and named each group by the common research fields, and lines indicate the other collaboration between authors. The line chart reflects the timeline view of the research group activity.

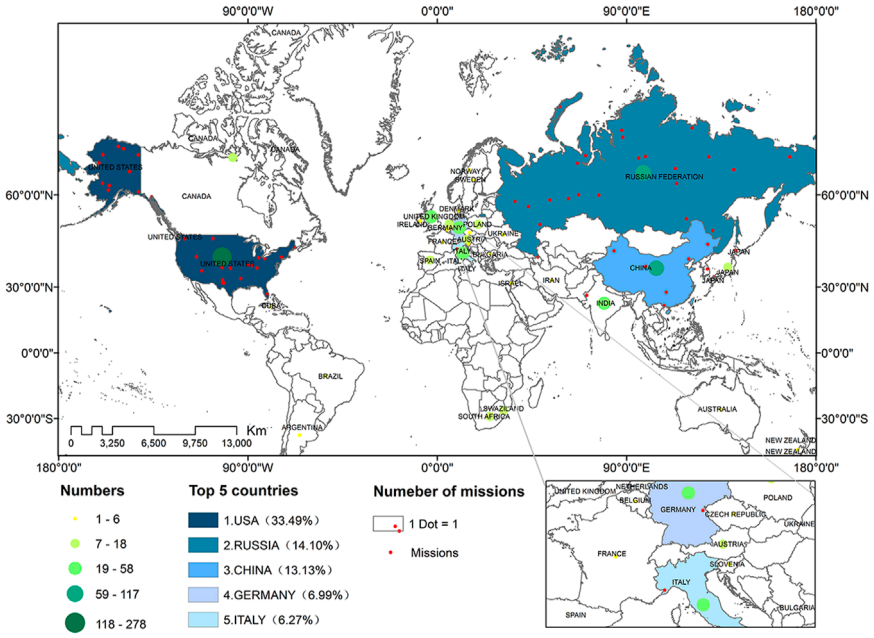


FIGURE 5
Geographical distribution of lunar mass movements publication volume level (Counted based on the affiliated institutions of the first author). The top 5 productive countries in lunar mass movements research are highlighted, the number of scatter points indicates the number of the lunar exploration missions each country launched.

1972, and after that, an increasing number of related publications were published, with an average of six articles per year for nearly 10 years. Earth and Planetary Sciences has been studied since 1969 and has been booming in recent years, with 75.3 percent of publications published in nearly 10 years, averaging 5.8. (See Figure 6).

The retrieved articles are from multiple scientific directions suggesting that the study of lunar mass movements is valuable in different scientific fields. An analysis from the time dimension reveals the development trend of several key disciplines in lunar mass movements research. Most of the articles are in the field of Astronomy, which is partly due to the absence of geology-focused journals in the past, but it can also indicate that the current research on lunar mass movements is mainly focused on this direction The rapid expansion of Earth and Planetary Sciences and many space missions in recent years validates the extensive application of comparative planetary methods in the study of lunar mass movements.

3.4 Quotation

The 965 papers in the database have been cited a total of 20,747 times between 1 January 1964, and 18 May 2022, including 409 self-citations. Each article was cited 21 times on average. Figure 7A demonstrates that the number of citations increases linearly over time. The total number of reads of these publications since 1996 was 109,094, with an average of 113 reads per publication, according to data collected by the system from ADS, Wos, and Scopus about the rate at which articles have been accessed. And 36,959 downloads were made in total, with a 38.3 download average. The data mentioned above illustrates the expansion of the impact of research on lunar mass movements, which is logical given the trend of synchronous rise in the number of articles.

We utilize six indexes, including the h-index[1], g-index[2], i10-Index[3], i100-Index, tori-Index[4], and read10-Index[5], to provide a summary of the citations, usage, and derived indicators of these articles. Because of author disambiguation issues, the result does not remove self-citations based on the author's name but instead uses a list-based removal of self-citations. Since 1962, the tori-index has increased rapidly, eventually peaking at 354.9, indicating that as time goes on, the research on the mass movements of the Moon and the contribution of earlier work to subsequent research grows. Additionally, the maximum values of the i10-index and i100-index

1 Hirsch's h-index is the largest number H such that H publications have at least H citations. It attempts to measure the productivity and impact of a researcher in a single number.

2 Given a set of articles ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number such that the top g articles received (together) at least g2 citations.

3 The iN-index is the number of publications with at least N citations.

4 The total research impact of a scholar (tori) is calculated using the reference lists of the citing papers, where self-citations are removed. The contribution of each citing paper is then normalized by the number of remaining references in the citing papers and the number of authors in the cited paper. The tori-index is defined as the amount of work that others have devoted to his/her research, measured in research papers.

5 The Read-10 index represents the present readership rate for an individual's papers published within the last 10 years, normalized according to the number of authors.

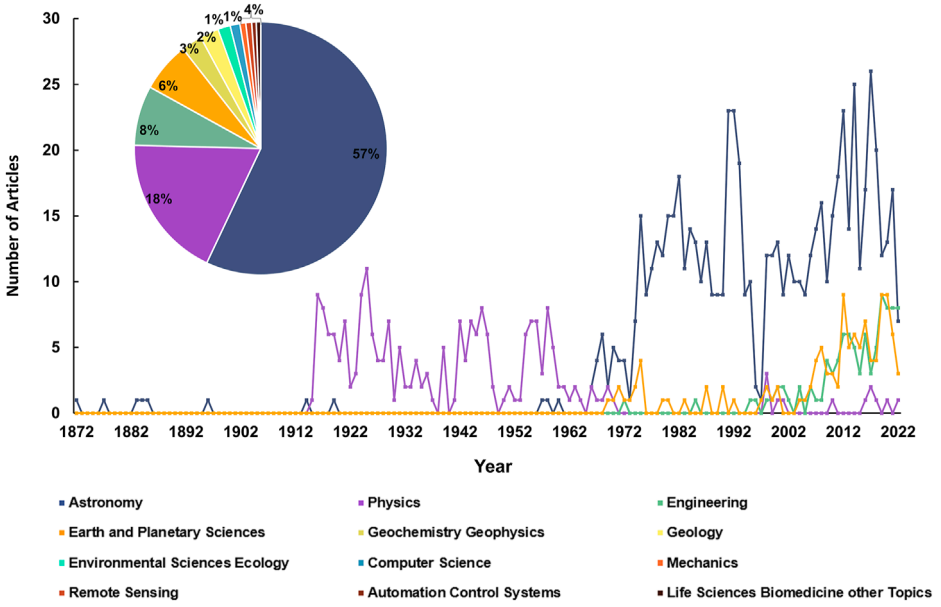


FIGURE 6
Disciplines involved in the research on lunar mass movements. The pie chart is characteristic of the discipline resulting from publication analysis. The line chart is the characteristics of the annual publication of different disciplines.

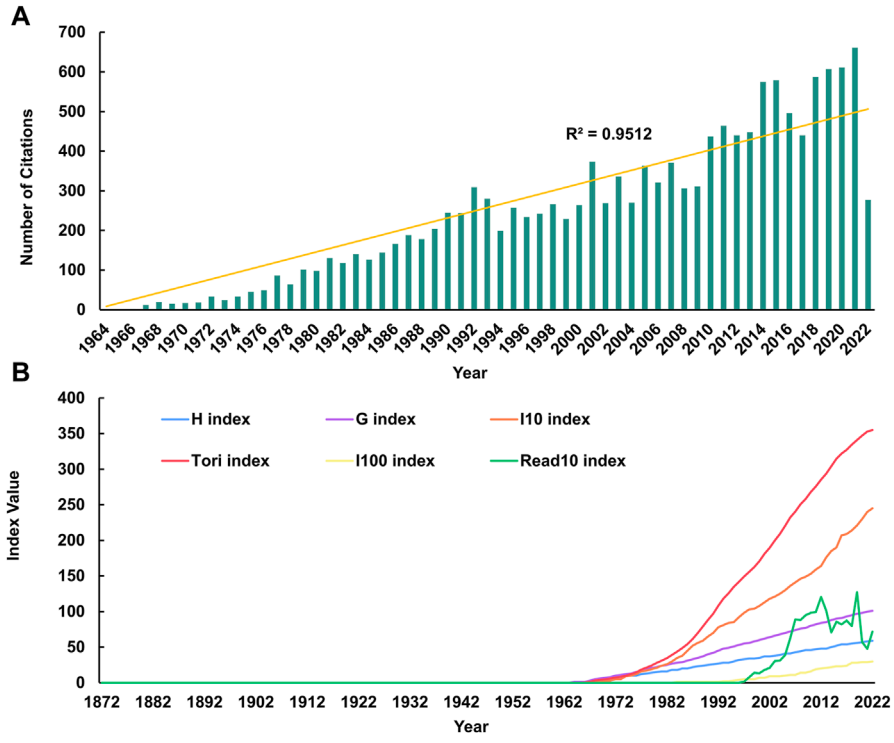


FIGURE 7
Overview of the citation of the literature on lunar mass movements: **(A)** The citation of related articles increases linearly over time (R^2 = 0.9512). **(B)** Characteristics of the index of citation per year.

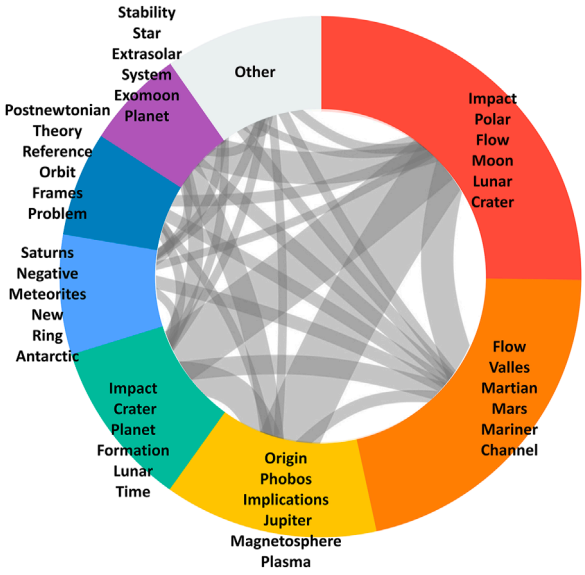


FIGURE 8
Cooperative relationship network of citation.

were each 245, while the maximum values of the h-index and g-index, with maximum values of 59 and 101, respectively, indicated a steadily ascending trend. The trend indicates that the number of articles with a high citation rate is increasing, implying growth of the academic impact of articles about the lunar mass movements. The Read10-index has fluctuated since 2005, indicating that the author's research on the mass movements of the Moon was volatile during the period. The statistics "0" in the chart are due to a data shortage (Figure 7B).

The co-citation network groups articles based on shared references across all publications and names them by searching for common, unique words in their titles. And the lines indicated shared references with papers from other groups, the size of the line reflects the frequency of co-citation (Figure 8). As shown, articles relating to the lunar impact crater and the Martian valley are frequently cited within the same article, implying that lunar mass movements research is commonly carried out by comparing the mass movements of Mars and the Moon. It can also be investigated by comparing the mass movements of other planets. Additionally, surface erosion, slope stability, and magnetosphere effect are some other directions that are researched frequently.

4 Discussion

4.1 Data for lunar mass movements research

The statistical analysis results of publication years and country distribution indicate that as lunar exploration technology matures, there is an increasing number of articles focused on studying lunar mass movements. This observation highlights the necessity of lunar exploration missions in providing fundamental lunar data for conducting research on lunar mass movements. At present, three ongoing lunar missions must be mentioned: the Lunar Reconnaissance Orbiter (LRO) by the National Aeronautics and Space Administrations (NASA, United States), the SELenological and ENgineering Explorer (SELENE-KAGUYA) by the Japan Aerospace Exploration Agency (JAXA, Japan), and the Chang'E missions by the Chinese National Space Administration (CNSA, China) (Scaioni et al., 2018) (Table 1).

The open-source Java Mission Planning and Analysis for Remote Sensing (JMARS) software, a WEB-GIS platform developed by the Arizona State University (Burris et al., 2022), allows for handling planetary remote-sensing datasets. The USGS's Integrated Software for Imagers and Spectrometers (ISIS; https://isis.astrogeology.usgs.gov//) could use to perform a standard calibration procedure for Lunar Reconnaissance Orbiter Camera (LORC) images (Chin et al., 2007).

The first lunar landslides the researchers encountered were several landslides around the Tsiolkovskiy crater (Boyce et al., 2020). This landslide is the only large, long-runout landslide on the Moon, this makes the Tsiolkovskiy landslide an excellent end-member for the study of the mechanics and conditions that control the emplacement of such rapid natural mass flows and their runout distance. Over the years, the topographic parameters, morphological characteristics, formation mechanism, and surrounding areas of regolith, ejecta, geological age, and rock abundance of the landslides in the Tsiolkovskiy crater have been well studied (Guest and Murray, 1969; Wu et al., 1972; Schultz, 1976b; Greenhagen et al., 2016; Morse et al., 2018; Boyce et al., 2020). Since then, numerous mass wasting features on a local scale have been noticed and reported based on a variety of lunar image data and terrain data (Table 2).

Even though various slope failure types have been examined on the Moon, there is still a lack of a comprehensive inventory of lunar landslides on a global scale. Table 3 lists the existing inventory of mass movements on the moon on a global scale.

From the list of mass movements, we found that researchers have used several morphological characteristic factors to describe the shape of the mass movements quantitatively (see Table 2). Moreover, Brunetti et al. (2015) compiled the inventory of 60 rock slides in the simple crater and using the planimetric area of the mapped failures calculated the probability density distribution of the landslide area. The measuring methods in detail list in Table 4 Below.

4.2 Methods used in lunar mass movements research

Three approaches are most commonly used to study mass movements on the Moon: 1) Image recognition, which is used to extract the boundary information of mass movements, thus studying their morphological, evolutionary, formational, distributional, and erosional characteristics; 2) Model simulation; physical models are usually constructed to simulate the lunar environment to study the development process and formation causes of mass movements; 3) Comparative planetology, by comparing the similar characteristics of different planets, the characteristics, influencing factors and classification standard of lunar mass movements can be well interpreted.

TABLE 1 Data access to lunar mass movements research.

Lunar missions	Instrument		Spatial resolution	Access
LRO	Lunar Reconnaissance Orbiter Camera (LORC)	Narrow-Angle Camera (NAC)	~0.5 m/pixel	http://target.lroc.asu.edu/q3/
		Wide-Angle Camera (WAC)	~100 m/pixel	
	Lunar Orbiter Laser Altimeter (LOLA)		~57 m vertical resolution: ~0.1 m	http://denali.gsfc.nasa.gov/
SELENE-KAGUYA	Lunar Imager/Spectrometer (LISM)	Terrain Camera (TC)	10 m	http://l2db.selene.darts.isas.jaxa.jp/
		Multi band Imager (MI)	VIS: 20 m	
			NIR 62 m	
		Spectral Profiler (SP)	562 × 400 m	
	Laser Altimeter (LALT)		5 m	
Chang'E missions	Three-line Array Charge Coupled Device Stereo Camera (CCD)		~1m–7 m	https://moon.bao.ac.cn/
	Laser Altimeter (LAM)		5 m	

TABLE 2 Available data catalogue of lunar mass movements in region.

Author	Location	Data	Characterization	Factors
Bray et al. (2010)	Crater Thales, Giordano Bruno	High-resolution (~0.5–2.0 m/pixel) images from the LOR-NAC	Melt deposits on the floors of complex craters are comprised of a mixture of melt and clastic debris, giving the majority of deposits a hummocky appearance	
Perumal. (2012)	Shrodinger basin	High-resolution images from the Chandrayaan-1 terrain mapping camera (TMC) and Lunar Reconnaissance Orbiter (LRO) images	Run-out flows induced seismic shaking as responsible for the down-slope mass movements. Crater counting dating provides a minimum age of landslides to be around 30 Ma	Slope
Otto et al. (2014)	The Tycho crater	The Low Altitude Mapping Orbit (LAMO) data from the NASA Dawn mission (~20 m/pixel) LOR-WAC (~100 m/pixel)	Distributed along the entire crater wall, flow-like features are observed, which are slides of low friction or liquid behavior developing striations	
Kumar et al. (2016)	Schrödinger basin	Chandrayaan-1 Terrain Mapping Camera (TMC) images, LOR-NAC&WAC, and the LOLA data set	Rim-crest is relatively smooth compared to hummocky	Length
				Slope
				Width
Li et al. (2018)	LaLande crater	LRO NAC mosaic and WAC mosaic	There are sweeping flows on the eastern crater wall	Area
Kokelaar et al. (2017)	Crater Kepler, Gambart B, Bessel, Censorinus, Riccioli CA, Virtanen F, Tralles A	High-resolution images (0.42–1.3 m/pixel) from the LORC, the LOLA, and "Kaguya Terrain Camera DEM Merge 60N60S 512 ppd"	Three types of dry-debris flow deposits are recognized: 1) multiple channel-and-lobe type, 2) single-surge polylobate type, 3) multiple-ribbon type	Length
				Slope
				Area
				Thickness
				Volume
Boyce et al. (2020)	The western rim of the Tsiolkovskiy crater	LROC images (NAC, WAC), Kaguya, and LOLA topographic data	The only giant landslide on the Moon with a lobate flow feature. The landslide components (i.e., north, south, west, and middle slides) formed almost simultaneously	Length Thickness
				Area
				Volume Mobility (i.e., L/H, H/L, L/V)

TABLE 3 Available data catalogue of lunar mass movements in globe.

Type	Number	Data	Method	Characterization
Falls, Slides, Flows, Creeps Xiao et al. (2013)	300	LOR-NAC (~0.5 m/pixel)	Visual interpretation	See Table 6
		The LOLA global elevation data (1,024 pixel/degree ~30 m/pixel in the equator)		
Rock slides Brunetti et al. (2015)	60	LOR-WAC (~100 m/pixel); Global WAC digital terrain model	Visual interpretation	The upper parts of the scarp areas of the landslides are visible on the crater rims or walls, and most of the displaced materials are deposited on the crater floors
		(~100 m/pixel)		
Rockfall Bickel et al. (2020)	136,610	LOR-NAC (~0.5 m/pixel)	RetinaNet	1) A boulder that usually has a higher albedo than the surrounding regolith
				2) A shadow that is cast by the boulder and appears right next to its parent boulder
				3) A nearly linear boulder track with light shadows at its deepest point and potentially along its rims
Granular Flows Bickel et al. (2022)	28,101	LOR-NAC (~0.5 m/pixel)	RetinaNet	The deposits of lunar flows usually form fans, fingers, rays, and/or lobes and represent distinct features that can be recognized from orbit, both morphologically and radiometrically, despite their small size

As research advances, image recognition techniques are gradually becoming semiautomatic or automatic from the manual. There are three main approaches: visual interpretation incorporating expert knowledge; recognition based on morphological features; and recognition based on machine learning or deep learning methods (Table 5). The primary method is visual interpretation with expert knowledge at present due to the diversity of types and complex morphological characteristics, which is slow and laborious. Technology for automatic identification is evolving, whereas the method is only capable of detecting a single type of mass movement at present. And recognition methods based on machine learning or deep learning lack sample libraries containing multi-scale morphological information of lunar mass movements. Therefore, more comprehensive automatic recognition methods need to be developed in combination with artificial intelligence technology in the future.

Physical model simulation compensates for the fact that researchers could not contact the lunar surface, helps us to imitate the process of movements, and explore the factors that influence mass movements and the effects on geomorphic evolution. By the discrete element method, Bui et al. (2009) introduce the self-locking ability between particles into the particle contact model to study the slope collapse of lunar soil. Since then, the discrete element method has become a common method for simulating lunar soil landslides, analyzing landslide processes and landslide mechanisms, and evaluating the engineering disaster index (Liao et al., 2018; Liever et al., 2018; Wu et al., 2020). To investigate the influence of the low gravity environment on particle fluidity, Huang and Mao (2013) used the drop tower to design a system to simulate a low gravity environment, which captured the configurations and variability in the flow materials by drop-tower tests. Zheng and Huang (2015) used silicon carbide as a lunar regolith simulant to study the flow patterns of lunar regolith by capturing configurations of the flow and calculating the engineering strains under the action of different chute angles.

As mentioned in the introduction, the mass movement process is widely distributed on Earth, Moon, Mars, Mercury, Venus, and some icy satellites and asteroids in the outer Solar System, with some of the same or similar characteristics, meaning that their forming may have similar properties. As a result, when studying the classification, formation, and influencing factors of mass movements on the Moon, researchers frequently refer to mass movements on other planets (Hargitai and Kereszturi, 2015). For example, an alcove, a channel, and a deposition apron are characteristics shared by lunar crater-wall landslides and Martian gullies. Fine granular material without liquid water collapsing results in the formation of Martian gullies. Also, on the Moon, there is no atmosphere and water, these observations suggest that dry landslides can form recess-channel-apron morphology on the planets (Bart, 2007). Mass-wasting such as slumps, slides, and flow-like features are both found in the Rheasilvia crater on Vesta and the Tycho crater on the Moon. Analyzing the similarities and differences of mass-wasting features on the Moon and asteroid Vesta concerning their gravity shows that slumping is triggered when gravity forms a torque on the

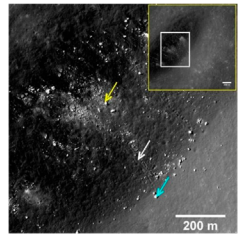
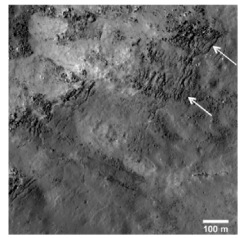
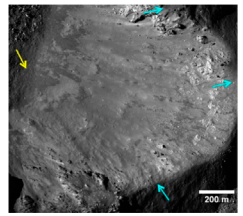
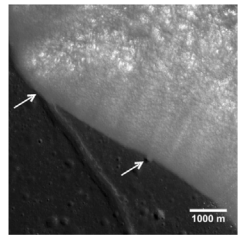
TABLE 4 Factors for measuring lunar mass movements morphology and mobility.

Factor		Description	Method	Role
Slope		The local slope angle nearby of the landslide scarps	Slope angles are tangents measured on the slope plots	Explore the relationship between slope instability and slope
Geometry	Area	Areas of the plan view of deposits	Measured using an ArcMap outline tool	Understand size, origin, emplacement, as well as the relationship with other terrains in the immediate vicinity
	Length	The maximum length from the rim	Measured the distance using the ruler in ArcMap	
	Width	The maximum width from the rim		
	Thickness	The thickness of the deposits	Derive thickness approximately from shadow lengths	
	Volume	The volume of the deposits	Segment on-slope areas and thicknesses were coupled to determine the segment volumes and these were then summed to find a total	
	Drop height	The difference in elevation between the source area and the deposit extreme tip	Measured by elevation difference based on altitude data using ArcMap	
Mobility	L/H	The ratio of length to drop height	Calculate the ratio of length to drop height	The H/L ratio is widely regarded as a proxy for the dynamic friction angle of the landslide. Study the mobility dependence on volume
	L/V	The ratio of length to volume	Calculate the ratio of length to volume	
	H/V	The ratio of drop height to volume	Calculate the ratio of drop height to volume	

TABLE 5 Methods for lunar mass movements identification.

Method	Author	Description
Visual interpretation	Schultz. (1976a)	Lunar surface features were interpreted by visual interpretation based on lunar orbiter photographs, accompanied by explanatory text
	Xiao and Komatsu. (2013)	Collecting over 300 typical mass wasting features by observed mass wasting features in LORC NAC images.
	Brunetti et al. (2015)	Mapped 60 rockfalls in simple impact craters, with scarps and deposits, have been distinguished, and some tracks, blocks, and cracks in the bodies have been mapped
	Kumar et al. (2016)	Found Boulder falls in the Schrödinger basin and mapped >1,500 boulders associated with trails and bouncing marks
	Kokelaar et al. (2017)	Seven lunar crater sites of granular avalanches are studied utilizing high-resolution images, for the first time, extensive remobilization of coarse talus by inundation with finer debris
Morphology	Craddock and Howard. (2000); Simon and Bruyn. (2007)	Using inequalities to fit crater boundaries to identify landslide locations and profiles
	Mahanti et al. (2014)	Chebyshev polynomial function is used to describe the crater shape in a quantitative and repeatable way
	Yordanov et al. (2016)	Based on the Chebyshev polynomial method to match the cross sections of impact craters, identified landslides by associating the asymmetry of the cross-section with the landslides
	Scaioni et al. (2018)	Based on the classification of morphological parameters by Chebyshev polynomials, fixing empirical absolute thresholding and statistical adaptive thresholding are adopted to identify slumps in simple impact craters
Deep learning	Bickel et al. (2020)	Analyzing an archive of more than 2 million high-resolution images from LORC using a fifth-generation Convolutional Neural Network (CNN M5), created the first global map of 136.610 lunar rockfall events

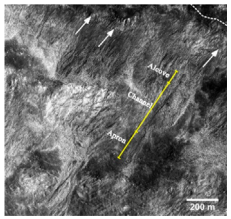
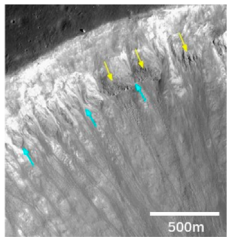
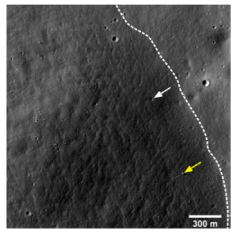
TABLE 6 Characteristics of different types of lunar mass movements [Modified from Xiao et al. (2013)]

Type		Definition	Distribution	Characterization	Example images Xiao et al. (2013)
Falls		Abrupt downslope movements by which rocks or regolith detach from steep slopes Highland and Bobrowsky, (2008)	A subhorizontal stratum along the rim of the rilles	Traces of rolling boulders are also typical indicators of recent rock falls	
Slides	Rock slides	Mass downslope movements along ruptures or relatively thin zones of intense shear strain, separate the sliding material from the more stable underlying planar Highland and Bobrowsky, (2008)	Most slides occur in Eratosthenes and Copernican-aged craters	The volume of displaced material gradually enlarges from the sliding front. Sub-parallel fractures grow in impact melt sheets due to gravitational dragging; the shattered rocks slide along slopes forming talus deposits at the foot of the slopes. Cruden and Varnes. (1996)	
	Debris slides			Debris slides displace material that has smaller grain sizes than rock slides	
Slumps		Sudden mass movements of large amounts of rocks and/or fine material for short distances Ritter et al. (2006)	Craters, tectonic scarps, and volcanic rilles	The movement is characterized by sliding along a concave-upward or planar surface, the displaced material in slumps usually has a larger volume Cruden and Varnes. (1996)	

(Continued on the following page)

10.3389/fspas.2023.1223642

TABLE 6 (*Continued*) **Characteristics of different types of lunar mass movements [Modified from Xiao et al. (2013)]**

Type		Definition	Distribution	Characterization	Example images Xiao et al. (2013)
Flows	Channeled flows	A form of continuous mass movement in which a combination of loose soil and rock fragments move downslope Highland and Bobrowsky, (2008)	Flow features are only observed in impact craters	Channeled flows on the Moon have a morphology of flow feature, usually composes of a gully/alcove in the source region. Malin and Edgett. (2000)	
	Sweeping flows			Have a morphology of flow feature, but does not form any channels	
Creeps		Creep is imperceptibly slow, steady, downward movement of soil and/or rocks Highland and Bobrowsky, (2008)	Almost all lunar surfaces	Form the lunar regolith ripples, appear convex in downslope directions, and are sub-parallel with each other. Adjacent ripples are about 20 m apart	

10.3389/fspas.2023.1223642

slumping body which exceeds the frictional forces and slumping develops more efficiently on bodies with higher gravity (Otto et al., 2014). Rock slides on the Moon are, on average, more massive than analogous rock slides on Mercury, according to a comparison of the size of landslides on the simple impact crater-wall on the Moon and Mercury (Brunetti et al., 2015). Therefore, it is hypothesized that authors a result of the properties of the rock material or possibly a result of the lunar surface gravity, which is weaker than those of Mercury's (Brunetti et al., 2015). Furthermore, the lunar mass movements classification system was updated from the Varnes (1978) classification system for terrestrial landslides based on movement characteristics and material properties (Cruden and Varnes, 1996).

4.3 Parameters of morphological and triggering analysis

As mentioned above, several variables were employed to quantify the morphological characteristics of mass movements including geometric parameters such as length, width, surface area, volume, and average thickness; mobility parameters such as L/H, L/V, and H/V; and other factors such as slope angle, which were typically determined using ArcMap software (Table 4). These morphometric parameters serve as indicators that reflect the effects of mass movements on topographic modification of the lunar surface. They are integrated into topographic analysis for quantitative comparison to highlight the difference of mass wasting within various environments, thus revealing the triggers and modification effects on the terrain of the lunar mass movements. For example, slope angles measurement for the host terrains of the mass wasting features aids in detecting the minimum slopes on which mass wasting can initiate. This provides insights into the slope evolution of lunar surface terrains under the influence of mass wasting (Xiao et al., 2013). Furthermore, comparing the details of topographic features on mass wasting in different craters enables exploration of whether mass wasting in mare settings differs significantly from that in highlands. Comparisons with terrestrial counterparts could also aid in understanding lunar surface processes (Kokelaar et al., 2017).

The modification of the surface topography by lunar mass movements focuses on two aspects: the modification of the sloping topography and the modification of the deposit topography. The lunar gravity gradually changed the slope height difference to a flatter topography by moving the slope material toward a region with a gentler topography. Therefore, mass movements generally occur preferentially over slopes forming lately, with a steeper slope angle (Xiao et al., 2013). Thus, the distribution of mass movements could provide a reference for relative age estimates in lunar topographic units, which was shown in Bickel et al. (2020) using the quantitative of rockfalls. In addition, the type and velocity of mass movements are different on the different side slopes. For example, older surfaces on the Moon generally have slopes much less than the angle of repose of lunar materials, where creeping of regolith is the dominate landslides (Xiao et al., 2013), and the speed at which the debris moves increases with slope steepness (Soderblom, 1970).

Compared to the same volume of landslides on the Earth, the lunar landslide has a much larger scale of mass wasting, resulting in a spectacular accumulation of landforms (Xiao et al., 2013). It is due to the Moon having a sixth gravity of the Earth, which means the material formed is less fragmented under the small gravity force of the Moon for the same material strength. Additionally, the weathered lunar regolith on the Moon's surface can lubricate coarse particles and reduce sliding friction, lengthening, extending the path of mass movements (Riller and Lieger, 2008), and giving the accumulation landform a slender appearance, as well as the regolith would damp the impact of individual particles, reducing fragmentation (Kokelaar et al., 2014). The density of lunar soil on the lunar slope is lower than that on the flat surface, and the bedrock erosion rate is higher than that on the flat surface because large-scale movement exposes the bedrock by clearing debris on the slope (Lin et al., 2020). In conclusion, the combined effect of small-scale microscopic gardening and mass movements will lessen the slope's fluctuation while burying the downhill and making it flat or fluctuate gently.

Studying the triggering of mass movements is inextricably linked to investigating its morphological features and spatial distribution traits. On the Moon, endogenic dynamic processes (such as volcanism, tectonism, tides, and body shrinkage) and seismic activity induced by impact cratering, as well as thermal stresses raised by diurnal temperature changes, are among the forces that could result in deformation and failure in slope materials (Watters et al., 2019; Head et al., 2020) (Figure 9).

Impact cratering includes asteroid impact, meteorite impact, etc. Different sized impact bodies have various impact effects, and seismic activity caused by impacts can have an impact on a much larger areas than the original impact craters [(Schultz, 1976a). The impact destroys the composition material's structure and original topography. The impact destabilizes the slope by causing a mass movement and changing its morphological characteristics, makes the slopes gentler and the surface regolith layer thicker and results in degraded craters (Bickel et al., 2022; Ikeda et al., 2022; Chandnani and Herrick, 2023)]. The impact of micrometeorites (0.1–2 mm) can also degrade the surface of the rock and possibly cause the granule on the surface to separate triggered debris flow and gardening produced agglutinate formation and related cohesion increase (Soderblom, 1970; Hörz et al., 1971; Basilevsky et al., 2013; Bandfield et al., 2014; Bickel et al., 2021; Bickel et al., 2022; Ikeda et al., 2022). Impact-related seismicity is shallow seismic, and tectonic activity can also be the cause of shallow seismic. Boulder falls are not caused by shallow moonquakes along lobate scarps but by ground shaking due to meteorite impacts (Ikeda et al., 2022).

The rate of diurnal temperature change is less than 2 K/min on the Moon ((Molaro and Byrne, 2012), but grain-scale stresses induced by diurnal temperature changes are on the order of 100 MPa (Molaro et al., 2015). This magnitude of stress is comparable to the typical strength of lunar rocks. Thus, the temperature changes lead to the grain reorientation caused by the differential expansion of the matter to induce mass movements. The strength of thermal weathering depends on the size of the rock mass's particles and albedo and thermal inertia, and some research has found that the thin regolith cover would prevent the initiation of thermal fracturing (Molaro et al., 2017; Ruesch et al., 2020). Besides thermal fracturing, thermal moonquakes, which happen only in extremely dry sand with a smooth surface may be another mechanism driving lunar landslides (Linasay, 1976; Bickel et al., 2021).

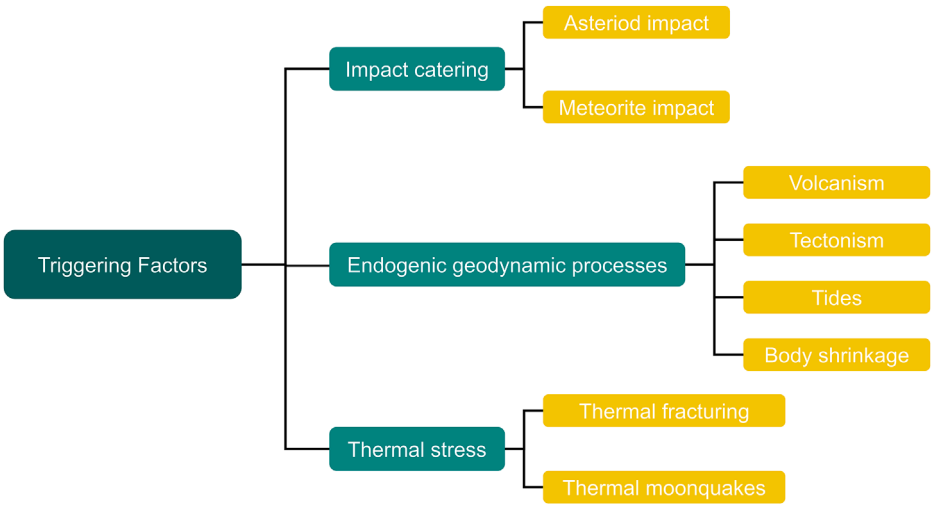


FIGURE 9
Triggering factors for mass movements on the Moon [Modified from Xiao. (2021)].

There are no affect factors like rainfall, melting snow, and frequent human activities of mass movements on the Moon compared to the mass movements on the Earth. Other factors, such as material composition, terrain conditions, lunar exploration and land missions, and space weathering, may play an important role in the formation and development of the lunar mass movements. In addition, whether the lunar soil contains water ice, what form of water storage, and how much water is still controversial, which will also have an important impact on the cause of the mass movements (Bart, 2007; McCubbin et al., 2015). At the same time, the mechanical properties of the rock composition material will affect the stability of the slope and the formation and development of the mass movements.

As lunar exploration engineering technology, such as the Artemis project (Barzilla et al., 2021) continues to advance, an increasing number of lunar orbiters and manned landers are impacting the lunar surface environment. According to recent news reports, scientists have found an orbiter component that has fallen into a fresh impact crater on the far side of the Moon. This intriguing finding has led us to speculate that external impact events of a similar nature could generate varying degrees of vibration across different scales of lunar surface regions. Such vibrations may potentially contribute to the occurrence of mass movements.

4.4 Future research direction

The research in this field describes a higher increase in the last 14 years (2009–2022), which has benefited from the dozens of successful manned and unmanned exploration Missions. In light of ongoing advancements in lunar exploration technology and an abundance of lunar exploration data, lunar mass movements have attracted the attention of an increasing number of researchers as an ideal window to study the transformation effects of internal and external forces on the morphology of the Moon's surface. However, the following problems still need to be solved.

At present, the recognition and mapping of mass movements still rely on visual interpretation of integrated expert knowledge, which is a time-consuming task (Nass et al., 2018), while suffering from artificial and technological limitations and biases. This contributes to the lack of catalogs on a global scale including all kinds of mass movements, and the lack of knowledge about the spatial distribution and density of mass movements. Research by Bickel et al. (2021), Bickel et al. (2022) used a machine learning-driven approach to solve this major limitation of human-driven, mapped 28,101 flow features and 136,610 boulder tracks between 60°N and S. These works opened the door to a statistically sound, global-scale analysis of the spatial distribution, density, and drivers of rockfalls and granular flows. However, a comprehensive inventory of other types of mass movements is still lacking. A morphological information sample database of multi-scale mass movements needs to be constructed to complete the global-scale database of mass movements, and machine learning or deep learning methods for identifying more types of lunar mass movements need to be experimented with. Then, a global-scale analysis of the spatial distribution, and density of all kinds of mass movements could be carried out.

Existing classification schemes for mass movements have unclear classification metrics. The current classification of the mass movements modified the classification system adopted by Varnes's (1978) for terrestrial landslides, which is classified by three indicators: movement process, slope morphology, and deposition characteristics (Kokelaar et al., 2017). Classifying lunar landslides based on the characteristics of terrestrial landslides would be incomprehensible. There are a series of problems, such as environmental simulation costs being high, explaining characteristic differences being difficult, and reference standards being inconsistent. It is urgent to establish a classification index system combining morphology, slip materials, and movement mode according to mass movements' morphology and spatial distribution characteristics. Rajaneesh et al. (2022) defined 15 critical parameters of the Martian landslide inventory viz., length (L), width (W),

relative relief (RR), area, L/W ratio, RR/L ratio, TI of landslide (mean, minimum, and maximum), and slope statistics (mean, minimum, and maximum), TI of the scarp, the slope of the scarp, and geology, which were analyzed using multiple discriminant analysis (MDA), it turns out that the parameters RR/L, L/W, and RR were suggested to be the parameters that primarily discriminates the Martian landslides. And then machine learning algorithms were used to evaluate the classification of landslides. This study demonstrates the potential of using machine learning models to classify extra-terrestrial landslides. Therefore, the main research direction of lunar landslide classification in the future is using machine learning algorithms to design a data-driven classification parameter system based on the spatial distribution characteristics and the morphological characteristics, integrating the environmental and geological information.

Triggers of the mass movements discussed above were short of compelling statistical evidence at present. This is partly limited by the lack of a comprehensive global-scale inventory. With the available data, only one or a few factors of a single type of mass movement could be conducted a quantitative assessment on a small scale by researchers, which cannot meet the demands of multi-scale big data analysis. The differences in the morphological characteristics and spatial distribution of mass movements were significant because its formation process is always influenced by the interaction of environmental and geological factors. In the study of terrestrial landslides, the landslide susceptibility model is usually constructed by multiple discriminant analysis of triggers such as terrain, hydrology, and human activities through algorithm models such as logistic model tree, random forest, support vector machine, and Bayesian (Chen et al., 2017; Pham et al., 2019; Sun et al., 2020; Xie et al., 2021). Thus, driven by machine learning and big data technology, there is a need to construct the mathematical model to illustrate the intersecting response mechanism of triggers for lunar landslides based on the multiple discriminant analysis of environmental, geological, and geomorphic on a multi-scale.

The study on the modification effect of mass movement on the lunar surface morphology is mostly a qualitative analysis, and without quantitative indicators and universality, as discussed above. In future studies, comparing the transformation scale under various types, terrain conditions, composition materials, space environment, and other conditions could be considered to develop a systematic evaluation index system, and to provide parameters for revealing the evolution process and mechanism of the lunar surface morphology.

It is anticipated that the study of lunar mass movements will advance significantly based on novel data sets acquired in the new era of lunar exploration, evolving computational methods, and GIS big data capabilities as a result of growing recognition of the significance of lunar mass movements research and the development of unmanned and manned lunar exploration projects.

5 Conclusion

The paper generates a database of articles about the lunar mass movements that reflects general information (e.g., data, tool, and method) related to the research of lunar mass movements and highlights the recent advances and the interest areas.

This study shows that the main and frontier research topics of lunar mass movements include the erosion of the surface by ejecta, geological phenomena, the lunar regolith, the evolution of planetary surface, the evolution of lunar impact craters, and thermodynamics. Icarus, JGR, and P&SS are the top three productive publishing journals. And research on lunar mass movements is dominated by a few countries, institutions, and scholars. California Institute of Technology, Arizona State University, and Johns Hopkins University are active institutions leading the lunar mass movements research, and the related institutions are mainly located in the United States, Russia, and China, considering their successful launch of lunar exploration missions as well as investments in research and development, it is not surprising. Frequent collaboration is carried out between scholars on the same or similar subjects to promote the development of research on the lunar mass movements.

Astronomy and physics are the primary disciplines, with research on the modification effect and degree of mass movements on the lunar surface being crucial to the exploration of the geological structure and evolution of the Moon. The citation index analysis indicates the growth of the academic impact of articles about the lunar mass movements. Moreover, co-citation network analysis, which provides important information about the research direction, revealed that comparative planetology was commonly used in the research on the characterization of lunar mass movements, and surface erosion, slope stability, and magnetosphere effects are some directions that are researched frequently.

Reviewing existing studies of lunar mass movements, we propose four unsolved issues. 1) The comprehensive database of lunar mass movements that records all types of landslides incidents on a global-scale has not been compiled. 2) The classification indicators to quantify the classification criteria for the lunar mass movement have not been definite. 3) The mathematical model to explain the triggering mechanism of mass movements on the Moon has not been built. 4) The quantitative indexes to evaluate the modification effect of mass movements on the lunar surface morphology have not been identified. Under the background of continuous innovation in lunar exploration technology and continuous enrichment of lunar exploration data, more types of mass movements should be identified and mapped on a global-scale with artificial intelligence technology to expand lunar landslide database, Multiple discriminant analysis and machine learning tools could be used to construct a lunar landslide classification system, and combined with feature datasets like crater size-frequency distribution, moonquake, impact flux, and topographic, driven by lunar landslide database, the causes and modification effects of mass movements could be clarified finally.

To sum up, this review provides a longitudinal survey of the published literature about the lunar mass movement in the last 151 years. It is hoped that these studies can inspire researchers from the field of lunar surface processes, especially new researchers, to understand the lunar mass movements.

Author contributions

JW conceived the idea of the study. KL wrote the paper. KL and JW interpreted the results. YS analyzed the data. All

authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by National Key Research and Development project (No. 2019YFE0123300), National Natural Science Foundation of China (No. 41801361), and the Fundamental Research Funds for the Central Universities (No. 2652019027).

Acknowledgments

The authors wish to give thanks to the valuable comments and suggestions given by two referees.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bandfield, J. L., Song, E., Hayne, P. O., Brand, B. D., Ghent, R. R., Vasavada, A. R., et al. (2014). Lunar cold spots: Granular flow features and extensive insulating materials surrounding young craters. *Icarus* 231, 221–231. doi:10.1016/j.icarus.2013.12.017

Bart, G. D. (2007). Comparison of small lunar landslides and martian gullies. *Icarus* 187, 417–421. doi:10.1016/j.icarus.2006.11.004

Barzilla, J., Semones, E., Lee, K., Johnson, A. S., Quinn, P., Whitman, K., et al. (2021). "Concept of operations and requirements for the Artemis mission," in *AGU fall meeting abstracts* (New Orleans, LA: AGU).

Basilevsky, A. T., Head, J. W., and Horz, F. (2013). Survival times of meter-sized boulders on the surface of the Moon. *Planet. Space. Sci.* 89, 118–126. doi:10.1016/j.pss.2013.07.011

Bhardwaj, A., and Sam, L. (2022). Reconstruction and characterisation of past and the most recent slope failure events at the 2021 rock-ice avalanche site in chamoli, Indian himalaya. *Remote Sens.* 14, 949. doi:10.3390/rs14040949

Bickel, V. T., Aaron, J., Manconi, A., and Loew, S. (2021). Global drivers and transport mechanisms of lunar rockfalls. *J. Geophys. Res. Planet.* 126 (10), 6824. doi:10.1029/2021JE006824

Bickel, V. T., Aaron, J., Manconi, A., Loew, S., and Mall, U. (2020). Impacts drive lunar rockfalls over billions of years. *Nat. Commun.* 11, 2862. doi:10.1038/s41467-020-16653-3

Bickel, V. T., Loew, S., Aaron, J., and Goedhart, N. (2022). A global perspective on lunar granular flows. *Geophys. Res. Lett.* 49 (12), 98812. doi:10.1029/2022GL098812

Boyce, J. M., Mouginis-Mark, P., and Robinson, M. (2020). The Tsiolkovskiy crater landslide, the moon: An LROC view. *Icarus* 337, 113464. doi:10.1016/j.icarus.2019.113464

Bray, V. J., Tornabene, L. L., Keszthelyi, L. P., McEwen, A. S., Hawke, B. R., Giguere, T. A., et al. (2010). New insight into lunar impact melt mobility from the LRO camera. *Geophys. Res. Lett.* 37, 4666. doi:10.1029/2010gl044666

Brunetti, M. T., Guzzetti, F., Cardinali, M., Fiorucci, F., Santangelo, M., Mancinelli, P., et al. (2014). Analysis of a new geomorphological inventory of landslides in Valles Marineris, Mars. *Earth Planet Sc. Lett.* 405, 156–168. doi:10.1016/j.epsl.2014.08.025

Brunetti, M. T., Xiao, Z., Komatsu, G., Peruccacci, S., and Guzzetti, F. (2015). Large rock slides in impact craters on the Moon and Mercury. *Icarus* 260, 289–300. doi:10.1016/j.icarus.2015.07.014

Bui, H. H., Kobayashi, T., Fukagawa, R., and Wells, J. C. (2009). Numerical and experimental studies of gravity effect on the mechanism of lunar excavations. *J. Terramechanics* 46, 115–124. doi:10.1016/j.jterra.2009.02.006

Burris, M., Dickenshied, S., Anwar, S., Anderson, Z., Hagee, W., Noss, D., et al. (2022). "Jmars 5: A modernized interface for planetary exploration analysis," in *AGU fall meeting abstracts* (Chicago: AGU).

Chandnani, M., and Herrick, R. R. (2023). Influence of target properties on wall slumping in lunar impact craters within the simple-to-complex transition. *J. Geophys. Res. Planets.* 128 (1), 7369. doi:10.1029/2022JE007369

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., et al. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* 151, 147–160. doi:10.1016/j.catena.2016.11.032

Chin, G., Brylow, S., Foote, M., Garvin, J., Kasper, J., Keller, J., et al. (2007). Lunar reconnaissance orbiter overview: The Instrument suite and mission. *Space Sci. Rev.* 129, 391–419. doi:10.1007/s11214-007-9153-y

Corominas, J., van Westen, C., Frattini, P., Cascini, L., Malet, J. P., Fotopoulou, S., et al. (2014). Recommendations for the quantitative analysis of landslide risk. *B Eng. Geol. Environ.* 73, 209–263. doi:10.1007/s10064-013-0538-8

Corsini, A., Borgatti, L., Cervi, F., Dahne, A., Ronchetti, F., and Sterzai, P. (2009). Estimating mass-wasting processes in active Earth slides - Earth flows with time-series of High-Resolution DEMs from photogrammetry and airborne LiDAR. *Nat. Hazard Earth Sys.* 9, 433–439. doi:10.5194/nhess-9-433-2009

Craddock, R. A., and Howard, A. D. (2000). Simulated degradation of lunar impact craters and a new method for age dating farside mare deposits. *J. Geophys. Res. Planet.* 105, 20387–20401. doi:10.1029/1999je001099

Cruden, D. M., and Varnes, D. J. (1996). in *Landslide types and processes*. Editors A. K. Turner, and R. L. Schuster (Washington, DC: National Academy Press), 36–75.

Deplus, C., Friant, A. L., Boudon, G., Komorowski, J. C., Villemant, B., Harford, C., et al. (2001). Submarine evidence for large-scale debris avalanches in the Lesser Antilles Arc. *Earth Planet Sc. Lett.* 192, 145–157. doi:10.1016/s0012-821x(01)00444-7

Greenhagen, B. T., Neish, C. D., Williams, J. P., Cahill, J. T., Ghent, R. R., Hayne, P. O., et al. (2016). Origin of the anomalously rocky appearance of Tsiolkovskiy crater. *Icarus* 273, 237–247. doi:10.1016/j.icarus.2016.02.041

Guest, J. E., and Murray, J. B. (1969). Nature and origin of Tsiolkovsky crater, lunar farside. *Planet Space Sci.* 17, 121–141. doi:10.1016/0032-0633(69)90128-7

Hargitai, H., and Kereszturi, Á. (2015). *Encyclopedia of planetary landforms*. New York, NY, USA: Springer, 1–2460.

Head, J. W., Wilson, L., Deutsch, A. N., Rutherford, M. J., and Saal, A. E. (2020). Volcanically induced transient atmospheres on the moon: Assessment of duration, significance, and contributions to polar volatile traps. *Geophys. Res. Lett.* 47, 89509. doi:10.1029/2020gl089509

Highland, L. M., and Bobrowsky, P. (2008). The landslide handbook: A guide to understanding landslides. *U. S. Geol. Surv. Circ.* 1325, 3–25. doi:10.3133/cir1325

Hörz, F., Hartung, J. B., and Gault, D. E. (1971). Micrometeorite craters on lunar rock surfaces. *J. Geophys. Res.* 76, 5770–5798. doi:10.1029/jb076i023p05770

Huang, Y., and Mao, W. (2013). First results derived from a drop-tower testing system for granular flow in a microgravity environment. *Landslides* 10, 493–501. doi:10.1007/s10346-013-0403-7

Ikeda, A., Kumagai, H., and Morota, T. (2022). Topographic degradation processes of lunar crater walls inferred from boulder falls. *J. Geophys. Res. Planets.* 127 (10), 7176. doi:10.1029/2021JE007176

Kereszturi, A. (2023). Polar ice on the moon. *Encycl. Lunar Sci.* 1, 9. doi:10.1007/978-3-319-05546-6_216-1

Kereszturi, A. (2014). Surface processes in microgravity for landing and sampling site selection of asteroid missions—suggestions for MarcoPolo-R. *Planet. Space. Sci.* 101, 65–76. doi:10.1016/j.pss.2014.06.005

Kereszturi, A., Tomka, R., Gläser, P. A., Pal, B. D., Steinmann, V., and Warren, T. (2022). Characteristics of de Gerlache crater, site of girlands and slope exposed ice in a lunar polar depression. *Icarus* 388, 115231. doi:10.1016/j.icarus.2022.115231

Kokelaar, B. P., Bahia, R. S., Joy, K. H., Viroulet, S., and Gray, J. M. (2017). Granular avalanches on the Moon: Mass-wasting conditions, processes, and features. *J. Geophys. Res. Planet.* 122, 1893–1925. doi:10.1002/2017je005320

Kokelaar, B. P., Graham, R. L., Gray, J. M., and Vallance, J. W. (2014). Fine-grained linings of leveed channels facilitate runout of granular flows. *Earth Planet Sc. Lett.* 385, 172–180. doi:10.1016/j.epsl.2013.10.043

Kumar, P. S., Sruthi, U., Krishna, N., Lakshmi, K. J. P., Menon, R., Amitabh, G., et al. (2016). Recent shallow moonquake and impact-triggered boulder falls on the Moon: New insights from the Schrodinger basin. *J. Geophys. Res. Planet.* 121, 147–179. doi:10.1002/2015je004850

Li, B., Ling, Z., Zhang, J., Chen, J., Liu, C., and Bi, X. (2018). Geological mapping of lunar highland crater Lalande: Topographic configuration, morphology and cratering process. *Planet Space Sci.* 151, 85–96. doi:10.1016/j.pss.2017.11.007

Liao, Y., Jiang, M., Chen, Y., Xi, B., and Tan, Y. (2018). Discrete element analysis of landslide mechanism on lunar soil induced by different slope angles. *J. Water Resour. Water Eng.* 29, 194–200. doi:10.11705/j.issn.1672-643X.2018.02.32

Liever, P. A., Gale, M. P., Mehta, R. S., and West, J. S. (2018). *Gas-granular simulation framework for spacecraft landing plume-surface interaction and debris transport analysis, in Granular Materials in Space Exploration*. Ohio: ASCE Earth & Space, 39–48. doi:10.1061/9780784481899.005

Lin, Q., Cheng, Q., Li, K., Xie, Y., and Wang, Y. (2020). Contributions of rock mass structure to the emplacement of fragmenting rockfalls and rockslides: Insights from laboratory experiments. *J. Geophys. Res. Sol. Ea.* 125, 19296. doi:10.1029/2019JB019296

Lindsay, J. (1976). "Energy at the lunar surfaces," in *Lunar stratigraphy and sedimentology*. Editors Z. Kopal, and A. G. Cameron (Amsterdam: Developments in Solar System and Space Science), 45–55.

Mahanti, P., Robinson, M. S., Humm, D. C., and Stopar, J. D. (2014). A standardized approach for quantitative characterization of impact crater topography. *Icarus* 241, 114–129. doi:10.1016/j.icarus.2014.06.023

Malin, M. C., and Edgett, K. S. (2000). Evidence for recent groundwater seepage and surface runoff on Mars. *Science* 288, 2330–2335. doi:10.1126/science.288.5475.2330

McCubbin, F. M., Vander Kaaden, K. E., Tartèse, R., Klima, R. L., Liu, Y., Mortimer, J., et al. (2015). Magmatic volatiles (H, C, N, F, S, Cl) in the lunar mantle, crust, and regolith: Abundances, distributions, processes, and reservoirs. *Am. Mineral.* 100, 1668–1707. doi:10.2138/am-2015-4934ccbyncnd

McEwen, A. S. (1989). Mobility of large rock avalanches - evidence from valles marineris, Mars. *Geology* 17, 1111–1114. doi:10.1130/0091-7613(1989)017<1111:molrae>2.3.co;2

Melosh, H. J. (2011). *Planetary surface processes*. Cambridge: Cambridge University Press.

Molaro, J., and Byrne, S. (2012). Rates of temperature change of airless landscapes and implications for thermal stress weathering. *J. Geophys. Res. Plane* 117, 413. doi:10.1029/2012JE004138

Molaro, J. L., Byrne, S., and Langer, S. A. (2015). Grain-scale thermoelastic stresses and spatiotemporal temperature gradients on airless bodies, implications for rock breakdown. *J. Geophys. Res. Planet.* 120, 255–277. doi:10.1002/2014je004729

Molaro, J. L., Byrne, S., and Le, J. L. (2017). Thermally induced stresses in boulders on airless body surfaces, and implications for rock breakdown. *Icarus* 294, 247–261. doi:10.1016/j.icarus.2017.03.008

Moore, J. G., Normark, W. R., and Holcomb, R. T. (1994). Giant Hawaiian landslides. *Annu. Rev. Earth Pl. Sc.* 22, 119–144. doi:10.1146/annurev.ea.22.050194.001003

Morse, Z. R., Osinski, G. R., and Tornabene, L. L. (2018). Morphologic mapping and analysis of Tsiolkovsky crater ejecta. *Lunar Plant Sci.* 56, 13650. doi:10.1111/maps.13650

Murdoch, N., Sánchez, P., Schwartz, S. R., and Miyamoto, H. (2015). Asteroid surface geophysics. arXiv preprint arXiv:1503.01931.

Nass, A., Di, K., Elgner, S., van Gasselt, S., Hare, T., Hargitai, H., et al. (2018). Planetary cartography–activities and current challenges. *Proc. ICA.* 1, 81–10. doi:10.5194/ica-proc-1-81-2018

Otto, K. A., Krohn, K., Balzer, J., Hoffmann, H., and Jaumann, R. (2014). *Comparison of mass-wasting features in craters on asteroid 4 Vesta and the moon*. (London: European Lunar Symposium), 15–16.

Parker, R. N., Densmore, A. L., Rosser, N. J., de Michele, M., Li, Y., Huang, R., et al. (2011). Mass wasting triggered by the 2008 Wenchuan earthquake is greater than orogenic growth. *Nat. Geosci.* 4, 449–452. doi:10.1038/ngeo1154

Perumal, S. K. (2012). "Gullies and landslides on the Moon: An example from Schrodinger basin," in *39th COSPAR scientific assembly* (Mysore, India: COSPAR), 1487.

Pham, B. T., Jaafari, A., Prakash, I., and Bui, D. T. (2019). A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling. *Bull. Eng. Geol. Environ.* 78, 2865–2886. doi:10.1007/s10064-018-1281-y

Pike, R. J. (1971). *Some preliminary interpretations of lunar mass-wasting process from Apollo 10 photography, analysis of Apollo 10 photography and visual observations*. Washington, D. C: Work of the US Gov. Public Use Permitted, 14–20.

Pradhan, S. P., and Siddique, T. (2019). "Mass wasting: An overview," in *Landslides: Theory, practice and modelling*. Editors S. P. Pradhan, V. Vishal, and T. Singh (Cham: OpenGeology). doi:10.1007/978-3-319-77377-3_1

Rajaneesh, A., Vishnu, C. L., Oommen, T., Rajesh, V. J., and Sajinkumar, K. S. (2022). Machine learning as a tool to classify extra-terrestrial landslides: A dossier from valles marineris, Mars. *Icarus* 376, 114886. doi:10.1016/j.icarus.2022.114886

Riller, U., and Lieger, D. (2008). Geological evidence for acoustic fluidization in large impact structures. *Large Meteor. Impacts Planet. Evol. IV* 1423, 3025.

Ritter, D. F., Kochel, C. R., and Miller, J. R. (2006). *Process geomorphology*. 4th edn. Long Grove: Waveland Press, 79–133.

Ruesch, O., Sefton-Nash, E., Vago, J. L., Kuppers, M., Pasckert, J. H., Khron, K., et al. (2020). *In situ* fragmentation of lunar blocks and implications for impacts and solar-induced thermal stresses. *Icarus* 336, 113431. doi:10.1016/j.icarus.2019.113431

Sam, L., and Bhardwaj, A. (2022). A remote sensing perspective on mass wasting in contrasting planetary environments: Cases of the moon and ceres. *Remote Sens.* 14, 1049. doi:10.3390/rs14041049

Scaioni, M., Yordanov, V., Brunetti, M. T., Melis, M. T., Zinzi, A., Kang, Z., et al. (2018). Recognition of landslides in lunar impact craters. *Eur. J. Remote Sens.* 51, 47–61. doi:10.1080/22797254.2017.1401908

Schenk, P. M., and Bulmer, M. H. (1998). Origin of mountains on io by thrust faulting and large-scale mass movements. *Science* 279, 1514–1517. doi:10.1126/science.279.5356.1514

Schultz, P. H. (1976a). Floor-fractured lunar craters. *Moon* 15, 241–273. doi:10.1007/bf00562240

Schultz, P. H. (1976b). *Moon morphology: Interpretations based on lunar orbiter photography*. Austin: University of Texas Press.

Shaller, P. J., and Komatsu, G. (1994). Landslides on Mars. *Landslide News* 8, 18–22.

Simon, J., and de Bruyn, J. R. (2007). Shape of impact craters in granular media. *Phys. Rev. E.* 76, 041306. doi:10.1103/physreve.76.041306

Soderblom, L. A. (1970). A model for small-impact erosion applied to the lunar surface. *J. Geophys. Res.* 75, 2655–2661. doi:10.1029/jb075i014p02655

Sun, D., Wen, H., Wang, D., and Xu, J. (2020). A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. *Geomorphology* 362, 107201. doi:10.1016/j.geomorph.2020.107201

Varnes, D. J. (1978). in *Slope movement types and processes*. Editors R. L. Schuster, and R. J. Krizek (Washington, D. C: National Academy of Sciences),11–33.

Waltham, D., Pickering, K. T., and Bray, V. J. (2008). Particulate gravity currents on Venus. *J. Geophys. Res. Planet.* 113, E02012. doi:10.1029/2007JE002913

Watters, T. R., Weber, R. C., Collins, G. C., Howley, I. J., Schmerr, N. C., and Johnson, C. L. (2019). Shallow seismic activity and young thrust faults on the Moon. *Nat. Geosci.* 12, 411–417. doi:10.1038/s41561-019-0362-2

Wolman, M. G., and Gerson, R. (1978). Relative scales of time and effectiveness of climate in watershed geomorphology. *Earth Surf. Proc.* 3, 189–208. doi:10.1002/esp.3290030207

Wu, S. S., Schafer, F. J., Jordan, R., Nakata, G. M., and Derick, J. L. (1972). Photogrammetry of Apollo 15 photography. *Apollo Prelim. Sci. rept.* 289, 36.

Wu, S., Wang, Y., Hou, X., Xue, P., and Long, L. (2020). Research on theoretical model of vertical impact of foot pad on lunar soil. *Manned Spacefl.* 26, 135–141. doi:10.3969/j.issn.1674-5825.2020.02.001

Xiao, Z., and Komatsu, G. (2013). Impact craters with ejecta flows and central pits on Mercury. *Planet Space Sci.* 83, 62–78. doi:10.1016/j.pss.2013.03.015

Xiao, Z. (2021). "Mass-movements on the moon," in *Treatise on geomorphology*. Editor F. John (Jack), Shroder (Academic Press: Elsevier, Oxford), 458–476. doi:10.1016/B978-0-12-818234-5.00047-X

Xiao, Z., Zeng, Z., Ding, N., and Molaro, J. (2013). Mass wasting features on the Moon - how active is the lunar surface? *Earth Planet Sc. Lett.* 376, 1–11. doi:10.1016/j.epsl.2013.06.015

Xie, W., Nie, W., Saffari, P., Robledo, L. F., Descote, P. Y., and Jian, W. (2021). Landslide hazard assessment based on Bayesian optimization–support vector machine in Nanping City, China. *Nat. Hazards.* 109 (1), 931–948. doi:10.1007/s11069-021-04862-y

Yordanov, V., Scaioni, M., Brunetti, M. T., Melis, M. T., Zinzi, A., and Giommi, P. (2016). Mapping landslides in lunar impact craters using Chebyshev polynomials and DEM'S. *Comm. VI* 41, 17–24. doi:10.5194/isprsarchives-XLI-B6-17-2016

Zheng, H., and Huang, Y. (2015). Model tests on flow slide of lunar regolith simulant. *Environ. Earth Sci.* 73, 4853–4859. doi:10.1007/s12665-014-3767-7

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 04 August 2023
DOI 10.3389/fspas.2023.1203845

Check for updates

OPEN ACCESS

EDITED BY
Alberto Fairén,
Spanish National Research Council (CSIC), Spain

REVIEWED BY
Manasvi Lingam,
Florida Institute of Technology, United States
William D. Orsi,
Ludwig Maximilian University of Munich, Germany

*CORRESPONDENCE
Ricardo Amils,
ramils@cbm.csic.es

RECEIVED 11 April 2023
ACCEPTED 25 July 2023
PUBLISHED 04 August 2023

CITATION
Escudero C and Amils R (2023), Hard rock dark biosphere and habitability. *Front. Astron. Space Sci.* 10:1203845. doi: 10.3389/fspas.2023.1203845

COPYRIGHT
© 2023 Escudero and Amils. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Hard rock dark biosphere and habitability

Cristina Escudero[1] and Ricardo Amils[2,3]*

[1]Geomicrobiology, Center for Applied Geoscience, University of Tübingen, Tübingen, Germany, [2]Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Universidad Autónoma de Madrid, Torrejón de Ardoz, Spain, [3]Planetology and Habitability Department, Centro de Astrobiología (CAB), INTA-CSIC, Torrejón de Ardoz, Spain

The discovery that most of the prokaryotic diversity and biomass on Earth resides in the deep subsurface, calls for an improved definition of habitability, which should consider the existence of dark biospheres in other planets and moons of the Solar System and beyond. The discovery of "interior liquid water worlds" on some ice moons with waterless surfaces has piqued wide astrobiological interest, but the sporadic mentions of the possibility of life in the deep subsurface of rocky planets in recent habitability reviews calls for a methodical effort to develop sufficient knowledge, both scientific and technological, to include the dark biospheres in our habitability assessments. In this review we analyze recent developments and the methodologies employed to characterize Earth's continental hard rock deep subsurface to both prepare the future exploration of the putative dark biosphere of Mars and to highlight its importance when evaluating planetary habitability.

KEYWORDS

hard rock deep subsurface, dark biosphere, habitability, Mars, liquid water, energy

Habitability

Habitability is defined as the potential of a given planetary environment to sustain life (Cockell et al., 2016 and references therein). Obviously, this concept depends on the conditions required for life to develop, among them energy, liquid solvent to facilitate mass transfer reactions and temperatures compatible with cell functions and the liquid state of the solvent (Kasting and Catling, 2003; Nisbet et al., 2007; Lammer et al., 2009; Cockell, 2014; Cockell, 2020; Lignam and Loeb, 2020; Lignam and Loeb, 2021). In addition, access to the elements required for the synthesis of cellular components is required. In the only reference system that we have, these are: carbon (C), hydrogen (H), oxygen (O), nitrogen (N), phosphorous (P) and sulfur (S). These conditions are determined by astronomical factors such as star type, planetary rotational and orbital characteristics, presence of moons, impact events and planetary factors such as mass, atmospheric and surface characteristics, plate tectonics and magnetic fields (Cockell et al., 2016; Cockell, 2020; Lingam and Loeb, 2021). Astrophysicists long ago defined the concept of planetary habitability zone (HZ) as the annular area around a main sequence star in which rocky planets with an Earth like mass can harbor liquid water in its surface (Maunder, 1913; Huang, 1959; Dole, 1964; Hart, 1979; Kasting et al., 1993). Among the high numbers of planets detected outside the Solar System (exoplanets) up to now, there is a significant proportion that meets these habitability conditions (https://exoplanets.nasa.gov/what-is-an-exoplanet/planet-types/terrestrial/). The discovery of extremophilic organisms able to thrive in extreme

environmental conditions of temperature, pH, ionic concentration, water activity, pressure, radiation, etc… has increased the probability of finding life in conditions that 50 years ago were unimaginable (Merino et al., 2019).

The astrobiological interest of habitability has been recently reviewed by several researchers (Cockell et al., 2016; Méndez et al., 2021). Cockell and collaborators constrained their habitability concept to known life to avoid the problem of defining life and because our ignorance of whether terrestrial life is universal. They distinguish between surface liquid water worlds and interior liquid water worlds, such as icy moons and terrestrial-type rocky planets where liquid water is only in their subsurface. This is the first mention on a habitability review of a possible internal source of energy generating liquid water in the interior of a planetary body. The focus of Méndez and collaborators is on Habitability Models for Astrobiology. In their review they recognize that "identifying water in the atmosphere of planets (in addition to other biosignature-relevant gases) is the only way to narrow down potential life-hosting targets, since subsurface life deep in the interior may not be able to modify the atmospheres of planets enough to be detectable remotely", pointing out the limitations of habitability exploration of rocky exoplanets (Méndez et al., 2021). Furthermore, the sentence quoted above is the only reference to deep subsurface life in the entire review. Interestingly, due to the experience of some of the authors in deep subsurface research, McMahon and collaborators elaborated the term "subsurface-habitability zone" (SSHZ) in 2013, acknowledging the proven existence of life in the deep subsurface of Earth as well as the possibility of interior liquid water on other planets depending on distance from the star and internal planetary heat (McMahon et al., 2013; MacMahon and Parnell, 2014). Despite its importance, this concept has not received the attention it deserved, most likely because research on hard rock deep subsurface is still very limited, and an in-depth discussion of the concepts required to analyze the habitability of a planet remains to be seen.

Although it is recognized the importance of the very prolific deep subsurface ocean research in the development of the dark biosphere concept, in this work only results from the deep oceanic crust are considered, because most of the work performed in the deep subsurface ocean has been done in sediments, requiring the presence of liquid water on the surface of the rocky planet, which was not the aim of the review. Furthermore, even if it is recognized the interest of the subsurface oceans detected in some ice moons from the Solar System, they are not considered also because do not fit in the hard rock deep subsurface concept, subject of the work. In this review we analyze the recent developments in characterizing the Earth continental hard rock deep subsurface, the methodologies used for its study as preparation for the exploration of the deep subsurface of Mars and to emphasize the need to include the dark biospheres in our habitability assessments.

The Earth hard rock dark biosphere

It took almost two hundred years to demonstrate Darwin's prediction of existence of life in the deep subsurface (Darwin, 1839). Although some pioneering observations were made at the beginning of the last century by taking advantage of samples obtained from oil drilling operations (Bastin et al., 1926), the consensus was that the detected microorganisms were contaminants resulting from the drilling operation. This is one of the biggest problems in this area of research (Lipman, 1931). The situation changed radically after the discovery of life associated to submarine hydrothermal vents (Corliss et al., 1979), which opened a very fruitful era of oceanic drilling projects searching for life in marine sediments and in the deep oceanic crust (Whelan et al., 1986; D´Hondt et al., 2007; Edwards et al., 2012a; Edwards et al., 2012b; Kallmeyer et al., 2012; Inagaki et al., 2015; Sueoka et al., 2019; Li et al., 2020), although the pioneering work of Morita and Zobell many years before should be underlined (Morita and Zobell, 1955). The methodologies developed to control the different sources of contamination during drilling was a crucial advance and dissipated skepticism around the study of the dark biosphere (Kieft, 2010). In 1992 Tomas Gold speculated, in his seminal paper, on the existence of life in the subsurface independent of radiation and the possibility of its presence in other planets (Gold, 1992). Since then, several studies have proved the existence of a great microbial diversity in the continental deep subsurface (also known as terrestrial deep subsurface) (Stevens and McKinley, 1995; Chapelle et al., 2002; Zhang et al., 2005; Suzuki et al., 2013; Wu et al., 2016; Momper et al., 2017b; Purkamo et al., 2018; Sherwood Lollar et al., 2019; Nuppunen-Puputti et al., 2022; Soares et al., 2023). "Life Underground" funded by the NASA Astrobiology Institute (NAI) and the "Center for Dark Energy Biosphere" funded by the National Science Foundation have been two interesting initiatives to foster deep subsurface research. Recent evaluations assert that most of the prokaryotic biomass and diversity on this planet is, in fact, within the deep subsurface (Colman et al., 2017; Bar-On et al., 2018; Magnabosco et al., 2018). Thus, incorporating dark biospheres into our habitability assessments should be a high priority.

Despite the important advances made in the methodologies to recover and analyze samples from the continental deep subsurface, the number of studies using pristine core samples from devoted drilling operations is still limited, mainly due to economic and technical limitations (Zhang et al., 2005; Fernández-Remolar et al., 2008; Gronstal et al., 2009; Wu et al., 2016; Amils et al., 2023; Templeton and Caro, 2023). For this reason, many researchers have taken advantage of samples obtained from different "subsurface windows", such as artesian wells, springs, radioactive waste disposal and underground research facilities, or deep mining operations. Most of them rely on groundwater, as it is much easier to collect and analyze, generating useful microbial diversity information. Nevertheless, these studies lack information on the relationship of microorganisms with the mineral components of the complex subsurface solid matrix in which they develop and on the interconnectivity required to understand the critical operation of deep subsurface biogeochemical cycles, essential to their successful performance (Stevens and McKinley, 1995; Pedersen, 1999; Chapelle et al., 2002; Murakami et al., 2002; Onstott et al., 2003; Sahl et al., 2008; Suzuki et al., 2013; Magnabosco et al., 2014; Momper et al., 2017a; Purkamo et al., 2018; Sherwood Lollar et al., 2019; Nuppunen-Puputti et al., 2022).

Unfortunately, the exploration of the deep subsurface biosphere in igneous ocean crust is very recent, mainly due to technical difficulties, thus very little microbiological information has been generated from the few studied areas (Edwards et al., 2012a; Li et al.,

2020), although preliminary results showed some general trends already observed in the hard rock continental deep subsurface (Li et al., 2020; Quemener et al., 2020).

Microbial cell density and diversity

The microbial cell number for the dark biosphere detected using groundwater samples from continental subsurface varies between 10^2 and 10^7 cells/mL (Pedersen, 2000; Basso et al., 2009; Itävaara et al., 2011). These values are much lower, up to 10^5 cells/gr, if solid, low porosity rocks with variable mineralogical content are considered (Onstott et al., 2003; Zhang et al., 2005; Fry et al., 2009; Breuker et al., 2011; Cockell et al., 2012; Dutta et al., 2018; Cockell et al., 2021; Amils et al., 2023). In the case of the oceanic crust the cellular density seems to be even lower, up to 10^3 cells/mL (Li et al., 2020). Considering these numbers and the suggestion that life can develop to a depth of 5–10 km, being temperature the most important limiting factor (Gold, 1992), we must conclude that the percentage of microbial life in the dark biosphere is considerable (Whitman et al., 1998; Kallmeyer et al., 2012; Bar-On et al., 2018; Magnabosco et al., 2018). Consequently, an important challenge is to identify the sources of energy that maintain this biomass.

Even though there is great variability of microbial populations identified in the continental deep subsurface, mainly due to the geological characteristics of the drilling locations and the methodologies used for its analysis, some general conclusions can be extracted from what we know so far. 1) The number of microorganisms decreases with depth (Moser et al., 2005; Itävaara et al., 2011; Cockell et al., 2012; McMahon and Parnell, 2014), although in some drilling operations this decrease is not observed (Amils et al., 2023). 2) Microbial diversity, in general, tends to decrease also with depth (Zhang et al., 2005; Lin et al., 2006; Chivian et al., 2008), although some exceptions have been reported (Itävaara et al., 2011; Amils et al., 2023). 3) Abundance and diversity of Bacteria is much higher than Archaea (Takai et al., 2001; Cockell et al., 2012; Ino et al., 2016; Lau et al., 2016; Rempfert et al., 2017; Amils et al., 2023). 4) Pseudomonadota, Actinomycetota and Bacillota have been described as the most common phyla in the continental deep subsurface (Zhang et al., 2005; Fry et al., 2009; Onstott et al., 2009; Magnabosco et al., 2016; Wu et al., 2016; Nuppunen-Puputti et al., 2022; Amils et al., 2023; Soares et al., 2023), although members of other phyla (Nitrospirota; Chloroflexota, Acidobacteriota and Deinococcota) and candidate phyla have been also reported. 5) New groups of unknown microorganisms have also been discovered (Takai et al., 2001; Gihring et al., 2006; Sahl et al., 2008).

Viruses have also been detected (Kyle et al., 2008; Eydal et al., 2009; Lau et al., 2014; Nyyssönen et al., 2014; Labonté et al., 2015; Rahlf et al., 2021), but regrettably, the information on the subsurface viral community is still very scarce impeding a clear vision of the role they play in the ecosystem. Fungi and nematodes are the only eukaryotes identified as members of the continental subsurface community, although the latter has only been identified in water samples from a subterranean mine (Borgonie et al., 2011). The fungal community, on the other hand, has been detected on multiple occasions in hard rock samples from different locations (Pedersen, 1997; Purkamo et al., 2013; Sohlberg et al., 2015; Ivarson et al., 2018) and recently in deep ocean crust (Quemener et al., 2020). Despite representing a very small fraction of the subsurface community and the little information currently available, they could play a very important role in the dissolution of recalcitrant organic matter and minerals, thus actively contributing to the energy input into the ecosystem, as have been suggested in marine sediments and in the deep ocean crust (Quemener et al., 2020; Reese et al., 2021).

Sources of energy in the continental dark biosphere

Four are the basic requirements for a habitable continental deep subsurface environment: energy, liquid water, adequate temperature and bioessential elements (Hoehler, 2004). The hard rock deep subsurface is characterized by the lack of solar radiation, in general, absence of O_2, and an increase in temperature and pressure with depth (Kieft, 2016). Obviously, the geological characteristics of the substrate: mineral composition, porosity, flow of water, presence of faults, will have an important influence in the development of microorganisms in the deep subsurface (Fredrickson et al., 1997; Pedersen, 2000; Amend et al., 2011; Li et al., 2020; Templeton and Caro, 2023).

Some authors maintain that most microorganisms from the continental deep subsurface are in the state of anabiosis, which means that they are not metabolically active but in a state of dormancy, with minimal energy consumption, which agrees with the observation of fast development of metabolically active microorganisms in enrichment cultures (D'Hondt et al., 2002; Rajala et al., 2015; Leandro et al., 2018). Others, using methodologies like Fluorescence *in situ* Hybridization (FISH), which can detect microorganisms with a significant number of ribosomes (Hoshino et al., 2008), suggest that the deep subsurface microorganisms are metabolically active (Escudero et al., 2018). Obviously, these observations are dependent on the characteristics of the drilled substrate, the type of samples, and the methodologies used, which makes generalizing extremely difficult.

To the best of our knowledge, no one has been able to measure the activity of microorganisms *in situ* in the rock continental subsurface. Viable microorganisms have been detected after long periods of time isolated in subsurface (Vreeland et al., 1998; Orsi et al., 2021) and it has been theorized that they could only survive under these conditions if they used some energy to avoid DNA and protein damage (Morita, 1999). Some researchers have used amino acid racemization to evaluate microbial deep subsurface doubling times (Steen et al., 2013). Indeed, some studies have corroborated that the strategy of repairing damaged cellular structures seems to be favored over dormancy in subsurface environments (Johnson et al., 2007) and that carbon turnover, at least in ocean sediments and continental groundwater samples, can vary from 1 year to several decades (Jørgensen, 2011; Onstott et al., 2014; Becraft et al., 2021). Thus, the possibility of extremely low levels of growth, in other words, geological duplication times, must also be considered in hard rock continental environments (Phelps et al., 1994; Steen et al., 2013).

An extremely critical and debated issue in the study of continental deep subsurface ecology is whether the energy sources should be endogenous or compatible with the existence of

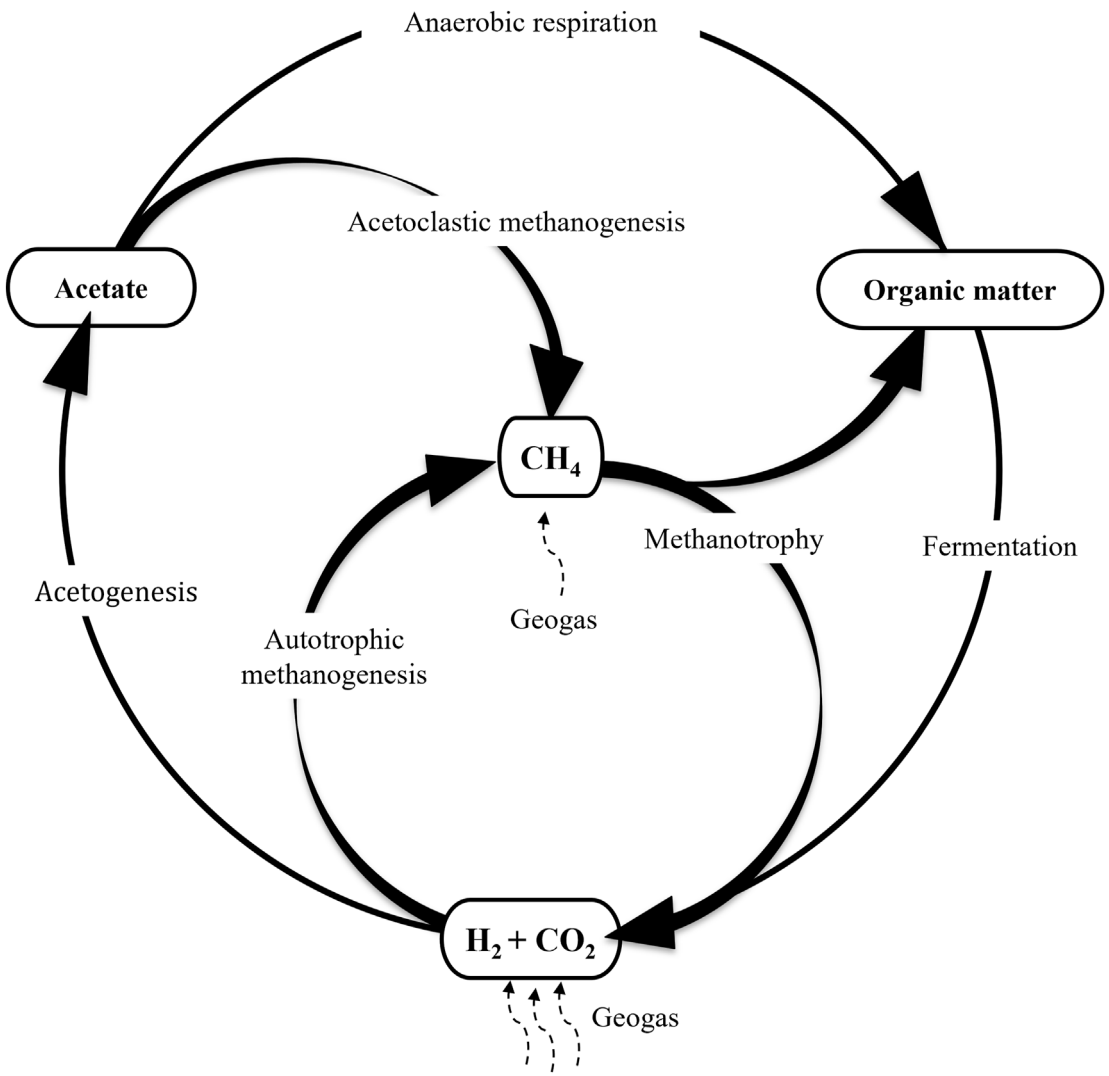


FIGURE 1
Proposed model in which H_2 is the main energy source for primary productivity in the deep subsurface. Figure modified from Pedersen (1997).

products generated in the surface. Purists maintain that only microorganisms that develop in the absence of any photosynthetic product should be considered members of the continental deep subsurface (Stevens and McKinley, 1995; Nealson et al., 2005; Orcutt et al., 2011; Momper et al., 2017b). These environments have been named Subsurface Lithoautotrophic Microbial Ecosystems (SLiMEs). But the fact that chemoheterotrophic microorganisms, i.e., microorganisms that use organic carbon as an energy source, are detected in abundance in the subsurface and the biomass must be recycled to keep geobiological cycles operative, makes this a difficult conundrum. Knowing that O_2 is rapidly consumed by aerobic and facultative microorganisms, we should consider the continental hard rock deep subsurface an anoxic environment, although exceptions may occur (Ruff et al., 2023), in which only anaerobic metabolisms prevail. Thus, anaerobic respiration is the main energy generation system, and due to the shortage of organic matter, minerals are the main source of electron donors and acceptors (Jones and Bennett, 2017; Rempfert et al., 2017; Amils et al., 2023; Templeton and Caro, 2023), although fermentation should also be considered an important recycling system (Sanz et al., 2021), as it facilitates the operation of the biogeochemical cycles (Amils et al., 2023). It is well established that different microorganisms can directly use minerals as electron donors and acceptors (Shock, 2009; El-Naggar et al., 2010) or dissolve minerals to release compounds to be used as energy source in chemolithotrophic metabolisms (Rogers et al., 1998; Edwards et al., 2000; Shelobolina et al., 2012; Vera et al., 2013; Dong et al., 2014; Osburn et al., 2014; Templeton and Caro, 2023).

Because H_2 is one of the most abundant gases in the continental deep subsurface it has been considered an important source of energy for chemolithoautotrophic microorganisms (Stevens and McKinley, 1995; Pedersen, 1997). According to the model advanced by these authors H_2 is the main driver of a system in which hydrogenotrophic methanogenesis and acetogenesis constitute the basis of the ecosystem (Figure 1), releasing methane (CH_4) and acetate and generating the biomass to be used by heterotrophic

microorganisms that, in turn, contribute to the maintenance of an active C cycle, which has been recently shown to be operating in the deep subsurface (Amils et al., 2023). Hydrogen could be produced abiotically including serpentinization and water radiolysis (Apps and van de Kamp, 1993; Stevens and McKinley, 1995; Lin et al., 2004; Sherwood-Lollar et al., 2007; Klein et al., 2020; Coskun et al., 2021; Leong et al., 2021; Templeton and Caro, 2023), but recently it has been shown that an important amount of H_2 in the subsurface of the Iberian Pyrite Belt is biologically produced (Sanz et al., 2021). This factor should be included in the analysis of other subsurface ecosystems to increase our understanding of the H cycle in the deep subsurface.

To some researchers H_2 is the most abundant source of energy in the deep subsurface which explains the continental dark biosphere's independence from the surface (Chapelle et al., 2002; Nealson et al., 2005; Brazelton et al., 2012; Lau et al., 2016). H_2, carbon dioxide (CO_2) and CH_4 have been reported in most of the drilling operations in which these gases have been measured, together with the identification of microorganisms able to use or produce them, which is a good back up for the suggested hypothesis (Pedersen, 2000; Moser et al., 2005; Itävaara et al., 2011; Wu et al., 2017; Amils et al., 2023). Nevertheless, other authors disagree with this view, claiming that in the deep subsurface, at least in part, there is a source of organic matter from the surface, such as petroleum deposits and sedimentary rocks (Fredrickson and Balkwill, 2006), in addition to the possibility of percolating water containing organic matter through pores and fractures of the rock. As mentioned earlier, the debate between defendants of each view is difficult to solve with the information available at the moment and the huge diversity of ecosystems studied. Obviously, this will not be a problem in the evaluation of habitability in other planetary systems.

Hydrogenotrophic methanogenesis was the first archaeal metabolic activity detected in the continental deep subsurface (Pedersen and Albinsson, 1992) and has been repeatedly identified in most drilling operations, directly, by detecting its metabolic product (CH_4) and/or the gases involved in its synthesis (H_2 and CO_2) or indirectly, by identifying the microorganisms responsible (Moser et al., 2005; Probst et al., 2014; Puente-Sánchez et al., 2014; Purkamo et al., 2015; Rempfert et al., 2017; Amils et al., 2023). Interestingly, in many drilling operations Cyanobacteria have been frequently identified (Onstott et al., 2003; Zhang et al., 2005; Bomberg et al., 2014; Purkamo et al., 2015; Ino et al., 2017; Rempfert et al., 2017; Puente-Sánchez et al., 2018; Amils et al., 2023), so today it is well established that members of this phyla, which obviously cannot make use of solar radiation as a source of energy, can develop in the deep subsurface using alternative sources of energy, such as H_2 (Puente-Sánchez et al., 2018).

In addition to H_2 and reduced organic matter in the continental deep subsurface, chemolithoautotrophic microorganisms can use CO_2 as carbon source and other energy sources such as reduced sulfur compounds (Amend and Teske, 2005; Gihring et al., 2006; Amend et al., 2011; Lau et al., 2016; Amils et al., 2023), iron (Sahl et al., 2008; Swanner et al., 2011; Shelobolina et al., 2012; Amils et al., 2023), nitrogen (Swanner et al., 2011; Nyyssönen et al., 2014; Lau et al., 2016; Amils et al., 2023), arsenic (Zhang et al., 2005; Sahl et al., 2008), manganese (Moser et al., 2005), or methane (Nyyssönen et al., 2012; Lau et al., 2016; Ino et al., 2017; Amils et al., 2023).

Methodologies currently used in the hard rock continental subsurface characterization

Several excellent reviews have been published covering the methodologies recommended for controlling the unavoidable contamination during drilling (Kieft, 2010; Wilkins et al., 2014). In this section we will concentrate on the methodologies that we consider fundamental for a full characterization of the operation of the hard rock continental deep subsurface. As mentioned, the deep subsurface is anoxic, thus it is extremely important to avoid any exposure of the drilled cores to atmospheric oxygen during the generation of the samples (Kieft, 2010). For this reason, after extraction, cores should be immediately placed in plastic bags and O_2 displaced with N_2. It is advisable to have access to an anaerobic glove box near the drilling site, so that powdered samples from the interior of the cores, using a rotary hammer drill, under sterile and temperature-controlled conditions can be generated and processed as soon as possible (Puente-Sánchez et al., 2018). This requirement will not be necessary for the characterization of Mars subsurface.

Most geomicrobiological subsurface information can be generated using different conventional analytical methodologies (for a recent description of most of these methodologies in hard rock deep subsurface drilling see Amils et al., 2023): Inductively Coupled Plasma Mass Spectrometry (ICP-MS) and Total Reflection X-Ray Fluorescence (TXRF) for elemental analysis of the solid substrate and the bioavailable elements; X-Ray Diffraction (XRD) for the mineral composition of the substrate; Scanning Electron Microscopy (SEM) for the visualization of mineral-microbe interactions; Confocal Raman Spectroscopy to evaluate mineral-microbe interaction (Escudero et al., 2020); Isotope Ratio Mass Spectrometry for identification of the involved metabolisms (Cabugao et al., 2022); ionic and gas chromatography for the analysis of substrates and products from different metabolisms; 16S rRNA gene cloning, 16S rRNA gene amplicon and metagenomic shotgun sequencing (Puente-Sánchez et al., 2018), immunological detection, FISH, Catalysed Reported Deposition Fluorescence *in situ* Hybridization (CARD-FISH), and lipid biomarker analysis (Li et al., 2020) for complementary evaluation of microbial diversity, each one with its own limitations; enrichment cultures, metatranscriptomics (Li et al., 2020), quantitative PCR (Jesser et al., 2015) metaproteomics (Bagnoud et al., 2016), single cell genomics (Becraft et al., 2021) and Time-of-flight Secondary Ionization Mass Spectrometry (Tof-SIMS) (Viang and Dong, 2012) for identification of specific metabolisms; isolation of microorganisms for phenotypic characterization; genome sequencing and gene annotation for the identification of potential metabolisms. The use of several complementary methodologies is strongly recommended to identify convergent results avoiding the bias introduced by each one (Amils et al., 2023).

Suitability of fluorescence *in situ* hybridization for hard rock deep subsurface characterization

Of the different methodologies used to identify the microbial diversity, the adaptation of rRNA-targeted fluorescence *in situ*

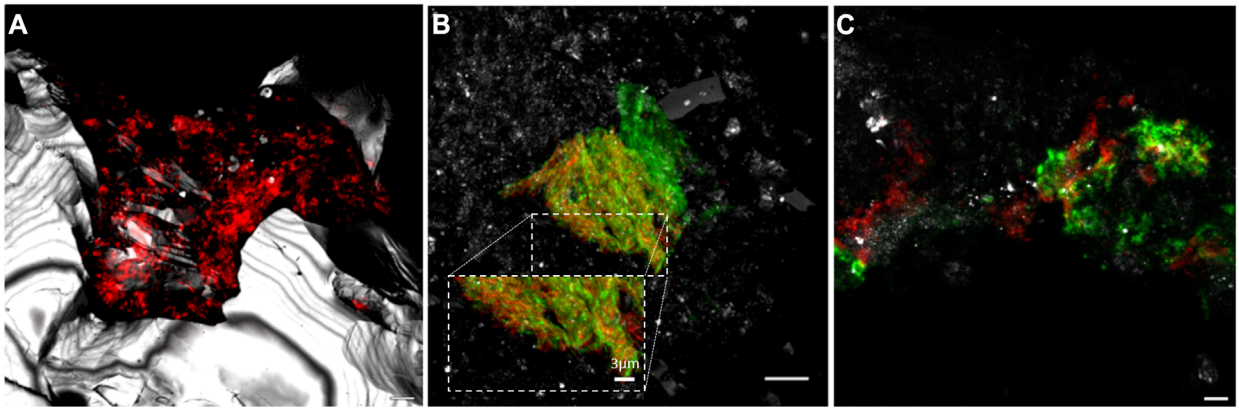


FIGURE 2
Microbial communities on mineral substrates of the Iberian Pyrite Belt subsurface visualized by Catalysed Reported Deposition Fluorescence *in situ* Hybridization (CARD-FISH) and Confocal Laser Scanning Microscopy (CLSM). **(A)** Bacteria domain at 139.4. **(B)** Interaction of *Acidovorax* (red) and *Sulfobacillus* (green) at 249.8 mbs. **(C)** *Acidiphillium spp.* biofilm at 228.4 mbs, in red, probe signal; in green, exopolysaccharides; in grey, reflection. Scale bars, 5 µm.

hybridization (rRNA-FISH) (Amand et al., 2005) to the analysis of microorganisms associated to semisolid and solid substrates was an important advance in the study of the hard rock deep subsurface (Hoshino et al., 2008). FISH technique allows the identification of microorganisms of interest, their quantification, and the analysis of their distribution in the sample (Figure 2). In fact, thanks to the possibility of performing multiple hybridizations on a single sample, this technique has been key to the identification of microbial associations and complementary metabolisms in both the oceanic (Schippers et al., 2005) and continental subsurface (Amils et al., 2023) An example is shown in Figure 2B, in which members of the genus *Acidovorax*, capable of oxidizing Fe(II) (Kappler et al., 2005), and the genus *Sulfobacillus*, which can couple the oxidation of reduced sulfur compounds to the reduction of Fe(III) (Justice et al., 2014), coexist together in the same microniche. Because FISH methodology requires fixation and denaturation of samples immediately after their extraction from within the cores, contamination is very unlikely. In addition, an important advantage over the rest of the methodologies is that it only requires a very small amount of sample, generating information on the presence of different microorganisms at microscopy size resolution, which cannot be obtained with the rest of the methodologies, as they require larger samples for their analysis (up to 100 g for DNA sequencing). Furthermore, FISH is a non-destructive methodology allowing the study of subsurface microorganisms in their natural environment: rock. Other used molecular biology methodologies such as "omics" yield significant bulk diversity results, but, unlike FISH, cannot provide information on the interconnection between the identified microorganisms and the mineral features of the complex matrix in which they inhabit.

Because, in general, the fluorescence probes for FISH are designed to hybridize with the ribosomal RNA, positive signals strongly suggest the presence of metabolically active microorganisms in the deep subsurface, a fundamental question in subsurface geomicrobiology (Schippers et al., 2005; Puente-Sánchez et al., 2018). The use of CARD-FISH allows amplification of the hybridization signal facilitating the distinction between real hybridizations and artifacts, such as the presence of autofluorescence minerals or the unspecific binding of the probe or dye to the diverse mineral content present in each sample (Escudero et al., 2018; Templeton and Caro, 2023). In addition, the combination of FISH with fluorescence lectin binding assay (FLBA) allowed the existence of biofilms made of exopolysaccharides in the oligotrophic deep subsurface to be demonstrated (Figure 2C), contrary to the generally accepted idea that, in these extreme oligotrophic conditions, microorganisms are unable to use their limited source of energy in the generation of these metabolically expensive structures (Escudero et al., 2018; Amils et al., 2023). If needed, microorganisms use their limited resources to produce biofilms, taking advantage of their useful properties such as protection against desiccation, diffusion control of metabolic products and substrates, facilitating the operation of complementary metabolisms and interconnecting microniches, among others (Flemming et al., 2016).

Sequence information is required for the design of specific fluorescence probes. An ideal situation would be to carry out the metagenomic analysis of the 16S rRNA genes present to facilitate the design of FISH probes. But in the absence of this information, a huge collection of published FISH probes of different specificity can facilitate the analysis (Amils et al., 2023). Due to its outstanding properties new FISH procedures have been developed targeting not only rRNA but also mRNA or single genes (Moraru et al., 2010).

The use of confocal laser scanning microscopy (CLSM) has been an important improvement in the detection of microorganisms occupying different focal planes in the mineral substrates. Correlative fluorescence and Raman microscopy (Raman-FISH) can be used to analyze the mineral-microorganism interactions, a fundamental relationship in the deep subsurface (Escudero et al., 2020). The introduction of super-resolution microscopy has been an important advance to overcome the limitations of using fluorescence methodologies for the analysis of complex subsurface samples (Moraru and Amann, 2012). The Nanoscale Secondary

Ion Mass Spectrometry (NanoSIMS) ability to measure stable isotopes and radioisotopes with appropriate half-lives has been used to image metabolically active microorganisms in complex communities such as those existing in the deep subsurface, and its coupling with Halogen *in situ* Hybridization (HISH) allow the simultaneous identification of microorganisms and the substrate uptake, providing important basic information on the operation of the deep subsurface (Musat et al., 2008).

Mars dark biosphere exploration

As mentioned and important limitation to estimate the subsurface habitability zone for rocky exoplanets is that deep subsurface life may not be able to modify the atmospheres of planets to be detectable remotely. This limitation does not exist in our Solar System, where exploration missions could be designed to collect information on the subsurface, or even better, to bring samples to Earth for analysis with more powerful and updated methodologies. This is the case for Mars, the planet that lost its surface water, and for which life at the surface is unlikely due to the extreme conditions such as lack of water, intense radiation, low temperatures, and strong oxidizing conditions (Margulis et al., 1979; Schofield et al., 1997; Rafkin et al., 2016; Martínez et al., 2017). The demonstration of Darwin´s prediction of life in hard rock deep subsurface of Earth changed this pessimistic point of view (Zhang et al., 2005; Fry et al., 2009; Breuker et al., 2011; Momper et al., 2017a; Dutta et al., 2018; Cockell et al., 2021) especially after the detection of liquid water in Mars subsurface (Orsei et al., 2018). There are an important number of studies suggesting the existence of life in the deep subsurface of Mars extrapolating results obtained on planet Earth (Boston et al., 1992; Jakosky and Shock, 1998; Cockell and Barlow, 2002; Michalski et al., 2018; Tarnas et al., 2018; Onstott et al., 2019; Sueoka et al., 2019; Stamenkovic et al., 2021; Sauterey et al., 2022). The demonstration of microbial life supporting the operation of the most important biogeochemical cycles (C, H, N, S, and Fe) at different depths in the hard rock deep subsurface of the Iberian Pyrite Belt (Amils et al., 2023), considered a good geological and mineralogical terrestrial Mars analogue (Fernández-Remolar et al., 2005), allows us to assert that the possible existence of past or even extant life in the subsurface of Mars must be considered (Price et al., 2018; 2022; Stamenkovic et al., 2019; Koike et al., 2020; Purkamo et al., 2020; Tarnas et al., 2021) and searched in future exploration missions. This will be the first space mission to explore the importance of the dark biosphere in the concept of planetary habitability.

Final considerations

The astrobiological interest of deep subsurface life is mainly concerned with broadening our concept of habitability which is fundamental to the search for life in the Universe. Our current concept of habitability does not take into consideration the possible existence of hard rock dark biospheres in rocky planets. Obviously, the presence of life in the hard rock continental deep subsurface has expanded our awareness of habitability significantly, although yet it is difficult to evaluate. McMahon and collaborators' introduction of the concept of a subsurface habitability zone was a pivotal advance in the field (McMahon et al., 2013). However, their calculations need more input based on an increase in our knowledge of the limits of life in the continental deep subsurface, which to date, are still only speculative. These include sources of energy available and the way in which basic elemental biogeochemical cycles, fundamental for successful dark biospheres, operate. An important limitation to our ability to estimate the subsurface habitability zone for an exoplanet is that currently we can only evaluate its mass, density, surface, and atmosphere as well as its distance from a star and the star's type, (Méndez et al., 2021). Future advances in the characterization of rocky exoplanets might allow solving this limitation.

We need to improve our still meager knowledge on the Earth igneous ocean crust and the continental dark biosphere to prepare the exploration of its possible existence on Mars or in the interior liquid water worlds in our Solar system, while we progress in the characterization of rocky exoplanets to incorporate them in the evaluation of their habitability, not only considering the possible existence of liquid water in its surface but the existence of conditions that could facilitate the habitability of its subsurface.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work has been financed by grants PID 2019-104812GB-I00 and TED 2021-129563B-I00 from the Spanish Ministerio de Ciencia e Innovación.

Acknowledgments

Authors want to thank R. Samalot for the English correction of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor AGF declared a past co-authorship with the author RA.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amand, A. L. S., Frank, D. N., De Groote, M. A., Basaraba, R. J., Orme, I. M., and Pace, N. R. (2005). Use of specific rRNA oligonucleotide probes for microscopic detection of *Mycobacterium tuberculosis* in cultures and tissue specimens. *J. Clin. Microbiol.* 43, 5369–5371. doi:10.1128/JCM.43.10.5369-5371.2005

Amend, J. P., McCollon, T. M., Hentscher, M., and Bach, W. (2011). Catabolic and anabolic energy for chemolithoautotrophs in deep-sea hydrothermal systems hosted in different rock types. *Geochim. Cosmochim. Acta* 75, 5736–5748. doi:10.1016/j.gca.2011.07.041

Amend, J. P., and Teske, A. (2005). Expanding frontiers in deep subsurface microbiology. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 219, 131–155. doi:10.1016/j.palaeo.2004.10.018

Amils, R., Escudero, C., Oggerin, M., Puente-Sánchez, F., Arce-Rodríguez, A., Fernández-Remolar, D., et al. (2023). Coupled C, H, N, S and Fe biogeochemical cycles operating in the continental deep subsurface of the Iberian Pyrite Belt. *Environ. Microbiol.* 25, 428–453. doi:10.1111/1462-2920.16291

Apps, J. A., and van de Kamp, P. C. (1993). Energy gases of abiogenic origin in the Earth's crust. *U. S. Geol. Surv. Prof. Pap.* 1570, 81–132.

Bagnoud, A., Chourey, K., Hettich, R. L., Bruijn, I., Andersson, A. F., Leupin, O. X., et al. (2016). Reconstructing a hydrogen-driven microbial metabolic network in Opalinus Clay rock. *Nat. Commun.* 7, 12770. doi:10.1038/ncomms12770

Bar-On, Y., Phillips, R., and Milo, R. (2018). The biomass distribution on Earth. *Proc. Nat. Acad. Sci. U. S. A.* 115, 6506–6511. doi:10.1073/pnas.1711842115

Basso, O., Lascourreges, J. F., Le Borgne, F., Le Goff, C., and Magot, M. (2009). Characterization by culture and molecular analysis of the microbial.diversity of a deep subsurface gas storage aquifer. *Res. Microbiol.* 160, 107–116. doi:10.1016/j.resmic.2008.10.010

Bastin, E. S., Greer, F. E., Merritt, C., and Moulton, G. (1926). The presence of sulphate reducing bacteria in oil field waters. *Science* 63, 21–24. doi:10.1126/science.63.1618.21

Becraft, E. D., Lau Vetter, M. C., Bezuidt, O. K., Brown, J. M., Labonté, J. M., Kauneckaite-Griguole, K., et al. (2021). Evolutionary stasis of a deep subsurface microbial lineage. *ISME J.* 15 (10), 2830–2842. doi:10.1038/s41396-021-00965-3

Bomberg, M., Nyyssönen, M., Nousiainen, A., Hultman, J., Paulin, L., Auvinen, P., et al. (2014). Evaluation of molecular techniques in characterization of deep terrestrial biosphere. *Open J. Ecol.* 4, 468–487. doi:10.4236/oje.2014.48040

Borgonie, G., García-Moyano, A., Litthauer, D., Bert, W., Bester, A., van Heerden, E., et al. (2011). Nematoda from the terrestrial deep subsurface of South Africa. *Nature* 474, 79–82. doi:10.1038/nature09974

Boston, J. P., Ivanov, M. V., and McKay, C. P. (1992). On the possibility of chemosynthetic ecosystems in subsurface habitats on Mars. *Icarus* 95, 300–308. doi:10.1016/0019-1035(92)90045-9

Brazelton, W. J., Nelson, B., and Schrenk, M. O. (2012). Metagenomic evidence for H_2 oxidation and H2 production by serpentinite-hosted subsurface microbial communities. *Front. Microbiol.* 2 (268), 268. doi:10.3389/fmicb.2011.00268

Breuker, A., Köweker, G., Blazejak, A., and Schippers, A. (2011). The deep biosphere in terrestrial sediments in the Chesapeake Bay area, Virginia, USA. *Front. Microbiol.* 2, 156. doi:10.3389/fmicb.2011.00156

Cabugao, K. G., Gushgari-Doyle, S., Chacon, S. S., Wu, X., Bhattacharyya, A., Bouskill, N., et al. (2022). Characterizing natural organic matter transformations by microbial communities in terrestrial subsurface ecosystems: A critical review of analytical techniques and challenges. *Front. Microbiol.* 13, 864895. doi:10.3389/fmicb.2022.864895

Chapelle, F. H., O'neill, K., Bradley, P. M., Methé, B. A., Ciufo, S. A., Knobel, L. L., et al. (2002). A hydrogen-based subsurface microbial community dominated by methanogens. *Nature* 415, 312–315. doi:10.1038/415312a

Chivian, D., Brodie, E. L., Alm, E. J., Culley, D. E., Dehal, P. S., DeSantis, T. Z., et al. (2008). Environmental genomics reveals a single-species ecosystem deep within Earth. *Science* 322, 275–278. doi:10.1126/science.1155495

Cockell, C. S. (2020). *Astrobiology: Understanding life in the universe*. 2nd edition, Hoboken: Wiley.

Cockell, C. S., and Barlow, N. G. (2002). Impact excavation and the search for subsurface life on Mars. *Icarus* 155, 340–349. doi:10.1006/icar.2001.6725

Cockell, C. S., Bush, T., Bryce, C., Direito, S., Fox-Powell, M., Harrison, J. P., et al. (2016). Habitability: A review. *Astrobiol* 16, 89–117. doi:10.1089/ast.2015.1295

Cockell, C. S., Schaefer, B., Wuchter, C., Coolen, M. J. L., Grice, K., Schnieders, L., et al. (2021). Shaping of the present-day deep biosphere at Chicxulub by the impact catastrophe that ended the cretaceous. *Front. Microbiol.* 12, 668240. doi:10.3389/fmicb,2021.668240

Cockell, C. S. (2014). Trajectories of martian habitability. *Astrobiol* 14, 182–203. doi:10.1089/ast.2013.1106

Cockell, C. S., Voytek, M. A., Gronstal, A. L., Finster, K., Kirshtein, J. D., Howard, K., et al. (2012). Impact disruption and recovery of the deep subsurface biosphere. *Astrobiol* 12, 231–246. doi:10.1089/ast.2011.0722

Colman, D., Poudel, S., Stamps, B. W., Boyd, E. S., and Spear, J. R. (2017). Microbial competition in porous environments can select against rapid biofilm growth. *Proc. Natl. Acad. Sci.* 114 (2), E161–E170. doi:10.1073/pnas.1525228113

Corliss, J. B., Dymond, J., Gordon, L. I., Edmond, J. M., von Herzen, R. P., Ballard, R. D., et al. (1979). Submarine thermal springs on the galapagos rift. *Science* 203, 1073–1083. doi:10.1126/science.203.4385.1073

Coskun, Ö. K., Vuillemin, A., Schubotz, F., Klein, F., Sichel, S. E., Eisenreich, W., et al. (2021). Quantifying the effects of hydrogen on carbon assimilation in a seafloor microbial community associated with ultramafic rocks. *ISME J* 16, 257–271. doi:10.1038/s41396-021-01066-x

Darwin, C. (1839). *Voyages of the adventure and Beagle, volume III–journal and remarks*. London: Henry Colburn, 1832–1836.

D'Hondt, S., Rutherford, S., and Spivack, A. J. (2002). Metabolic activity of subsurface life in deep-sea sediments. *Science* 295, 2067–2070. doi:10.1126/science.1064878

Dole, S. H. (1964). *Habitable planets for man*. New York: Blaisdell.

Dong, Y., Sanford, R. A., Locke, R. A., Cann, I. K., Mackie, R. I., and Fouke, B. W. (2014). Fe-oxide grain coatings support bacterial Fe-reducing metabolisms in 1.7-2.0 km-deep subsurface quartz arenite sandstone reservoirs of the Illinois Basin (USA). *Front. Microbiol.* 5 (511), 511. doi:10.3389/fmicb.2014.00511

Dutta, A., Dutta Gupta, S., Gupta, A., Sarkar, J., Roy, S., Mukherjee, A., et al. (2018). Exploration of deep terrestrial subsurface microbiome in Late Cretaceous Deccan traps and underlying Archean basement, India. *India Sci. Rep.* 8, 17459. doi:10.1038/s41598-018-35940-0

D´Hondt, S., Inagaki, F., Ferdelman, T., Jorgensen, B. B., Kato, K., Kemp, P., et al. (2007). Exploring subseafloor life with the integrated ocean drilling program. *Sci. Drill.* 5, 26–37. doi:10.5194/sd-5-26-2007

Edwards, K. J., Becker, K., and Colwell, F. (2012a). The deep, dark energy biosphere: Intraterrestrial life on Earth. *Annu.Rev.Earth Planet. Sci.* 40, 551–568. doi:10.1146/annurev-earth-042711-105500

Edwards, K. J., Bond, P. L., Gihring, T. M., and Banfield, J. F. (2000). An archaeal iron oxidizing extreme acidophile important in acid mine drainage. *Science* 287, 1796–1799. doi:10.1126/science.287.5459.1796

Edwards, K. J., Fisher, A. T., and Wheat, C. G. (2012b). The deep subsurface biosphere in igneous ocean crust: Frontier habitats for microbiological exploration. *Front. Microbiol.* 3, 8. doi:10.3389/fmicb.2012.00008

El-Naggar, M. Y., Wanger, G., Leung, K. M., Yuzvinsky, T. D., Southam, G., Yang, J., et al. (2010). Electrical transport along bacterial nanowires from Shewanella oneidensis MR-1. *Proc.Natl. Acad. Sci. U. S. A.* 107, 18127–18131. doi:10.1073/pnas.1004880107

Escudero, C., del Campo, A., Ares, J. R., Sánchez, C., Martínex, J. M., Gómez, F., et al. (2020). Visualizing microorganisms-mineral interactions in the iberian pyrite Belt subsurface: The *Acidovorax* case. *Front. Microbiol.* 11, 572104. doi:10.3389/fmicb.2020.572104

Escudero, C., Vera, M., Oggerin, M., and Amils, R. (2018). Active microbial biofilms in deep poor porous continental subsurface rocks. *Sci. Rep.* 8 (1), 1538. doi:10.1038/s41598-018-19903-z

Eydal, H. S., Jägevall, S., Hermansson, M., and Pedersen, K. (2009). Bacteriophage lytic to *Desulfovibrio aespoeensis* isolated from deep groundwater. *ISME J.* 3, 1139–1147. doi:10.1038/ismej.2009.66

Fernández-Remolar, D. C., Prieto-Ballesteros, O., Rodríguez, N., Gómez, F., Amils, R., Gómez- Elvira, J., et al. (2008). Underground habitats in the río tinto basin: A model for subsurface life habitats on Mars. *Astrobiol* 8, 1023–1047. doi:10.1089/ast.2006.0104

Fernández-Remolar, D., Morris, R. V., Gruener, J. E., Amils, R., and Knoll, A. (2005). The Río Tinto Basin, Spain: Mineralogy, sedimentary geobiology, and implications for interpretation of outcrop rocks at Meridiani Planum, Mars. *Earth Planet. Sci. Lett.* 240, 149–167. doi:10.1016/j.epsl.2005.09.043

Flemming, H.-C., Wingender, T., Szewzyk, P., Steinberg, P., Rice, S. A., and Kjelleberg, S. (2016). Biofilms: An emergent form of bacterial life. *Nat. Rev. Microbiol.* 14, 563–575. doi:10.1038/nrmicro.2016.94

Fredrickson, J. K., and Balkwill, D. L. (2006). Geomicrobial processes and biodiversity in the deep terrestrial subsurface. *Geomicrobiol. J.* 23, 345–356. doi:10.1080/01490450600875571

Fredrickson, J. K., McKinley, J. P., Bjornstad, B. N., Long, P. E., Ringelberg, D. B., White, D. C., et al. (1997). Pore-size constraints on the activity and survival of subsurface bacteria in a late cretaceous shale-sandstone sequence, northwestern New Mexico. *Geomicrobiol. J.* 14, 183–202. doi:10.1080/01490459709378043

Fry, N. K., Fredrickson, J. K., Fishbain, S., Wagner, M., and Stahl, D. A. (2009). Population structure of microbial communities associated with two deep, anaerobic, alkaline aquifers. *Appl. Environ. Microbiol.* 63, 1498–1504. doi:10.1128/AEM.63.4.1498-1504.1997

Gihring, T., Moser, D., Lin, L. H., Davidson, M., Onstott, T., Morgan, L., et al. (2006). The distribution of microbial taxa in the subsurface water of the Kalahari Shield, South Africa. *Geomicrobiol. J.* 23, 415–430. doi:10.1080/01490450600875696

Gold, T. (1992). The deep, hot biosphere. *Proc. Natl. Acad. Sci.* 89, 6045–6049. doi:10.1073/pnas.89.13.6045

Gronstal, A. L., Voytek, M. A., Kirshtein, J. D., Nicole, M., Lowit, M. D., and Cockell, C. S. (2009). Contamination assessment in microbiological sampling of the Eyreville core, Chesapeake Bay impact structure. *Geol. Soc. Am. Spec. Pap.* 458, 951–964. doi:10.1130/2009.2458(41)

Hart, M. H. (1979). Habitable zones about main sequence stars. *Icarus* 37, 351–357. doi:10.1016/0019-1035(79)90141-6

Hoehler, T. M. (2004). Biological energy requirements as quantitative boundary conditions for life in the subsurface. *Geobiol* 2, 205–215. doi:10.1111/j.1472-4677.2004.00033.x

Hoshino, T., Yilmaz, L. S., Noguera, D. R., Daims, H., and Wagner, M. (2008). Quantification of target molecules needed to detect microorganisms by fluorescence *in situ* hybridization (FISH) and catalyzed reporter deposition-FISH. *Appl. Environ. Microbiol.* 74, 5068–5077. doi:10.1128/AEM.00208-08

Huang, S. S. (1959). Occurrence of life in the universe. *Am. Sci.* 47, 397–402.

Inagaki, F., Hinrichs, K.-U., Kubo, Y., Bowles, M. W., Heuer, V. B., Hong, W.-L., et al. (2015). Exploring deep microbial life in coal-bearing sediment down to ~2.5 km below the ocean floor. *Science* 349, 420–424. doi:10.1126/science.aaa6882

Ino, K., Hernsdorf, A. W., Konno, U., Kouduka, M., Yanagawa, K., Kato, S., et al. (2017). Ecological and genomic profiling of anaerobic methane-oxidizing archaea in a deep granitic environment. *ISME J.* 12, 31–47. doi:10.1038/ismej.2017.140

Ino, K., Konno, U., Kouduka, M., Hirota, A., Togo, Y. S., Fukuda, A., et al. (2016). Deep microbial life in high-quality granitic groundwater from geochemically and geographically distinct underground boreholes. *Environ. Microbiol. Rep.* 8, 285–294. doi:10.1111/1758-2229.12379

Itävaara, M., Nyyssönen, M., Kapanen, A., Nousiainen, A., Ahonen, L., and Kukkonen, I. (2011). Characterization of bacterial diversity to a depth of 1500 m in the Outokumpu deep borehole, Fennoscandian Shield. *FEMS Microbio.l Ecol.* 77, 295–309. doi:10.1111/j.1574-6941.2011.01111.x

Ivarson, M., Bengtson, S., Drake, H., and Francis, W. (2018). Fungi in deep subsurface environments. *Adv. Appl. Microbiol.* 102, 83–116. doi:10.1016/bs.aambs.2017.11.001

Jakosky, B. M., and Shock, E. L. (1998). The biological potential of Mars, the early Earth, and Europa. *J. Geophys. Res.* 103, 19359–19364. doi:10.1029/98je01892

Jesser, K. J., Fullerton, H., Hager, K. W., and Moyer, C. L. (2015). Quantitative PCR analysis of functional genes in iron-rich microbial mats at an active hydrothermal vent system (Lō'ihi Seamount, Hawai'i). *Appl. Environ. Microbiol.* 81 (9), 2976–2984. doi:10.1128/AEM.03608-14

Johnson, S. S., Hebsgaard, M. B., Christensen, T. R., Mastepanov, M., Nielsen, R., Munch, K., et al. (2007) Ancient bacteria show evidence of DNA repair. *Proc. Natl. Acad. Sci. U. S. A.* 4, 104(36):14401–14405. doi:10.1073/pnas.0706787104

Jones, A. A., and Bennett, P. C. (2017). Mineral ecology: surface specific colonization and geochemical drivers of biofilm accumulation, composition, and phylogeny. *Front. Microbiol.* 8, 491. doi:10.3389/fmicb.2017.00491

Jørgensen, B. B. (2011). Deep subseafloor microbial cells on physiological standby. *Proc. Natl. Acad. Sci.* 108 (45), 18193–18194. doi:10.1073/pnas.1115421108

Justice, N. B., Norman, A., Brown, C. T., Singh, A., Thomas, B. C., and Banfield, J. F. (2014). Comparison of environmental and isolate Sulfobacillus genomes reveals diverse carbon, sulfur, nitrogen, and hydrogen metabolisms. *BMC genomics* 15, 1107. doi:10.1186/1471-2164-15-1107

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., and D´Hont, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc. Nat. Acad. Sci. U. S. A.* 109, 16213–16216. doi:10.1073/pnas.1203849109

Kappler, A., Schink, B., and Newman, D. K. (2005). Fe(III) mineral formation and cell encrustation by the nitrate-dependent Fe(II)-oxidizer strain BoFeN1. *Geobiol* 3, 235–245. doi:10.1111/j.1472-4669.2006.00056.x

Kasting, J. F., and Catling, D. (2003). Evolution of a Habitable Planet. *Ann. Rev. Astron. Astrophys.* 41, 429–463. doi:10.1146/annurev.astro.41.071601.170049

Kasting, J. F., Whitmire, D. P., and Reynolds, R. T. (1993). Habitable zones around main sequence stars. *Icarus* 101, 108–128. doi:10.1006/icar.1993.1010

Kieft, T. L. (2016). "Microbiology of the deep continental biosphere. *Their world: A diversity of microbial environments.* 1, 225–249. doi:10.1007/978-3-319-28071-4_6

Kieft, T. (2010). "Sampling the deep sub-surface using drilling and coring techniques," in *Handbook of hydrocarbon and lipid microbiology* (Springer), 3427–3441.

Klein, F., Tarnas, J. D., and Bach, W. (2020). Abiotic sources of molecular hydrogen on Earth. *Elements* 16 (1), 19–24. doi:10.2138/gselements.16.1.19

Koike, M., Nakada, R., Kajitani, I., Usui, T., Tamenori, Y., Sugahara, H., et al. (2020). *In-situ* preservation of nitrogen-bearing organics in Noachian Martian carbonates. *Nat. Comm.* 11 (1), 1988. doi:10.1038/s41467-020-15931-4

Kyle, J. E., Eydal, H. S., Ferris, F. G., and Pedersen, K. (2008). Viruses in granitic groundwater from 69 to 450 m depth of the Äspö hard rock laboratory, Sweden. *ISME J.* 2, 571–574. doi:10.1038/ismej.2008.18

Labonté, J. M., Field, E. K., Lau, M., Chivian, D., Van Heerden, E., Wommack, K. E., et al. (2015). Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. *Front. Microbio,l* 6, 349. doi:10.3389/fmicb.2015.00349

Lammer, H., Bredehoft, J. H., Coustenis, A., Khodachenko, M. L., Kaltenegger, L., Grasset, O., et al. (2009). What makes a planet habitable? *Astron. Astrophys. Rev.* 17, 181–249. doi:10.1007/s00159-009-0019-z

Lau, M. C., Cameron, C., Magnabosco, C., Brown, C. T., Schilkey, F., Grim, S., et al. (2014). Phylogeny and phylogeography of functional genes shared among seven terrestrial subsurface metagenomes reveal N-cycling and microbial evolutionary relationships. *Front. Microbiol.* 5, 531. doi:10.3389/fmicb.2014.00531

Lau, M. C., Kieft, T. L., Kuloyo, O., Linage-Alvarez, B., Van Heerden, E., Lindsay, M. R., et al. (2016). An oligotrophic deep-subsurface community dependent on syntrophy is dominated by sulfur-driven autotrophic denitrifiers. *Proc. Nat. Acad. Sci.* 113, E7927–E7936. doi:10.1073/pnas.1612244113

Leandro, T., Rodrígez, N., Sanz, J. L., da Costa, M. S., and Amils, R. (2018). Study of methanogenic enrichment cultures of rock cores from the deep subsurface of the Iberian Pyritic Belt. *Heliyon* 4, e00605. doi:10.1016/j.heliyon.2018.e00605

Leong, J. A. M., Howells, A. E., Robinson, K. J., Cox, A., Debes, R. V., Fecteau, K., et al. (2021). Theoretical prediction versus environmental observations on serpentinization fluids: lessons from the Smail Ophiolite in Oman. *J. Geophys. Res. Solid Earth* 126, e2020JBO20756. doi:10.1029/2020JBO20756

Li, J., Mara, P., Schubotz, F., Sylvan, J. B., Burgaud, G., Klein, F., et al. (2020). Recycling and metabolic flexibility dictate life in the lower oceanic crust. *Nature* 579, 250–255. doi:10.1038/s41586-020-2075-5

Lin, L. H., Slater, G. F., Sherwood Lollar, B., Lacrampe-Couloume, G., and Onstott, T. C. (2004). The yield and isotopic composition of radiolytic H2, a potential energy source for the deep subsurface biosphere. *Geochim. Cosmochim. Acta* 69 (4), 893–903. doi:10.1016/j.gca.2004.07.032

Lin, L. H., Wang, P. L., Rumble, D., Lippmann-Pipke, J., Boice, E., Pratt, L. M., et al. (2006). Long-term sustainability of a high-energy, low-diversity crustal biome. *Science* 314, 479–482. doi:10.1126/science.1127376

Lingam, M., and Loeb, A. (2021). *Life in the Cosmos. From biosignatures to thechnosignatures*. Boston: Harvard University Press.

Lingam, M., and Loeb, A. (2020). Potential for liquid water biochemistry deep under the surfaces of the Moon, Mars and beyond. *Astrophys. J. Lett-* 90, L11. doi:10.3847/2041-8213/abb608

Lipman, C. B. (1931). Living microorganisms in ancient rocks. *J. Bacteriol.* 22, 183–198. doi:10.1128/jb.22.3.183-198.1931

Magnabosco, C., Lin, L. H., Dong, H., Bomberg, M., Ghiorse, W., Stan-Lotter, H., et al. (2018). The biomass and biodiversity of the continental subsurface. *Nat. Geosci.* 11, 707–717. doi:10.1038/s41561-018-0221-6

Magnabosco, C., Ryan, K., Lau, M. C., Kuloyo, O., Lollar, B. S., Kieft, T. L., et al. (2016). A metagenomic window into carbon metabolism at 3 km depth in Precambrian continental crust. *ISME J.* 10, 730–741. doi:10.1038/ismej.2015.150

Magnabosco, C., Tekere, M., Lau, M. C., Linage, B., Kuloyo, O., Erasmus, M., et al. (2014). Comparisons of the composition and biogeographic distribution of the bacterial communities occupying South African thermal springs with those inhabiting deep subsurface fracture water. *Front. Microbiol.* 5, 679. doi:10.3389/fmicb.2014.00679

Margulis, L., Mazur, P., Barghoom, E. S., Halvorson, H. O., Jukes, T. H. J., and Kaplan, I. R. (1979). TheViking Mission: implications for life in Vallis Marineris area. *Science* 305, 78–81. doi:10.1126/science.1097549

Martínez, G. M., Newman, C. N., deVicente-Retortillo, A., Fischer, E., Rennó, N. O., Richardson, M. I., et al. (2017). Erratum to: The Modern Near-Surface Martian Climate: A Review of *in-situ* Meteorological Data from Viking to Curiosity. *Space Sci. Rev.* 212, 339–340. doi:10.1007/s11214-017-0368-2

Maunder, E. W. (1913). *Are the planets inhabited*. London: Harper and Brothers.

McMahon, S., O´Malley-James, J., and Parnell, J. (2013). Circumstellar habitable zones for Deep terrestrial biospheres. *Planet. Space Sci.* 85, 312–318. doi:10.1016/j.pss.2013.07.002

McMahon, S., and Parnell, J. (2014). Weighing the deep continental biosphere. *FEMS Microbiol. Ecol.* 87, 113–120. doi:10.1111/1574-6941.12196

Méndez, A., Rivera-Valentín, E. G., Schulze-Makuch, D., Filberto, J., Ramírez, R. M., Wood, T. E., et al. (2021). Habitability models for Astrobiology. *Astrobiol* 21 (8), 1017–1027. doi:10.1089/ast.2020.2342

Merino, N., Aronson, H. S., Bojanova, D. P., Feyhl-Busca, J., Wong, M. L., Zhang, S., et al. (2019). Living at the extremes: Extremophiles and the limits of life in a planetary context. *Front. Microbiol.* 10, 780. doi:10.3389/fmicb.2019.00780

Michalski, J. R., Onstott, T. C., Mojzsis, S. J., Mustard, J., Chan, Q. H. S., Niles, P. B., et al. (2018). The Martian subsurface as a potential window into the origin of life. *Nat. Geosc.* 11, 21–26. doi:10.1038/s41561-017-0015-2

Momper, L., Jungbluth, S. P., Lee, M. D., and Amend, J. P. (2017a). Energy and carbon metabolisms in a deep terrestrial subsurface fluid microbial community. *ISME J.* 11, 2319–2333. doi:10.1038/ismej.2017.94

Momper, L., Reese, B. K., Zinke, L., Wanger, G., Osburn, M. R., Moser, D., et al. (2017b). Major phylum-level differences between pore fluid and host rock bacterial communities in the terrestrial deep subsurface. *Environ. Microbiol. Rep.* 9, 501–511. doi:10.1111/1758-2229.12563

Moraru, C., Amann, R., Cremer, M., and Schermelleh, L. (2012). Crystal ball: fluorescence *in situ* hybridization in the age of super-resolution microscopy. *Syst. Appl. Microbiol.* 35, 43–64. doi:10.1007/978-1-62703-137-0_4

Moraru, C., Lam, P., Fuchs, B. M., Kuypers, M. M., and Amann, R. (2010). GeneFISH - an *in situ* technique for linking gene presence and cell identity in environmental microorganisms. *Environ. Microbiol.* 12, 3057–3073. doi:10.1111/j.1462-2920.2010.02281.x

Morita, R. Y. (1999). Is H2 the universal energy source for long-term survival? *Microb. Ecol.* 38, 307–320. doi:10.1007/s002489901002

Morita, R., and Zobell, C. (1955). Occurrence of bacteria in pelagic sediments collected during the Mid- Pacific Expedition. *Deep-Sea Res.* 3, 66–73. doi:10.1016/0146-6313(55)90036-8

Moser, D. P., Gihring, T. M., Brockman, F. J., Fredrickson, J. K., Balkwill, D. L., Dollhopf, M. E., et al. (2005). *Desulfotomaculum* and *Methanobacterium* spp. Dominate a 4- to 5-Kilometer-Deep Fault. *Appl. Environ. Microbiol.* 71, 8773–8783. doi:10.1128/AEM.71.12.8773-8783.2005

Murakami, Y., Fujita, Y., Naganuma, T., and Iwatsuki, T. (2002). Abundance and viability of the groundwater microbial communities from a borehole in the Tono uranium deposit area, central Japan. *Microb. Environ.* 17, 63–74. doi:10.1264/jsme2.2002.63

Musat, N., Halm, H., Winterholler, B., Hoppe, P., Peduzi, S., Hillion, F., et al. (2008). A single-cell view on the ecophysiology of anaerobic phototrophic bacteria. *Proc. Natl. Acad. Sci.* 105, 17861–17866. doi:10.1073/pnas0809329105

Nealson, K. H., Inagaki, F., and Takai, K. (2005). Hydrogen-driven subsurface lithoautotrophic microbial ecosystems (SLiMEs): do they exist and why should we care? *Trends Microbiol.* 13, 405–410. doi:10.1016/j.tim.2005.07.010

Nisbet, E., Zahnle, K., Gerasimov, M. V., Helbert, J., Jaumann, R., Hofman, B. A., et al. (2007). Creating habitable zones. At all scales, from planets to mud-micro-habitats, on Earth and on Mars. *Space aci. Rev.* 129, 79–121. doi:10.1007/978-0-387-74288-5_4

Nuppunen-Puputti, M., Kietäväinen, R., Raulio, M., Soro, A., Purkamo, L., Kukkonen, I., et al. (2022). Epilithic microbial community functionality in deep oligotrophic continental bedrock. *Front. Microbiol.* 13, 826048. doi:10.3389/fmicb.2022.826048

Nyyssönen, M., Bomberg, M., Kapanen, A., Nousiainen, A., Pitkänen, P., and Itävaara, M. (2012). Methanogenic and sulphate-reducing microbial communities in deep groundwater of crystalline rock fractures in Olkiluoto, Finland. *Geomicrobiol. J.* 29, 863–878. doi:10.1080/01490451.2011.635759

Nyyssönen, M., Hultman, J., Ahonen, L., Kukkonen, I., Paulin, L., Laine, P., et al. (2014). Taxonomically and functionally diverse microbial communities in deep crystalline rocks of the Fennoscandian shield. *ISME J.* 8, 126–138. doi:10.1038/ismej.2013.125

Onstott, T. C., Ehlmann, B. L., Sapers, H., Coleman, M., Ivarsonn, M., Marlow, J. J., et al. (2019). Paleo-rock-hosted life on Earth and the search on Mars: a review and strategy for exploration. *Astrobiol* 19 (10), 1230–1262. doi:10.1089/ast.2018.1960

Onstott, T., Magnabosco, C., Aubrey, A. D., Burton, A. S., Dworkin, J. P., Elsila, J. E., et al. (2014). Does aspartic acid racemization constrain the depth limit of the subsurface biosphere? *Geobiology* 12 (1), 1–19. doi:10.1111/gbi.12069

Onstott, T., McGown, D. J., Bakemans, C., Ruskeeniemi, T., Ahonen, L., Telling, J., et al. (2009). Microbial Communities in Subpermafrost Saline Fracture Water at the Lupin Au Mine, Nunavut, Canada. *Microb. Ecol.* 58, 786–807. doi:10.1007/s00248-009-9553-5

Onstott, T., Moser, D. P., Pfiffner, S. M., Fredrickson, J. K., Brockman, F. J., Phelps, T., et al. (2003). Indigenous and contaminant microbes in ultradeep mines. *Environ. Microbiol.* 5, 1168–1191. doi:10.1046/j.1462-2920.2003.00512.x

Orcutt, B. N., Sylvan, J. B., Knab, N. J., and Edwards, K. J. (2011). Microbial ecology of the dark ocean above, at, and below the seafloor. *Microbiol. Mol. Biol. Rev.* 75, 361–422. doi:10.1128/MMBR.00039-10

Orsei, R., Lauro, S. E., Pettinelli, E., Cicchetti, A., Coradini, M., Di Paolo, U., et al. (2018). Radar evidence of subglacial liquid water on Mars. *Science* 361, 490–493. doi:10.1126/science.aar7268

Orsi, W. D., Magritsch, T., Vargas, S., Coskun, O. K., Vuillemin, A., Höhna, S., et al. (2021). Genome Evolution in Bacteria Isolated from Million-Year-Old Subseafloor Sediment. *mBio* 12, e0115021–21. doi:10.1128/mBio01150-21

Osburn, M. R., Larowe, D. E., Momper, L. M., and Amend, J. P. (2014). Chemolithotrophy in the continental deep subsurface: Sanford Underground Research Facility (SURF), USA. *Microbiol* 5, 610. doi:10.3389/fmicb.2014.00610

Pedersen, K., and Albinsson, Y. (1992). Possible Effects of Bacteria on Trace Element Migration in Crystalline Bed-Rock. *Radiochim. Acta* 58 (59), 365–370. doi:10.1524/ract.1992.5859.2.365

Pedersen, K. (2000). Exploration of deep intraterrestrial microbial life: current perspectives. *FEMS Microbiol. Lett.* 185, 9–16. doi:10.1111/j.1574-6968.2000.tb09033.x

Pedersen, K. (1997). Microbial life in deep granitic rock. *FEMS Microbiol. Rev.* 20, 399–414. doi:10.1111/j.1574-6976.1997.tb00325.x

Pedersen, K. (1999). Subterranean microorganisms and radioactive waste disposal in Sweden. *Eng. Geol.* 52, 163–176. doi:10.1016/s0013-7952(99)00004-6

Phelps, T., Murphy, E., Pfiffner, S., and White, D. (1994). Comparison between geochemical and biological estimates of subsurface microbial activities. *Microb. Ecol.* 28, 335–349. doi:10.1007/BF00662027

Price, A., Macey, M. C., Pearson, V. K., Schwenzer, S. P., Ramkissoon, N. K., and Olsson-Francis, S. P. (2022). Oligotrophic growth of nitrate-dependent Fe^{2+}-oxidizing microorganisms under simulated early martian conditions. *Front. Microbiol.* 13, 800219. doi:10.3389/fmicb.2022.800219

Price, A., Pearson, V. K., Schwenzer, S. P., Miot, J., and Olsson-Francis, K. (2018). Nitrate-dependent iron oxidation: a potential Mars metabolism. *Front. Microbiol.* 9, 513. doi:10.3389/fmicb.2018.00513

Probst, A. J., Birarda, G., Holman, H. Y. N., DeSantis, T. Z., Wanner, G., Andersen, G. L., et al. (2014). Coupling Genetic and Chemical Microbiome Profiling Reveals Heterogeneity of Archaeome and Bacteriome in Subsurface Biofilms That Are Dominated by the Same Archaeal Species. *PLoS One* 9 (6), e99801. doi:10.1371/journal.pone.0099801

Puente-Sánchez, F., Arce-Rodríguez, A., Oggerin, M., García-Villadangos, M., Moreno-Paz, M., Blanco, Y., et al. (2018). Viable cyanobacteria in the deep continental subsurface. *Proc. Natl. Acad. Sci.* 115, 10702–10707. doi:10.1073/pnas.1808176115

Puente-Sánchez, F., Moreno-Paz, M., Rivas, L., Cruz-Gil, P., García-Villadangos, M., Gómez, M., et al. (2014). Deep subsurface sulfate reduction and methanogenesis in the Iberian Pyrite Belt revealed through geochemistry and molecular biomarkers. *Geobiol* 12, 34–47. doi:10.1111/gbi.12065

Purkamo, L., Bomberg, M., Nyyssönen, M., Kukkonen, I., Ahonen, L., and Itävaara, M. (2015). Heterotrophic communities supplied by ancient organic carbon predominate in deep Fennoscandian bedrock fluids. *Microb. Eco.l* 69, 319–332. doi:10.1007/s00248-014-0490-6

Purkamo, L., Bomberg, M., Nyyssönen, M., Kukkonen, I., Ahonen, L., Kietäväinen, R., et al. (2013). Dissecting the deep biosphere: retrieving authentic microbial communities from packer-isolated deep crystalline bedrock fracture zones. *FEMS Microbiol. Ecol.* 85, 324–337. doi:10.1111/15746941.12126

Purkamo, L., Kietäväinen, R., Miettinen, H., Sohlberg, E., Kukkonen, I., Itävaara, M., et al. (2018). Diversity and functionality of archaeal, bacterial and fungal communities in deep Archaean bedrock groundwater. *FEMS Microbiol. Ecol.* 94, 8. doi:10.1093/femsec/fiy116

Purkamo, L., Kietäväinen, R., Nuppumen-Puputto, M., Bomberg, M., and Cousins, C. (2020). Ultradeep Microbial Communities at 4.4 km within Crystalline Bedrock: Implications for Habitability in a Planetary Context. *Life* 10, 2. doi:10.3390/life10010002

Quemener, M., Mara, P., Schubotz, F., Beaudoin, D., Li, W., Pachiadaki, M., et al. (2020). Meta-omics highlights the diversity, activity and adaptations of fungi in deep oceanic crust. *Environ. Microbiol.* 22 (9), 3950–3967. doi:10.1111/1462-2920.15181

Rafkin, S. C. R., Pla-García, J., Kahre, M. A., Gómez-Elvira, J., Hamilton, V. E., Marín, M., et al. (2016). The meteorology of Gale Crater as determined from Rover Environmental Monitoring Station observations and numerical modeling. Part II: Interpretation. *Icarus* 280, 114–138. doi:10.1016/j.icarus.2016.01.031

Rahlf, J., Turzynski, V., Esser, S. P., Monsees, I., Bornemann, T. L. V., Figueroa-Gonzalez, P. A., et al. (2021). Lytic archaeal viruses infect abundant primary producers in Earth´s crust. *Nat. Commun.* 12, 4642. doi:10.1038/s41467-021-24803-4

Rajala, P., Bomberg, M., Kietäväinen, R., Kukkonen, I., Ahonen, L., Nyyssönen, M., et al. (2015). Rapid reactivation of deep subsurface microbes in the presence of C-1 compounds. *Microorg* 3, 17–33. doi:10.3390/microorganisms3010017

Reese, B. K., Sobol, M. S., Bowles, M. W., and Hinrichs, K. U. (2021). Redefining the subsurface biosphere: characterization of fungi isolated from energy-limited marine deep subsurface sediment. *Front. Fungal Biol.* 2, 727543. doi:10.3389/ffunb.2021.727543

Rempfert, K. R., Miller, H. M., Bompard, N., Nothaft, D., Matter, J. M., Kelemen, et al. (2017). Geological and geochemical controls on subsurface microbial life in the Samail Ophiolite, Oman. *Front. Microbiol.* 8, 56. doi:10.3389/fmicb.2017.00056

Rogers, J., Bennett, P., and Choi, W. (1998). Feldspars as a source of nutrients for microorganisms. *Am. Mineral.* 83, 1532–1540. doi:10.2138/am-1998-11-1241

Ruff, S. E., Humez, P., Hrabe de Angelis, I., Diao, M., Nightingale, M., Chao, S., et al. (2023) Hydrogen and dark oxygen drive microbial productivity in diverse groundwater ecosystems. *Nat. Commun.* 14(1), 3194. doi:10.1038/s41467-023-38523-4

Sahl, J. W., Schmidt, R., Swanner, E. D., Mandernack, K. W., Templeton, A. S., Kieft, T. L., et al. (2008). Subsurface microbial diversity in deep-granitic-fracture water in Colorado. *Appl. Environ. Microbiol.* 74, 143–152. doi:10.1128/AEM.01133-07

Sanz, J. L., Rodríguez, N., Escudero, C., and CarrizoAmils, D. R. (2021). Biological production of H_2, CH_4 and CO_2 in the Deep subsurface of the Iberian Pyrite Belt. *Environ. Microbiol.* 23 (7), 3913–3922. doi:10.1111/1462-2920.15561

Sauterey, B., Charnay, B., Affholder, A., Mazevet, S., and Ferriére, R. (2022). Early Mars habitability and global cooling by H2-based methanogens. *Nat. Astron.* 6, 1263–1271. doi:10.1038/s41550-022-01786-w

Schippers, A., Neretin, L. N., Kallmeyer, J., Ferdelman, T. G., Cragg, B. A., Parkes, J., et al. (2005). Prokaryotic cells of the deep sub-seafloor biosphere identified as living bacteria. *Nature* 433, 861–864. doi:10.1038/nature03302

Schofield, J. T., Barnes, J. R., Crisp, D., Haberle, R. M., Larsen, S., Magalhães, J. A., et al. (1997). The Mars Pathfinder Atmospheric Structure Investigation/Meteorology (ASI/MET) Experiment. *Science* 278, 1752–1758. doi:10.1126/science.278.5344.1752

Shelobolina, E., Xu, H., Konishi, H., Kukkadapu, R., Wu, T., Blöthe, M., et al. (2012). Microbial lithotrophic oxidation of structural Fe (II) in biotite. *Appl. Environ. Microbiol.* 78, 5746–5752. doi:10.1128/AEM.01034-12

Sherwood Lollar, G. S., Warr, O., Telling, J., Osburn, M. R., and Sherwood Lollar, B. (2019). 'Follow the Water': Hydrogeochemical Constraints on Microbial Investigations 2.4 km Below Surface at the Kidd Creek Deep Fluid and Deep Life Observatory. *Hydrogeochem. constrains Microb. investigations 2.4 km bellow Surf. A. T. Kidd Creek Deep Fluid Deep Life Observatory. Geomicrobiol. J.* 36, 859–872. doi:10.1080/01490451.2019.1641770

Sherwood-Lollar, B., Voglesonger, K., Lin, L. H., Lacrampe-Couloume, G., Abrajano, T. A., Onstott, T. C., et al. (2007). Hydrogeologic Controls on Episodic H_2 Release from Precambrian Fractured Rocks—Energy for Deep Subsurface Life on Earth and Mars. *Astrobiol* 7 (6), 971–986. doi:10.1089/ast.2006.0096

Shock, E. L. (2009). Minerals as energy sources for microorganisms. *Econ. Geol.* 104, 1235–1248. doi:10.2113/gsecongeo.104.8.1235

Soares, A., Edwards, A., An, D., Bagnoud, A., Bradley, J., Barnhart, E., et al. (2023). A global perspective on bacterial diversity in the terrestrial deep subsurface. *Microbiol* 169, 001172. doi:10.1099/mic.0.001172

Sohlberg, E., Bomberg, M., Miettinen, H., Nyyssönen, M., Salavirta, H., Vikman, M., et al. (2015). Revealing the unexplored fungal communities in deep groundwater of crystalline bedrock fracture zones in Olkiluoto, Finland. *Front. Microbiol.* 6, 573. doi:10.3389/fmicb.2015.00573

Stamenkovic, V., Beegle, L. W., Zacny, K., Arumugam, D. D., Baglioni, P., Barba, W., et al. (2019). The next frontier for planetary and human exploration. *Nat. Astron.* 3, 116–120. doi:10.1038/s41550-018-0676-9

Stamenkovic, V., Lynch, K., Boston, P., Tarnas, J., Edwards, C. D., Sherwood-Lollar, B., et al. (2021). Deep Trek: Science of Subsurface Habitability and Life on Mars. *Bull. Am. Astronomical Soc.* 53 (4). doi:10.3847/25c2cfeb.dc18f731

Steen, A. D., Jorgensen, B. B., and Lomstein, B. (2013). Abiotic racemization kinetics of amino acids in marine sediments. *PLoS ONE* 8 (8), e71648. doi:10.1371/journal.pone.0071648

Stevens, T. O., and McKinley, J. P. (1995). Lithoautotrophic microbial ecosystems in deep basalt aquifers. *Science* 270, 450–455. doi:10.1126/science.270.5235.450

Sueoka, Y., Yamashita, S., Kouduka, M., and Suzuki, Y. (2019). Deep microbial colonization in saponite-bearing fractures in aged basaltic crust: implications for subsurface life on Mars. *Front. Microbiol.* 10, 2793. doi:10.3389/fmicb.2019.02793

Suzuki, S., Si, I., Wu, A., Cheung, A., Tenney, A., Wanger, G., et al. (2013). Microbial diversity in The Cedars, an ultrabasic, ultrareducing, and low salinity serpentinizing ecosystem. *Proc. Natl. Acad. Sci.* 110, 15336–15341. doi:10.1073/pnas.1302426110

Swanner, E. D., Nell, R. M., and Templeton, A. S. (2011). Ralstonia species mediate Fe-oxidation in circumneutral, metal-rich subsurface fluids of Henderson mine, CO. *Co. Chem. Geol.* 284, 339–350. doi:10.1016/j.chemgeo.2011.03.015

Takai, K., Moser, D. P., DeFlaun, M., Onstott, T. C., and Fredrickson, J. K. (2001). Archaeal diversity in waters from deep South African gold mines. *Appl. Environ. Microbiol.* 67, 5750–5760. doi:10.1128/AEM.67.21.5750-5760.2001

Tarnas, J. D., Mustard, J. F., Sherwood-Lollar, B., Bramble, M. S., Cannon, K. M., Palumbo, et al. (2018). Radiolytic H2 production on Noachian Mars: Implications for habitability and atmospheric warming. *Earth Planet. Sci. Lett.* 502, 133–145. doi:10.1016/j.epsl.2018.09.001

Tarnas, J. D., Mustard, J. F., Sherwood-Lollar, B., Cannon, K. M., Lorand, J-P., Onstott, T. C., et al. (2021). Earth-like habitable environments in the subsurface of Mars. *Astrobiol* 21 (6), 741–756. doi:10.1089/ast.2020.2386

Templeton, A. S., and Caro, T. A. (2023). The rock-hosted biosphere. *Annu. Rev. Earth Planet. Sci.* 51, 493–519. doi:10.1146/annurev-earth-031920-081957

Vera, M., Schippers, A., and Sand, W. (2013). Progress in bioleaching: fundamentals and mechanisms of bacterial metal sulfide oxidation—part A. *Appl. Microbiol. Biotechnol.* 97, 7529–7541. doi:10.1007/s00253-013-4954-2

Viang, H., and Dong, H. (2012). Biogeochemistry and geomicrobiology in extreme environments: Preface. *Geosci. Front.* 3 (3), 269–271. doi:10.1016/j.gsf.2012.03.001

Vreeland, R. H., Piselli, A. F., Jr, McDonnough, S., and Meyers, S. (1998). Distribution and diversity of halophilic bacteria in a subsurface salt formation. *Extremophiles* 2, 321–331. doi:10.1007/s007920050075

Whelan, J., Oremland, R., Tarafa, M., Smith, R., Howarth, R., and Lee, C. (1986). Evidence for sulfate-reducing and methane-producing microorganisms in sediments from sites 618, 619, and 622. *Initial Rep. deep sea Drill. Proj.* 96, 767–775. doi:10.2973/dsdp.proc.96.147.1986

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc. Nat. Acad. Sci. U. S. A.* 95 (12), 6578–6583. doi:10.1073/pnas.95.12.6578

Wilkins, M., Daly, R., Mouser, P., Trexler, R., Wrighton, K., Sharma, S., et al. (2014). Trends and future challenges in sampling the deep terrestrial biosphere. *Front. Microbiol.* 5, 481. doi:10.3389/fmicb.2014.00481

Wu, X., Holmfeldt, K., Hubalek, V., Lundin, D., Åström, M., Bertilsson, S., et al. (2016). Microbial metagenomes from three aquifers in the Fennoscandian shield terrestrial deep biosphere reveal metabolic partitioning among populations. *ISME J.* 10, 1192–1203. doi:10.1038/ismej.2015.185

Wu, X., Pedersen, K., Edlund, J., Eriksson, L., Åström, M., Andersson, A. F., et al. (2017). Potential for hydrogen-oxidizing chemolithoautotrophic and diazotrophic populations to initiate biofilm formation in oligotrophic, deep terrestrial subsurface waters. *Microbiome* 5, 37. doi:10.1186/s40168-017-0253-y

Zhang, G., Dong, H., Xu, Z., Zhao, D., and Zhang, C. (2005). Microbial diversity in ultra-high-pressure rocks and fluids from the Chinese Continental Scientific Drilling Project in China. *Appl. Environ. Microbiol.* 71, 3213–3227. doi:10.1128/AEM.71.6.3213-3227.2005

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 17 August 2023
DOI 10.3389/fspas.2023.1211784

Check for updates

OPEN ACCESS

EDITED BY
Ryan C. Fortenberry,
University of Mississippi, United States

REVIEWED BY
Marcin Gronowski,
Institute of Physical Chemistry, Poland
Giuseppe Cassone,
National Research Council (CNR), Italy

*CORRESPONDENCE
Cristina Puzzarini,
cristina.puzzarini@unibo.it

RECEIVED 25 April 2023
ACCEPTED 31 July 2023
PUBLISHED 17 August 2023

CITATION
Puzzarini C, Alessandrini S, Bizzocchi L, Melosso M and Rivilla VM (2023), From the laboratory to the interstellar medium: a strategy to search for exotic molecules in space.
Front. Astron. Space Sci. 10:1211784.
doi: 10.3389/fspas.2023.1211784

COPYRIGHT
© 2023 Puzzarini, Alessandrini, Bizzocchi, Melosso and Rivilla. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

From the laboratory to the interstellar medium: a strategy to search for exotic molecules in space

Cristina Puzzarini[1]*, Silvia Alessandrini[1], Luca Bizzocchi[1], Mattia Melosso[1] and Víctor M. Rivilla[2]

[1]ROT&Comp Lab, Department of Chemistry "Giacomo Ciamician", University of Bologna, Bologna, Italy, [2]Centro de Astrobiología (CAB, CSIC-INTA), Madrid, Spain

The chemistry of the interstellar medium occurs under extreme conditions and can lead to the formation of exotic molecules. These are species that on Earth are unstable and/or highly reactive. Their discovery in space is usually based on the astronomical observation of their rotational fingerprints, which requires an accurate laboratory investigation. This is based on a strategy that starts from the interplay of experiment and theory. State-of-the-art quantum-chemical calculations are used to predict the relevant spectroscopic information required to guide the spectral recording, analysis and assignment. Rotational spectra measurements are then performed in the centimeter-/millimeter-/submillimeter-wave region, thereby exploiting efficient on-the-fly production protocols for exotic molecules. Subsequently, the spectral analysis leads to accurate spectroscopic parameters, which are then used for setting up accurate line catalogs for astronomical searches and detections. This review is based on the strategy developed and the results obtained at the ROT&Comp Lab of the University of Bologna.

KEYWORDS

astrochemistry, interstellar medium, unstable/reactive species, quantum chemistry, rotational spectroscopy, astronomical observations

1 Introduction

The chemical composition of the interstellar medium (ISM) systematically changes while the physical evolution proceeds from diffuse clouds toward star-forming regions (Yamamoto, 2017). Regardless of the environment, the chemistry of the interstellar space occurs under extreme conditions, with temperatures ranging between 10 and hundreds of K (with a few exceptions such as shocked regions where the gas kinetic temperature temporarily increases up to a few thousands of K), very low densities varying from 1 to 10^8 cm^{-3}, and ionizing radiation (Yamamoto, 2017; Tielens, 2021). Because of such extreme conditions, the chemical reactions that involve either ions or radicals as one of the reactants are effective and efficient. These reactions are unusual or even unfeasible under terrestrial conditions. The ISM is also characterized by turbulence and shocked regions (Yamamoto, 2017; Tielens, 2021). Shock-wave-driven chemistry contributes to widen the unusual chemical reactivity and leads to the onset of both transient and stable molecular species, with different shock-wave velocities triggering different chemical processes (Cassone et al., 2018; García de la Concepción et al., 2021). In short, the ISM is

characterized by a chemistry that can lead to the formation of unstable and/or highly reactive molecules, here broadly denoted as "exotic" (Negron-Mendoza and Ramos-Bernal, 2001; Yamamoto, 2017; Puzzarini, 2020; Herbst, 2021). These are radicals and ions, but also closed-shell unsaturated systems such as imines and carbon chains. Owing to the very low density of the ISM, these reactive species survive long enough to be detected (Müller et al., 2005; McGuire, 2018; Rivilla et al., 2018; Tennyson, 2019; Bizzocchi et al., 2020; McGuire, 2022; Alberton et al., 2023a). For the sake of completeness, we extend the classification of exotic molecules to unusual species, such as deuterated systems and functionalized building-blocks of polycyclic atomatic hydrocarbons (PAHs).

Molecules can be observed through their electronic, vibrational or rotational transitions at optical/ultraviolet, infrared or centi-/milli-/submillimeter wavelengths, respectively. However, more than 90% of the gas-phase molecules discovered in space so far have been detected owing to their rotational fingerprints. For this reason, in this review, laboratory spectroscopy refers to rotational spectroscopy and astronomical observations are those carried out with radiotelescopes in the microwave region, from centimeter to submillimeter wavelengths. In this respect, in the last decade, the improved sensitivity of ground-based facilities such as ALMA (Wootten and Thompson, 2009), GBT (Prestage et al., 2009), and Yebes (Tercero et al., 2021) have led to unexpected results in terms of number and complexity of molecules detected. While astronomical line surveys of a given astrophysical object should lead to a complete census of its molecular content, unknown features are still present and this is mainly due to the lack of the corresponding spectroscopic information. There is, therefore, a great need of investigating new potential interstellar molecules. The interest is mainly on short-living systems that have not been studied yet or have been poorly characterized in the lab.

In this review, the journey from the laboratory to space is addressed. To describe this journey, the strategy developed at the ROT&Comp Lab of the Department of Chemistry "Giacomo Ciamician", University of Bologna, and the results there obtained are considered by way of example. The target is limited to gas-phase molecular systems, and emphasis is put on the methodology to be exploited, with particular focus on unstable species, for obtaining very accurate line catalogs, which are then employed in astronomical searches. Illustrative systems range from potential prebiotic species such as (*Z*)-1,2-ethenediol (Melosso et al., 2022; Rivilla et al., 2022) and allylimine (Alberton et al., 2023a) to radical species, passing through deuterated ammonia, a long carbon chain, and substituted benzene. The first two systems are typical examples of exotic molecules: closed-shell species that are, however, too reactive to be isolated and stored, thus requiring to be produced on-the-fly while performing spectroscopic measurements. The same applies to the radical here considered: a sulfur-bearing species, sulfeno (HSO; Cazzoli et al., 2016; Marcelino, N. et al., 2023), which provides a piece of the "missing sulfur" puzzle. For a stable and abundant molecule such as ammonia, the mono-deturated isotopologue (NH_2D) has been considered (Melosso et al., 2020b; Melosso et al., 2021), which had to be prepared *ad hoc* inside the cell and allows for pointing out the importance of accounting for hyperfine effects in the rotational spectrum for the correct interpretation of astronomical spectra. Finally, a long carbon chain (DC_7N; Puzzarini et al., 2023) and isomers of cyanoethynylbenzene (Spaniol et al., 2023) have been taken as examples of larger systems.

This review paper is organized as follows. In the next section, the methodology is presented in detail, with a subsection dedicated to each step of the strategy as outlined in Figure 1. Then, the results section addresses the case studies mentioned above. It has to be noted that, while at the ROT&Comp Lab (in collaboration with V.M. Rivilla) a great effort has been taken to rationalize the strategy here presented, this is not at all unique and several research groups employ methodologies that rely on the interplay of experiment and theory for the spectroscopic characterization which is then exploited for astronomical searches. Examples in this direction (the list being not at all exhaustive)

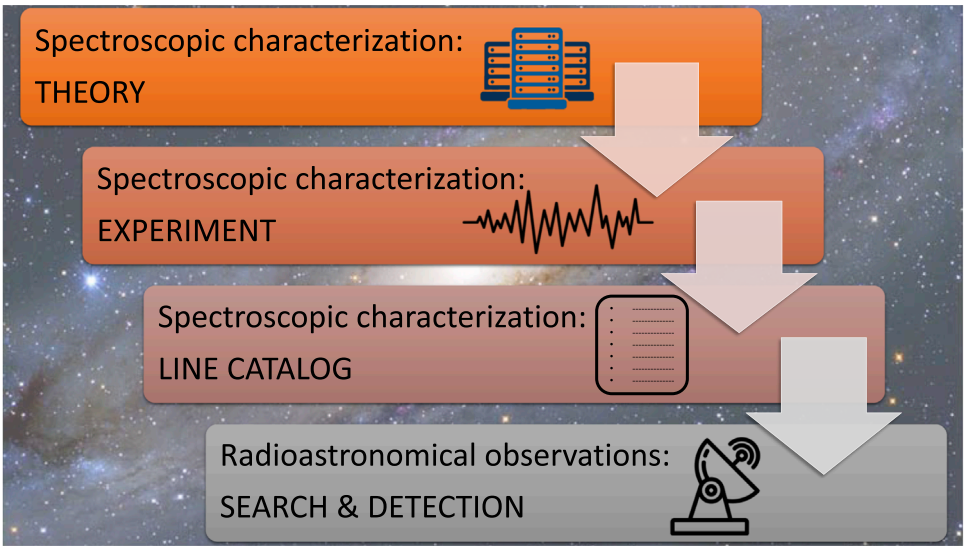


FIGURE 1
From the laboratory to space: Main steps of the strategy leading to the detection of gas-phase exotic species in the ISM.

are provided by Belloche et al. (2017); Bermúdez et al. (2018); Bizzocchi et al. (2020); McCarthy et al. (2021); Lee et al. (2021); Cabezas et al. (2022a); Changala et al. (2022); Koelemay et al. (2022); Alberton et al. (2023b); Cernicharo et al. (2023); Silva et al. (2023).

2 Methodology

The main steps of the strategy pursued for the detection of new molecules in the ISM are introduced in Figure 1, with "new" denoting a molecular species for which the rotational spectroscopic information is entirely missing or not sufficient for astronomical purposes. As explained in the Introduction, these are often exotic molecules. The protocol starts from an accurate characterization of the spectroscopic parameters of interest by means of quantum-chemical calculations. These are used to simulate the rotational spectrum, the simulation allowing to guide, support and complement the next step: the experiment. This latter starts with the set up of the conditions for the efficient production of the target molecular species. Based on the simulated spectrum, the best frequency ranges are then chosen. The analysis of the experimental measurements is carried out by relying on the comparison between computed and recorded spectra. Such an analysis then leads to the accurate determination of the spectroscopic parameters required for setting up a line catalog to be used for guiding astronomical searches. The assignment of a sufficient number of unblended features in the astronomical survey to the rotational transitions belonging to the molecule under consideration leads to its detection and allows the derivation of its column density. These steps are detailed in the following sections.

2.1 Spectroscopic characterization: theory

The spectroscopic parameters required for predicting the rotational spectrum of a given molecule are its rotational and centrifugal distortion constants (Gordy and Cook, 1984; Puzzarini et al., 2010; Barone et al., 2021). In some cases, magnetic and/or electric interactions can occur and lead to the so-called hyperfine structure of the rotational spectrum (Gordy and Cook, 1984; Puzzarini et al., 2010; Barone et al., 2021). Even if, in most of the cases, these effects are not observable in astronomical spectra (i.e., they do not affect the transition frequencies in astronomical surveys), there are cases where hyperfine interactions play a role (Puzzarini et al., 2010; Guzmán et al., 2012; Barone et al., 2021; Melosso et al., 2021).

According to vibrational perturbation theory to second order (VPT2; Mills, 1972), rotational constants of a given vibrational state (B^i_{vib}) consist of two terms:

$$B^i_{vib} = B^i_e - \sum_r \alpha^i_r \left(\vee_r + \frac{d_r}{2} \right). \quad (1)$$

B^i_e is the equilibrium rotational constant relative to the ith inertial axis ($i = a, b, c$ so that $B^a_e = A_e$, $B^b_e = B_e$, $B^c_e = C_e$), and only depends on the equilibrium structure and the isotopic composition. The second term is the vibrational correction (required to go from the equilibrium to the vibrational state of interest), which is obtained by a sum, running over the r vibrational modes ($\vee_r$ being the vibrational quantum number associated to the rth vibrational mode and d_r its degeneracy), of the vibration-rotation interaction constants (α^i_r). Because of the very low temperatures typical of the ISM, we are usually interested in the rotational constants of the vibrational ground state (see Melosso et al., 2020a for an example where vibrational excited states are considered of relevance), Eq. 1 thus becoming:

$$B^i_0 = B^i_e - \frac{1}{2} \sum_r d_r \alpha^i_r = B^i_e + \Delta B^i_{vib}. \quad (2)$$

In the equation above, the equilibrium term contributes to about 99% of the vibrational ground-state rotational constant, B^i_0 (Puzzarini et al., 2010; Puzzarini and Stanton, 2023). Therefore, from a computational point of view, most of the effort is put on the accurate determination of the equilibrium structure that straightforwardly provides B^i_e (Puzzarini et al., 2008; Puzzarini et al., 2010; Alessandrini et al., 2018). The vibrational correction ΔB^i_{vib} requires the calculation of the α^i_r constants, and thus implies anharmonic force field computations (Puzzarini et al., 2010; Puzzarini et al., 2019; Puzzarini and Stanton, 2023).

To reach high accuracy in equilibrium structure determinations, basis-set and electron-correlation effects should be taken into account. This can be accomplished by exploiting quantum-chemical composite schemes, which account for the different contributions separately in order to improve their evaluation. Indeed, they are computed at the highest possible level and then combined to obtain the best theoretical estimate, thereby exploiting the additivity approximation. In the literature, several composite schemes have been proposed to find the best compromise between accuracy and computational cost (Császár et al., 1998; Tajti et al., 2004; Heckert et al., 2005; Heckert et al., 2006; Karton et al., 2006; Huang and Lee, 2009; Puzzarini et al., 2010; Puzzarini et al., 2019; Puzzarini and Barone, 2011; Karton and Martin, 2012; Peterson et al., 2012; Cheng et al., 2017; Morgan et al., 2018; Lupi et al., 2021).

For this specific review, we mention the so-called CBS+CV "gradient scheme", which is entirely based on coupled-cluster (CC) theory (Heckert et al., 2005; Heckert et al., 2006), and is implemented in the CFOUR quantum-chemistry package in a black-box manner (Matthews et al., 2020). In this approach, the additivity approximation is exploited at the energy-gradient level, which has to be minimized for obtaining the equilibrium structure. This scheme employs the CCSD(T) method, the acronym standing for the CC singles and doubles approximation augmented by a perturbative treatment of triple excitations (Raghavachari et al., 1989). In the CBS+CV approach, CBS means that the extrapolation to the complete basis set (CBS) limit is accomplished, while CV refers to the incorporation of the core-valence correlation contribution. In more details, the energy gradient contains three terms. The first two contributions are the gradients for the extrapolation of the HF-SCF and CCSD(T) correlation energies to the CBS limit, which are performed separately. For the former, the exponential extrapolation formula by Feller (1993) is employed, while the n^{-3} (n being the basis-set cardinal number) extrapolation scheme by Helgaker et al. (1997) is used for the CCSD(T) correlation energy. Since the latter extrapolation is performed within the frozen-core (fc) approximation, the

third term incorporates the missing CV correlation effects, the corresponding corrective term being obtained as difference between all-electron (ae) and fc CCSD(T) calculations in the same core-valence basis set. This latter is the cc-p(w)CV*n*Z set (Woon and Dunning Jr, 1995; Peterson and Dunning Jr, 2002). For the extrapolations to the CBS limit, the correlation-consistent valence cc-pV*n*Z bases are instead used (Dunning Jr, 1989; Wilson et al., 1996). To further improve the equilibrium structure, it is requires to go beyond the CCSD(T) model, and thus to incorporate corrections due to a full treatment of triples (fT) and quadruple excitations (fQ). This leads to the composite scheme usually denoted as CCSD(T)/CBS+CV+fT+fQ. As demonstrated in Puzzarini et al. (2008) and Alessandrini et al. (2018), CBS+CV and CCSD(T)/CBS+CV+fT+fQ equilibrium rotational constants have a relative accuracy better than 0.1% and 0.05%, respectively. If the dimension of the molecule is increased and the CBS+CV model is not affordable (such as in the case of substituted benzene or small building blocks of biomolecules), accurate equilibrium structures can be obtained by resorting to the so-called "cheap-composite" scheme (Puzzarini and Barone, 2011; Lupi et al., 2021), which is denoted by the acronyms ChS or junChS depending on the type of basis sets employed (junChS employs the partially augmented jun-cc-pV*n*Z basis sets by Papajak et al., 2011). This starts from fc-CCSD(T)/(jun-)cc-pVTZ calculations and incorporates the extrapolation to the CBS limit and the CV correlation effects at a lower level of theory (i.e., using second-order Møller-Plesset perturbation theory, MP2; Møller and Plesset, 1934). For larger systems for which even the (jun)ChS approach is not affordable (such as small PAHs), accurate equilibrium geometries can be obtained by means of the so-called "Lego brick" approach (Melli et al., 2021; Ye et al., 2022). Within it, the system is seen as formed by smaller fragments (i.e., the "Lego bricks") whose accurate equilibrium geometries are available. The template molecule (TM) approach (Piccardo et al., 2015) is then used to account for the modifications occurring when moving from the isolated fragment to the same fragment within the molecular system under consideration. For the TM approach, we resort to the double-hybrid rev-DSDPBEP86 (revDSD) density functional (Santra et al., 2019), also incorporating dispersion corrections (D3BJ; Grimme et al., 2010; Grimme et al., 2011), in conjunction with the jun-cc-pVTZ triple-zeta basis set. For the accurate determination of the equilibrium structure of the fragments, we employ the so-called semi-experimental (SE) structures (Pulay et al., 1978; Piccardo et al., 2015; Demaison et al., 2016), which are determined by a linear least-squares fit of the SE equilibrium rotational constants of different isotopologues (these are obtained by correcting the experimental B_0^i constants for the computed ΔB_{vib}^i corrections; Pulay et al., 1978). To correct the linkage between different fragments, the linear regression (LR) model is then employed (Penocchio et al., 2015), thereby exploiting the database from Ceselin et al. (2021). In the following, we refer to the "Lego brick" approach as TM+LR, or only TM whenever LR is not applied.

For the vibrational corrections to rotational constants, as mentioned above, this contribution requires anharmonic force field calculations. Since ΔB_{vib} accounts for less than 1% of B_0, for its evaluation, exploitation of CC-based composite scheme is not required all. Indeed, the fc-CCSD(T)/cc-pVTZ level of theory is more than suitable. Actually, lower levels of theory involving the MP2 method or double-hybrid functionals (Grimme, 2006; Santra et al., 2019) perform very well in this respect (Puzzarini et al., 2019; Ye et al., 2022). As far as centrifugal-distortion constants are concerned, their evaluation involves force field computations as well and, thus, they can be obtained as byproducts of the calculations leading to the evaluation of ΔB_{vib} (Puzzarini et al., 2010; Puzzarini et al., 2019). In detail, for the quartic terms, the knowledge of the harmonic force field is required, while the sextic terms need the anharmonic one (up to cubic force constants).

The approaches described above are able to provide predictions for rotational constants with a relative accuracy of 0.1%, while centrifugal distortion constants usually show deviations from experiment ranging from 5% to 10%. Overall, rotational transitions are predicted with uncertainties that mainly reflect the accuracy of the computed rotational constants. As pointed out in Puzzarini et al. (2008), Puzzarini et al. (2010) and Alessandrini et al. (2018), pushing theory to the limit allows one to reduce the relative errors on transition frequencies to less than 0.05%. However, this means that for a line at 100 GHz, the uncertainty is 50 MHz. While such an accuracy is suitable for guiding experiment in the lab, it is not at all sufficient for supporting astronomical searches. Furthermore, unlike vibrational spectroscopy, rotational spectroscopy does not show systematic effects that can be recovered by, e.g., *ad hoc* scaling factors.

For closed-shell species, electric and/or magnetic interactions occur whenever they contain one or more atoms with non-zero nuclear spin I. These split rotational energy levels and, consequently, rotational transitions giving rise to the hyperfine structure of the rotational spectrum. The interaction of major interest is the nuclear quadrupole coupling that takes place whenever $I \geq 1$: the quadrupole moment of the quadrupolar nucleus interacts with the electric field gradient at the nucleus itself. From a computational point of view, the prediction of nuclear quadrupole coupling constants requires the calculation of the electric field gradient at the quadrupolar nucleus:

$$\chi_{ij} = eQq_{ij}, \tag{3}$$

where i and j refer to the inertial axes. eQ is the nuclear quadrupole moment and q_{ij} represents the ij-th element of the electric-field gradient tensor (Puzzarini et al., 2010). Based on the literature on this topic (Puzzarini et al., 2009; Puzzarini et al., 2010; Puzzarini et al., 2019), ae-CCSD(T) computations in combination with core-valence basis sets and incorporating vibrational corrections provide quantitative predictions. In the case of open-shell species, the situation is more involved because the coupling of the rotational angular momentum with the electronic spin momentum should be incorporated in the treatment (Gordy and Cook, 1984; Cazzoli et al., 2016). If more than one unpaired electrons are present, the coupling between the electronic spins should also been considered (Gordy and Cook, 1984; Cazzoli et al., 2016; Bizzocchi et al., 2022).

2.2 Spectroscopic characterization: experiment

As it follows from the discussion above, theory is not able to provide line catalogs with the accuracy required by radioastronomy.

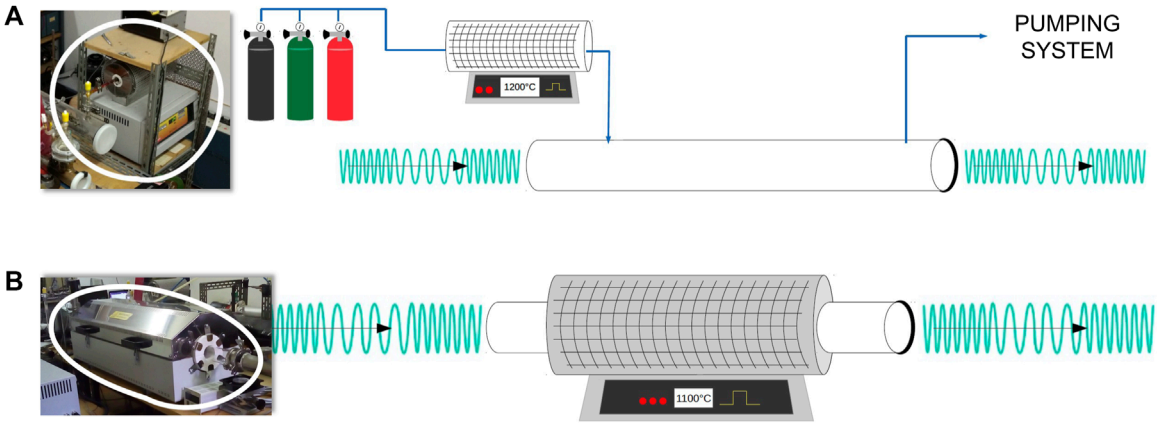


FIGURE 2
The two FVP systems available at the ROT&Comp Lab (the millimeter-/submillimeter-wave radiation is represented by green waves). **(A)** external FVP apparatus. **(B)** *in situ* FVP system. In both panels, pictures of the FVP oven are given on the left side.

For this reason, a mandatory step of our strategy is the experimental spectroscopic characterization of the molecular species to be searched in the ISM.

On general grounds, rotational spectroscopy is performed from the centimeter-wave to the submillimeter-wave region. The rotational spectrometers at the ROT&Comp lab[1] are frequency-modulation millimeter/sub-millimeter-wave (mm/sub-mm) spectrometers, which work in the 75–1.6 THz and 70–990 GHz frequency ranges. For interested readers, the former is described in detail in Cazzoli et al. (2012, 2016), Melosso et al. (2019a) and Melosso et al. (2019b), while for the latter the production phase has just started. To produce unstable molecules, such as ions and radicals, but also species that are highly reactive under terrestrial conditions, such as imines and carbon chains, the mm/sub-mm spectrometer is combined with either a flash vacuum pyrolysis (FVP) system (Melosso et al., 2018; Jiang et al., 2021; Melli et al., 2022) or a DC discharge apparatus (Cazzoli et al., 2012; Cazzoli et al., 2016). The FVP technique is mainly used for the production of closed-shell species, while DC discharge is mostly employed for the generation of radicals and ions (Puzzarini et al., 2023).

Concerning the FVP technique, two different apparatuses can be used, which are schematized in Figure 2. The first system is external to the spectrometer, it consists of a 30 cm long tubular oven surrounding a quartz tube directly connected to the inlet of the spectrometer cell (see Figure 2A). The maximum temperature reached by this oven is 1,200°C. The second apparatus is defined as an *in situ* system (see Figure 2B) because a 1.5 m long quartz tube is employed as absorption cell and is surrounded by a 90 cm tubular furnace (Jiang et al., 2021). In analogy to the first system, a maximum temperature of 1,200°C can be reached. For both apparatuses, in view of instability or reactivity of the target molecules, measurements are performed in dynamical conditions, which means that a tenuous flow of fresh pyrolysis products is continuously assured by the vacuum system.

For the production of ionic and radical species, the mm/sub-mm spectrometer is equipped with a negative glow-discharge cell (Cazzoli et al., 2012; Cazzoli et al., 2016), with measurements being performed at low temperatures to improve the signal-to-noise ratio. The experimental apparatus is schematized in Figure 3. Together with a suitable mixture of precursors, a buffer gas (usually, He or Ar) is often employed and a DC discharge ranging from a few mA to some tens of mA is applied. Analogously to FVP, measurements are carried out in dynamical conditions. During measurements, a longitudinal magnetic field can be applied: for ions, it is used for extending the negative glow discharge (thus improving ions production); for radicals, Zeeman effect disentangles their signal from those of the closed-shell species (such as precursors or undesired products).

The resolution of rotational spectra is mainly limited by two effects, collisions and Doppler effect, which determine a broadening of lines. The collisional broadening, also denoted as pressure broadening, can be reduced by working at a sufficiently low pressure. The Lamb-dip technique allows for getting rid of the Doppler broadening, thus leading to a resolution of a few tens of kHz (see right panel of Figure 4) and a measurement accuracy of 0.5–5 kHz (Cazzoli et al., 2004; Cazzoli et al., 2009; Cazzoli et al., 2015; Puzzarini et al., 2009). Lamb-dip technique requires three conditions to be fulfilled.

- measurements are carried out at very low pressures in order to have rotational lines only broadened by the Doppler effect;
- measurements are performed using radiation power sufficiently intense to lead to partial saturation;
- the radiation should go back and forward inside the cell (double-pass arrangement) in order to have a pump-and-probe effect.

To exploit the double-pass configuration in the mm/sub-mm spectrometer, a roof-top mirror is located at the end of the cell to reflect the radiation back and rotate its polarization by 90°, while a wire-grid polarizer is placed at the other end. The overall result is that

1 https://site.unibo.it/rotational-computational-spectroscopy/en

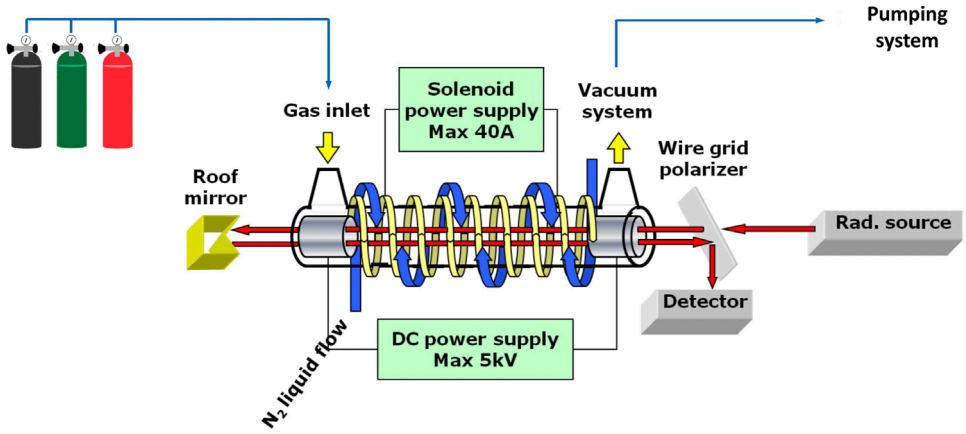


FIGURE 3
Scheme of the DC discharge system of the frequency-modulated mm-/submm spectrometer at the ROT&Comp Lab.

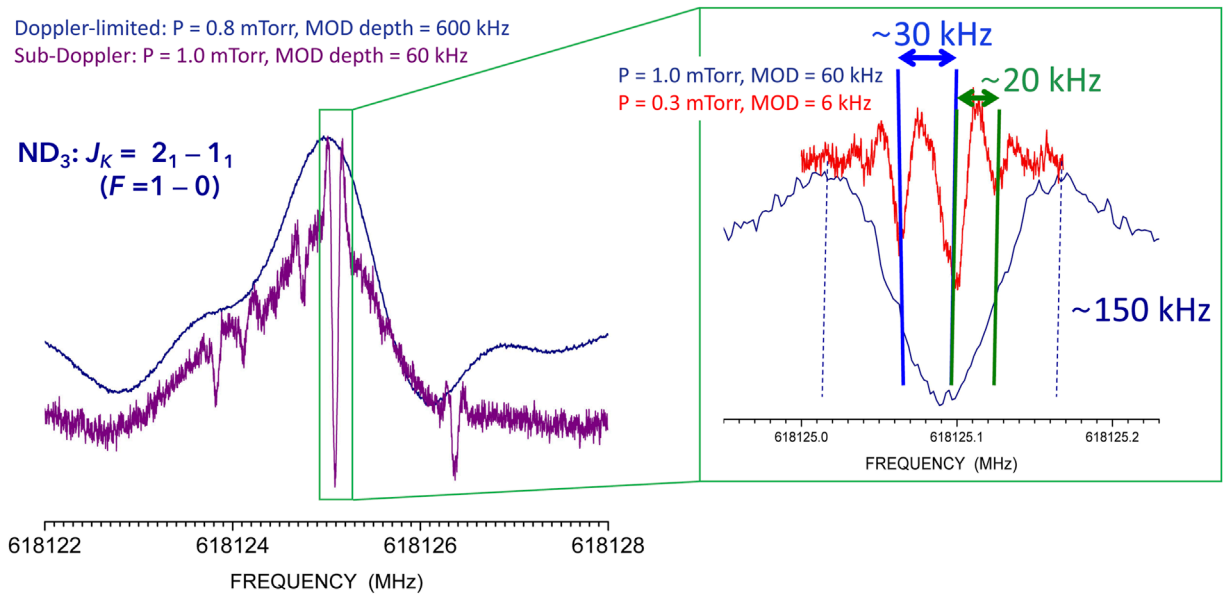


FIGURE 4
Lamb-dip recording of the $F = 1 \leftarrow 0$ hyperfine component of the $J_K = 2_1 \leftarrow 1_1$ transition of ND_3. Left panel: Comparison between the Doppler-limited (dark blue line) and Lamb-dip (violet line) recordings. Right panel: zoom of the central hyperfine component (dark blue line) showing improved resolution (red line).

the radiation enters the cell passing through the polarizer, whereas the radiation back from the roof-top mirror with its polarization rotated by 90° is reflected at right angle onto the detector.

While hyperfine structures arising from strong interactions are evident in Doppler-limited rotational spectra (see left panel of Figure 4), weaker effects might be concealed by the Doppler profile and need sub-Doppler resolution to be disclosed (see Figure 4). Furthermore, when several interactions take place, the Doppler-limited profiles are distorted and preclude accurate frequency determinations. Figure 4 showing part of the hyperfine structure of the $J_K = 2_1 \leftarrow 1_1$ transition of ND_3 (J,K being the rotational quantum numbers for symmetric rotors), provides a significant example. In the case of ND_3, both N and D are quadrupolar nuclei; however, nitrogen gives rise to a stronger quadrupole coupling than deuterium and the corresponding hyperfine components are well resolved, In the left panel, the $F = 1 \leftarrow 0$ hyperfine component is depicted, F being the quantum number associated to the hyperfine interaction due to N. The Doppler-limited (dark blue) and sub-Doppler (violet) profiles are compared: it is evident that a complex hyperfine structure due to the three D nuclei is hidden in the former. The right panel shows that the hyperfine structure of the left panel is even more complicated (red profile): the resolution of the Lamb-dip technique can be further pushed by adjusting the pressure inside the cell and the modulation amplitude (MOD). As evident in the right

panel, features separated by ~20 kHz are well resolved. The overall consequence is that, whenever the Lamb-dip technique is exploited (even in absence of hyperfine structure), the transition frequency is retrieved with great accuracy. It has to be noted that the spectral analysis needs to account for all interactions occurring. In the present case, these include (in addition to nitrogen and deuterium quadrupole couplings) nuclear spin-rotation interactions for all nuclei and all possible dipolar spin-spin couplings (Puzzarini et al., 2009; Puzzarini et al., 2015; Melosso et al., 2020b; Melosso et al., 2021).

2.3 Spectroscopic characterization: line catalog

Application of the contact transformation perturbation procedure to the Watson Hamiltonian leads to a block-diagonal effective Hamiltonian, whose pure rotational and centrifugal-distortion contributions are here considered (Aliev and Watson, 1985):

$$\mathbf{H}_{rot} = \mathbf{H}_{02} + \mathbf{H}_{04} + \mathbf{H}_{06}, \tag{4}$$

where $\mathbf{H}_{04}$ and $\mathbf{H}_{06}$ are the quartic and sextic centrifugal-distortion terms, respectively, and $\mathbf{H}_{02}$ is the rigid-rotor Hamiltonian:

$$\mathbf{H}_{02} = \sum_i B_e^i \mathbf{J}_i^2, \tag{5}$$

where B_e^i is the equilibrium rotational constant mentioned above.

Even if the rigid-rotor Hamiltonian $\mathbf{H}_{02}$ is the most important term, a realistic and accurate description of the rotational problem requires to go beyond the rigid-rotor approximation, thus including in the treatment at least the $\mathbf{H}_{04}$ Hamiltonian, which is expressed in terms of the quartic centrifugal distortion parameters (this requires some contact transformations; interested readers are referred–for example–to Aliev and Watson, 1985). Moving to higher terms, a simple general formula for $\mathbf{H}_{06}$ was obtained by Aliev and Watson (1976). Different results are then obtained depending on the Hamiltonian reduction chosen (Watson, 1968; Watson, 1977; Aliev and Watson, 1985). Moreover, the effect of molecular vibrations should be taken into account. Therefore, B_e^i in Eq. 5 should be replaced by the effective rotational constant of Eq. 2.

If hyperfine interactions are present, the Hamiltonian incorporates their contribution:

$$\mathbf{H} = \mathbf{H}_{rot} + \mathbf{H}_{hfs}, \tag{6}$$

where $\mathbf{H}_{rot}$ is the rotational part of the Hamiltonian operator introduced above, also including centrifugal distortion. $\mathbf{H}_{hfs}$ accounts for all possible hyperfine interactions, with the nuclear quadrupole coupling being the one of interest in this context. This latter is described by making use of the traceless tensor $\boldsymbol{\chi}$ of Eq. 3. Other possible hyperfine interactions are the nuclear spin-rotation (SR) interaction (described in terms of the $\boldsymbol{C}$ tensor) and, if more than one nuclei with nonzero $\mathbf{I}$ are present, the dipolar spin-spin coupling. Hyperfine interactions lead to the definition of one or more additional quantum numbers (usually denoted with F or, if more than one, F_x with x denoting a coupling order or a coupling atom).

Moving to radical species, the effective Hamiltonian operator contains at least two contributions:

$$\mathbf{H} = \mathbf{H}_{rot} + \mathbf{H}_{fs}, \tag{7}$$

where $\mathbf{H}_{rot}$ is the usual rotational Hamiltonian term as above. $\mathbf{H}_{fs}$ is the fine structure Hamiltonian, which is—at least—due to electron spin-rotation interaction (described through the electron spin-rotation tensor $\boldsymbol{\epsilon}$). In all cases, the rotational angular momentum $\mathbf{N}$ and the electronic spin momentum $\mathbf{S}$ couple to lead to $\mathbf{J}$ (therefore, for radicals, the rotational quantum number is N and not J). If more than one electron is present, then $\mathbf{H}_{fs}$ also incorporates the electron spin-spin interaction (through the electron spin–spin coupling tensor $\boldsymbol{\lambda}$ which is symmetric and traceless). If nonzero nuclear spins are present in the molecule, in addition to the hyperfine interactions mentioned above, a hyperfine-structure term describing the interaction between the electron spin and the nuclear spins should be incorporated in the Hamiltonian above (Gordy and Cook, 1984; Bizzocchi et al., 2022).

The spectral analysis, guided by the simulated spectrum, leads to the assignment of the rotational transitions in terms of quantum numbers. These rotational transitions are then fitted using the effective Hamiltonian mentioned above. In the fitting procedure, the computed spectroscopic parameters provide the starting point, while the outcome is their experimental counterpart. In turn, the experimental spectroscopic parameters are employed to accurately predict all possible rotational transitions up to a given frequency limit which depends on quantum numbers sampled in the actual measurements. This collection of transitions together with their spectroscopic information is the line catalog to be used in astronomical searches. Such an application requires that the uncertainties affecting the transition frequencies are smaller than or, in the worst case, equal to 100–200 kHz, because astronomical line surveys are dense and only accurate rest frequency values allow for discriminating the features belonging to different molecules.

2.4 Radioastronomical observations: search and detection

To search for the transitions of the molecule under consideration, first of all, spectral surveys of the target molecular cloud are obtained. Then, the rotational spectroscopy information from the laboratory work needs to be incorporated into a suitable program that generates synthetic spectra under the assumption of local thermodynamic equilibrium (LTE) conditions. To properly evaluate possible line contamination by other molecules, the emissions from the molecules already identified toward that cloud are considered in the simulation. A tool able to provide the best nonlinear least-squares LTE fit to these data is then exploited, with the parameters of the LTE model being the molecular column density (N), the excitation temperature ($T_{\rm ex}$), the observed radial velocity of the astronomical object ($v_{\rm LSR}$), and full width at half maximum (FWHM) of the Gaussian line profiles (see, e.g., Martín et al., 2019).

A complementary analysis to retrieve N and $T_{\rm ex}$ is offered by the "rotational diagram method" (Goldsmith and Langer, 1999) which, for multiple transitions, plots their $\log(N_u/g_u)$ value (where N_u/g_u is calculated from the integrated intensity) against

the corresponding upper-state energy, E_u. The linear regression slope, $1/k_B T_{ex}$ (k_B being the Boltzmann constant), provides the excitation temperature, whereas the intercept, $N/Q(T)$, gives the column density once the molecular partition function at the derived temperature ($Q(T)$) is computed. It is thus crucial to have accurate values of $Q(T)$ to accurately determine N. In addition to the rotational and hyperfine contributions to $Q(T)$ that can be easily obtained from the spectroscopic characterization (carried out to obtained the line catalog), the vibrational term should also be taken into account because its contribution can be greater than 1 already at low temperature (above 20–30 K). This requires the accurate knowledge of all fundamental vibrational frequencies of the molecule under consideration. This requirement can be fulfilled by means of experimental studies or state-of-the-art quantum-chemical computations (thereby exploiting composite schemes analogous to those introduced in the theory section). More details on the method can be found, in Goldsmith and Langer (1999).

Since at the very low density of the ISM the molecular energy level populations are rarely at the LTE, collisions compete with radiative processes (Kłos and Lique, 2017). Therefore, the interpretation of observed interstellar spectra in terms of molecular abundances requires the knowledge of collisional rate coefficients of the molecule under consideration, with the most abundant perturbing species (namely, molecular hydrogen) being considered. While a few experimental studies allow the determination of such parameters (see, e.g., Orr, 1995; James et al., 1998; Roueff and Lique, 2013; Hays et al., 2022), quantum inelastic scattering calculations offer an accurate and more accessible alternative (Roueff and Lique, 2013; Kłos and Lique, 2017). The first step of this approach is the determination of the corresponding collisional potential energy surface (Roueff and Lique, 2013; Tonolo et al., 2021; Tonolo et al., 2022). The collisional calculations provide a set of inelastic cross sections as a function of the collision energy and, starting from these quantities, the corresponding excitation and de-excitation rate coefficients are obtained for the desired temperature range (Roueff and Lique, 2013; Kłos and Lique, 2017; Tonolo et al., 2021; Tonolo et al., 2022).

3 Results

To present the strategy described in the Methodology section at work, illustrative examples have been chosen. The first two studies addressed are those that led to the detection of (*Z*)-1,2-ethenediol (Melosso et al., 2022; Rivilla et al., 2022) and the tentative detection of allylimine (Alberton et al., 2023a) toward the G+0.693-0.027 molecular cloud, located at the Galactic Center. The discussion then moves to touch the HSO radical species. This latter has been recently detected towards several cold dark clouds (Marcelino, N. et al., 2023). The subsequent example is NH_2D, which is employed to point out the importance of accounting for hyperfine effects in astronomical spectra. Finally, as examples of larger systems, DC_7N and isomers of cyanoethynylbenzene are considered.

3.1 (*Z*)-1,2-ethenediol: a precursor of the RNA world in space

Ethenediol exists in the *E* and *Z* stereoisomeric forms, with the latter being more stable by about 20 kJ/mol (Karton and Talbi, 2014). In turn, as shown in Figure 5A, the *Z* isomer has two possible configurations, the preferred one being the *syn, anti* form (the *anti, anti* conformer lies higher in energy; see Figure 5A). (*Z*)-1,2-ethenediol belongs to the $C_2H_4O_2$ family, whose most stable isomer is acetic acid, followed by methyl formate, 1,1-ethenediol and glycolaldehyde (Karton and Talbi, 2014). Among these, prior to Melosso et al. (2022); Rivilla et al. (2022), 1,2- and 1,1-ethenediol were the only species still escaping detection in space, the reason being the lack of any rotational spectroscopy characterization.

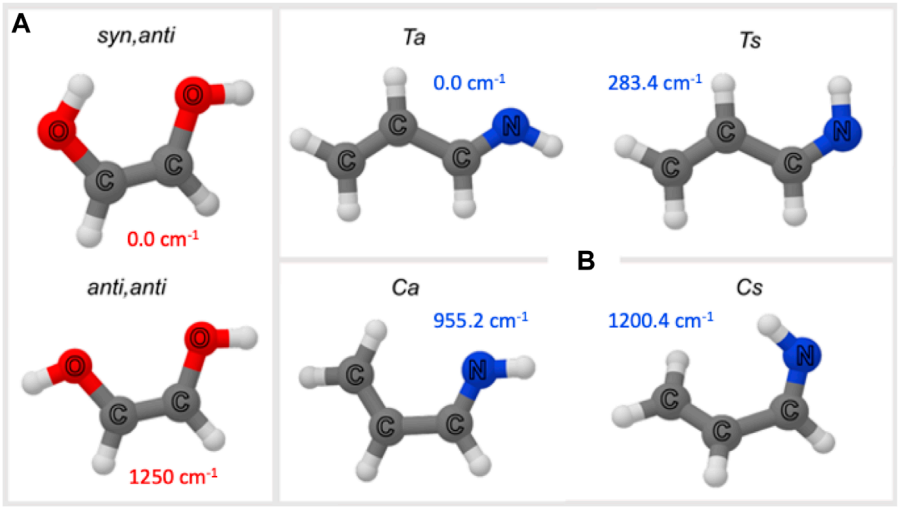


FIGURE 5
Molecular structure of (*Z*)-1,2-ethenediol **(A)** and allylimine isomers **(B)**. Relative energies are also provided (for (*Z*)-1,2-ethenediol: B2PLYP/aug-cc-pVTZ level of theory from Melosso et al., 2022; for allylimine: CBS+CV level of theory from Alberton et al., 2023a).

While we here present the strategy introduced above to 1,2-ethenediol, we mention that work is still in progress on the 1,1 isomer.

Prior to the investigation by Melosso et al. (2022), no information was available for the rotational parameters of 1,2-ethenediol. Therefore, in Melosso et al. (2022), after a preliminary scan of the potential energy surface, the equilibrium structure (and thus the equilibrium rotational constants) of the *syn, anti* form of (*Z*)-1,2-ethenediol has been evaluated at the CBS+CV level. Vibrational corrections to rotational constants as well as centrifugal distortion constants have been computed at the fc-MP2/cc-pVTZ level. Such calculations allowed the prediction of the rotational spectrum to be used for guiding the recording and the analysis of the experimental counterpart.

(*Z*)-1,2-ethenediol is a highly reactive species that requires to be produced on-the-fly while performing spectroscopic measurements. For its production, the external FVP system was employed. For this purpose, bis-exo-5-norbornene-2,3-diol was considered as suitable precursor and purposely synthesized (for details, see Melosso et al., 2022). (*Z*)-1,2-ethenediol was produced by flowing the vapours of the precursor through the quartz tube heated at 750°C. The pressure inside the absorption cell was kept at ~4 μbar. These conditions were obtained by monitoring the intensity of the (*Z*)-1,2-ethenediol transitions and adjusted to obtain the best signal-to-noise ratio (S/N). As shown in Figure 6, the oven temperature of 750 °C provides the best S/N. The temperature of the precursor also affects the production of (*Z*)-1,2-ethenediol, as evident in the right panel of Figure 6. While the best S/N is obtained by heating the sample at 80°C, the working temperature was maintained at 50°C because a mild heating decreases the intensity of the interfering lines produced by glycolaldehyde (Melosso et al., 2022). Indeed, even in the best conditions, signals belonging to glycolaldehyde and cyclopentadiene are observed. The former is produced by isomerization of (*Z*)-1,2-ethenediol, while the latter is a pyrolysis co-product (Melosso et al., 2022).

Figure 1 of Melosso et al. (2022) provides an example of comparison between the simulated (based on computed spectroscopic parameters) and recorded spectra in the 255–260 GHz frequency range. That figure demonstrates that the computational strategy is very effective: it allows for obtaining a very good prediction of the rotational spectrum, thus enormously facilitating the spectral assignment. A further confirmation of the correct identification of (*Z*)-1,2-ethenediol was offered by the presence of a tiny splitting in most of the absorption features, which is a clear evidence of a tunneling motion. This was somewhat expected in view of the similarity with ethylene glycol (Christen et al., 1995). In detail, the –OH moieties undergo a concerted large amplitude tunneling motion between two equivalent positions of the *syn, anti* form of (*Z*)-1,2-ethenediol, which splits each rotational energy level into two.

The analysis of the recorded features was performed using a rotational-torsional Watson S-reduced Hamiltonian (Watson, 1977), with terms up to 8th power of the rotational angular momentum being incorporated in the Hamiltonian. This analysis led to the accurate determination of several spectroscopic parameters. For example, the rotational constants were obtained with an

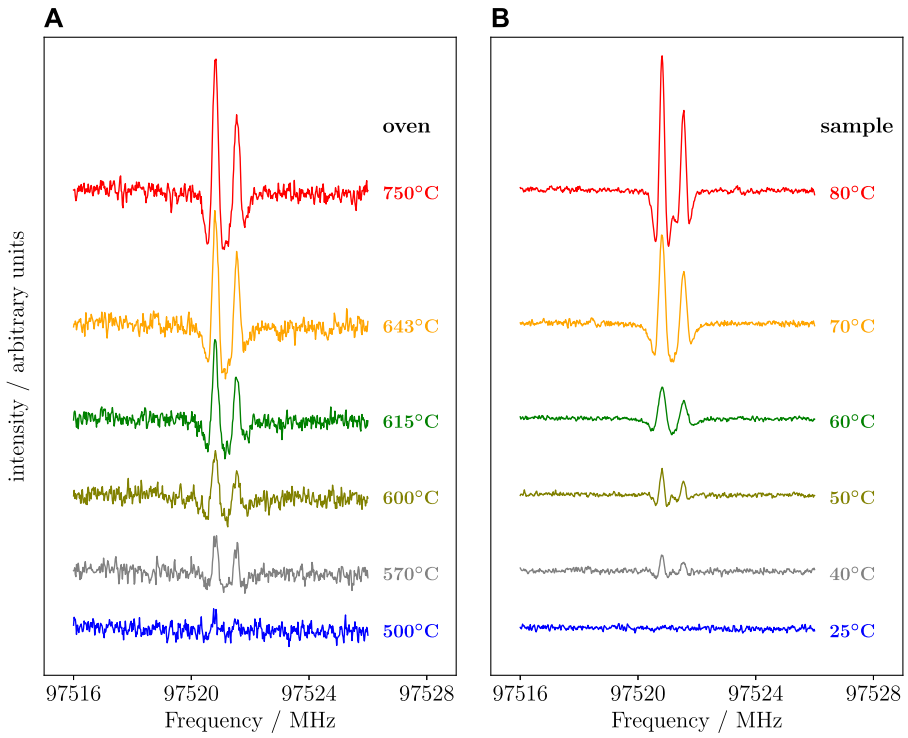


FIGURE 6
Recordings of the $9_{2,8} \leftarrow 8_{2,7}$ transition of the (*Z*)-1,2-ethenediol obtained at different pyrolysis conditions. **(A)** recording at different oven temperatures with that of precursor fixed at 55°C. **(B)** recording at different sample temperatures with the oven kept at 750°C.

average error of 0.0016%, the uncertainty of the quartic centrifugal distortion parameters was found around 0.018%, and the energy difference between the two inversion states was determined at kHz accuracy (for all details the reader is referred to Melosso et al., 2022). From these spectroscopic constants, a line catalog was generated using the SPCAT routine of the CALPGM suite (Pickett, 1991) in the format of the CDMS and JPL catalogues (Pickett et al., 1998), and was ready to be used in astronomical line-analysis tools. In detail, the fit of the recorded lines between 80 and 125 GHz (115 distinct frequency values) allowed the accurate determination of the rotational constants and few centrifugal distortion terms. These combined with the computed values for the remaining centrifugal distortion parameters predicted the transitions in the 240–375 GHz range with an average error of 600 kHz. The fit of the recorded lines up to 260 GHz (263 distinct frequency values) reduced the average error to 77 kHz for the transitions lying in the 260–375 GHz interval. The incorporation of the recorded lines up to 375 GHz led to the final fit (578 distinct lines). The corresponding parameters were used to set up the line catalog, whose predictions are expected to be very accurate up to 500 GHz and to meet the requirements for guiding astronomical observations (see section 2.3) up to 750 GHz for transitions involving low K_a values.

(*Z*)-1,2-ethenediol has been searched toward the G+0.693-0.027 molecular cloud located in the Galactic Center by exploiting high-sensitivity unbiased spectral surveys obtained with the Yebes 40 m (Guadalajara, Spain) and the IRAM 30 m (Granada, Spain) telescopes (Rivilla et al., 2022). This led to the assignment of 18 unblended or slightly blended transitions (arising from nine pairs of unresolved transitions). To correctly evaluate possible line contaminations by other molecules, the emission from the more than 120 molecules already identified toward the G+0.693-0.027 molecular cloud were considered. To derive the physical parameters of the (*Z*)-1,2-ethenediol emission, a nonlinear least-squares LTE fit to the data was performed using MADCUBA[2]. This led to an excitation temperature of 8.5 ± 0.6 K (in line with the typical values found in G+0.693-0.027) and a column density of $(1.8 \pm 0.1)\times10^{13}$ cm^{-2} (Rivilla et al., 2022), with $N(H_2)$ being 1.35×10^{23} cm^{-2} (Martín et al., 2008).

3.2 Allylimine: a prebiotic molecule

Allylimine can be seen as a sort of substituted butadiene where a terminal =CH_2 group is substituted by the =NH moiety. Therefore, in analogy to butadiene, it exists in two isomers: the *cis* (*C*) and *trans* (*T*) species (see Figure 5B). Additionally, for each of them, the position of the iminic H with respect to the closest H of the -CH=CH_2 group leads to the *syn* (*s*) and *anti* (*a*) forms (see Figure 5B). Consequently, allylimine has four geometric isomers, which are denoted as *Ta*, *Ts*, *Ca*, and *Cs* (see Figure 5B). All these forms except the latter (*Cs*) have a planar structure. At the CBS+CV level, the *Ta* form is the most stable followed by *Ts* lying 282.4 cm^{-1} above *Ta* (Alberton et al., 2023a). Contrary, the *cis* species are highly unstable (*Ca* and *Cs* lie at 955.3 and 1,200.4 cm^{-1} with respect to *Ta*, respectively) and were not experimentally observed in the spectroscopic investigation reported in Alberton et al. (2023a). Therefore, in the following, only *Ta* and *Ts* are considered.

In Alberton et al. (2023a), the CBS+CV level of theory was employed for the evaluation of their equilibrium rotational and quartic centrifugal distortion constants. While a reliable estimate of rotational constants was already available owing to the experimental studies by Penn (1978) and Brown et al. (1981) in the centimeter-wave region (8–26 GHz), these calculations were crucial to provide accurate predictions of the centrifugal terms. The impact of these latter parameters is demonstrated in Figure 7, which shows how the transition frequencies of *Ts* at 261 GHz predicted using the rotational constants by Penn (1978) (red line) do not reproduce the experimental observations (black line). The simulated spectrum noticeably improves by incorporating the computed centrifugal distortion constants (yellow spectrum), while it agrees perfectly with the recorded spectrum once the distortion terms have been determined experimentally (green spectrum). Noted is that in Figure 7 some experimental features do not have the predicted counterparts because they do not belong to *Ts*. Among the spectroscopic parameters that are relevant for predicting/analyzing the rotational spectrum of allylimine, one has to consider the nuclear quadrupole coupling constants. In fact, nitrogen is a quadrupolar nucleus ($I = 1$) and a hyperfine structure is originated (see Section 2.3). The nuclear quadrupole coupling constants were experimentally determined by Brown et al. (1981) and, in Alberton et al. (2023a), computed at the ae-CCSD(T)/cc-pCVQZ level, also including ae-MP2/cc-pCVTZ vibrational corrections.

From the experimental side, in Alberton et al. (2023a), the rotational spectrum of allylimine was recorded in selected frequency regions from 84 to 307 GHz. As most of imines, allylimine is a reactive species that cannot be synthesized and stored for measurements. Instead, in analogy to (*Z*)-1,2-ethenediol, it required to be produced on-the-fly. At the ROT&Comp Lab, measurements were performed using the *in situ* FVP system, with the optical elements of the spectrometer arranged in a double-pass configuration in order to increase the absorption path-length. Allylimine was thus produced directly inside the absorption cell by using diallylamine as precursor and a temperature of 500°C.

The computed parameters together with the available experimental information were used to simulate the rotational spectrum in the millimeter-/submillimeter-wave region, as shown in Figure 7 (yellow trace). This figure provides a flavor of how the interplay experiment-theory enormously facilitated the spectral analysis. This was carried out using the CALPGM suite of programs (Pickett, 1991) and a S-reduced Watson-type rotational Hamiltonian combined with a hyperfine-structure Hamiltonian describing the nitrogen quadrupole coupling. From the comparison of the latest results with those previously reported in Brown et al. (1981), it is noted that all spectroscopic parameters were greatly improved. Concerning the rotational constants, for *Ta*, the uncertainties were reduced by more than two orders of magnitude, with differences between the two sets of constants exceeding the associated errors. For *Ts*, an analogous improvement has to be noted, with discrepancies even more accentuated. A similar situation is observed for the nuclear quadrupole coupling constants. However, for *Ts*, the differences are as large as a few hundreds of kHz and, consequently,

2 Madrid Data Cube Analysis on ImageJ is a software developed at the Center of Astrobiology (CAB) in Madrid: https://cab.inta-csic.es/madcuba/

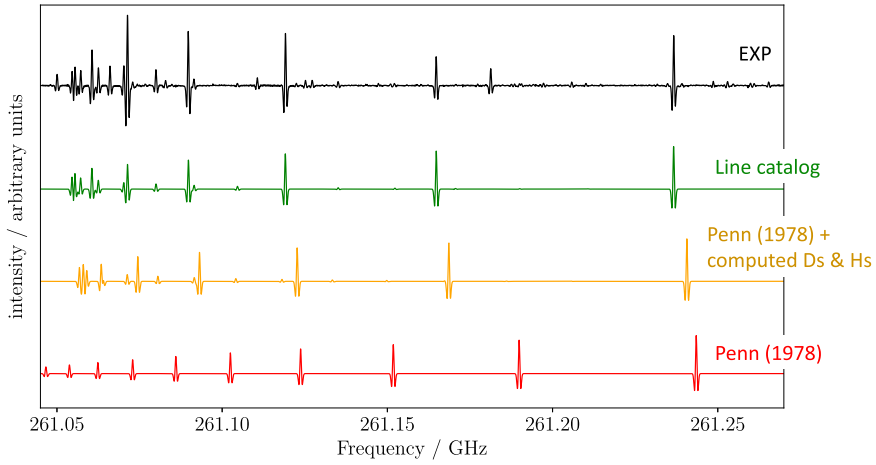


FIGURE 7
Excerpt of the rotational spectrum of *Ts*-allylimine (in black) and comparison with various simulations. The colored traces have been obtained by using: the rotational constants from Penn (1978) (red trace); the rotational constants from Penn (1978) augmented by the computed quartic and sextic centrifugal distortion constants from Alberton et al. (2023a) (yellow trace); line catalog derived in Alberton et al. (2023a) (green trace).

the computed parameters turned out to be more accurate than the old experimental determination. Quartic and sextic centrifugal distortion constants were instead determined for the first time and, as mentioned above, the calculated values turned out extremely useful for predicting the line positions in the millimeter-wave region (Figure 7). From the improved spectroscopic parameters, a line catalog for guiding astronomical observations has been generated.

In analogy to (*Z*)-1,2-ethenediol, the *Ta* and *Ts* allylimine isomers have been searched toward the G+0.693-0.027 molecular cloud. The spectroscopic parameters above have been incorporated into the MADCUBA package (Martín et al., 2019) in order to generate, under the LTE assumption, the synthetic spectra for *Ta* and *Ts* to be compared with the observed spectra. For each isomer, such a comparison led to the tentative assignment of eight transitions that were found unblended or slightly blended with emissions from other species already identified in this source. This analysis allowed the derivation of a column density $N = (1.5 \pm 0.6) \times 10^{13}$ cm^{-2} for *Ta* and $N \sim 0.5 \times 10^{13}$ cm^{-2} for *Ts*, thus resulting in a [*Ta*]/[*Ts*] isomer ratio of about 3. Interestingly, this is much lower than the estimated value of 15 based on thermodynamic considerations. Thus, the [*E*]/[*Z*] isomeric ratio of imines in the ISM is still a topic of great interest which deserves further investigations.

3.3 Sulfeno: a piece of the "missing sulfur" puzzle

The HSO radical has been the subject of a few rotational studies. Endo et al. (1981) provided the first spectral characterization of this radical together with that of its deuterated isotopologue. In Cazzoli et al. (2016), the investigation of the rotational spectrum of HSO was extended up to the THz region, thus improving and enlarging the original dataset of spectroscopic parameters. The work of Cazzoli et al. (2016) was also supported by state-of-the-art quantum-chemical calculations, with particular focus on sextic centrifugal distortion constants as well as fine and hyperfine parameters (Cazzoli et al., 2016).

In Cazzoli et al. (2016), HSO was prepared by DC discharging (~10–40 mA) a 1:2 mixture of H_2S and O_2, with starting pressures of about 3–10 mTorr of H_2S, but without any buffer gas. In addition, sulfur adsorbed on cell walls because of the use of sulfur-bearing chemicals in previous unrelated experiments allowed the production of HSO by DC discharging (50–60 mA) a flow of ~35–50 mTorr of H_2O. In both cases, measurements were performed in a continuous flow of gas, in order to constantly provide fresh precursors, and using the apparatus shown in Figure 3. Zeeman effect (due to application of a magnetic field of several Gauss) allowed to confirm that the recorded features belonged to a radical species. The line catalog (at 1, 2, and 3 mm) obtained in Cazzoli et al. (2016) was also employed to carry out a search for HSO in two high-mass star-forming regions, namely, Orion KL and Sagittarius (Sgr) B2, and in the cold dark cloud Barnard 1 (B1-b). However, no HSO lines were found above the confusion limit of the surveys. In case of the dense core B1-b, in Cazzoli et al. (2016), the search did not involve the most favorable transitions, which—at very low temperatures—lie at lower frequencies than those considered.

Very recently, the spectral line surveys performed at the Yebes 40 m telescope at 7 mm allowed the discovery of HSO towards several cold dark clouds, namely, B1-b, TMC-1, L183, L483, L1495B, L1527, and Lupus-1A (Marcelino, N. et al., 2023), thus demonstrating that HSO is widespread in cold dense cores. The detection was mainly confirmed by the observation of one or more fine/hyperfine components of the fundamental transition lying at 39 GHz, which were not accurately predicted by the line catalog from Cazzoli et al. (2016), the average uncertainty on the fine/hyperfine components being ~200 kHz. Therefore, the measured frequencies (from astronomical spectra) were used to improve the spectroscopic parameters and thus the line list. Observations performed at 3 mm using the IRAM 30 m telescope (improved with respect to those reported in Cazzoli et al., 2016)

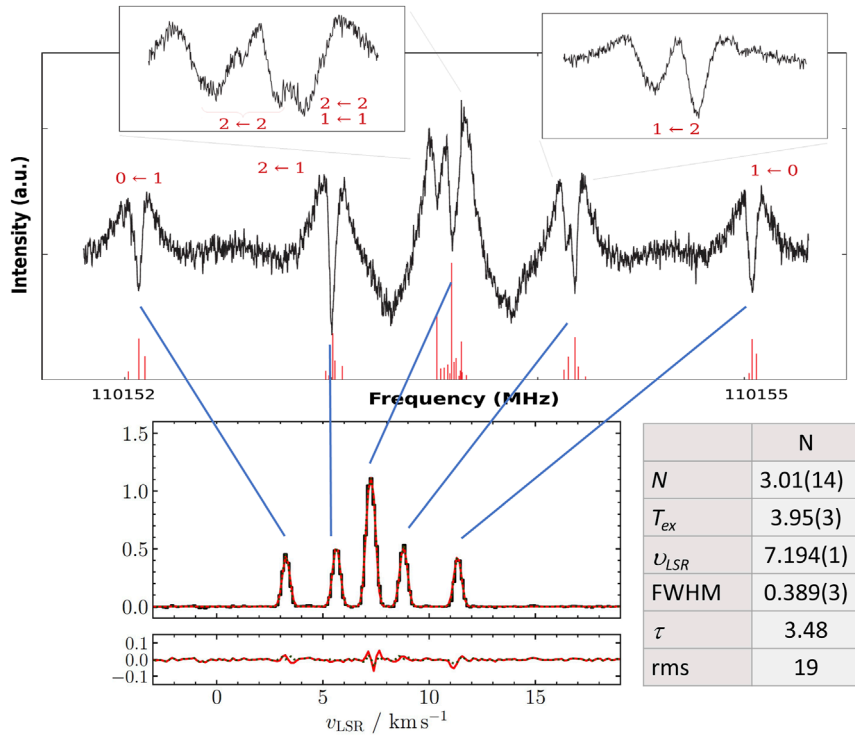
were used to complement the Yebes data (Marcelino, N. et al., 2023). In all cases, the HSO column densities were derived using LTE models able to reproduce the observed spectra. The retrieved N values range between 7.0×10^{10} cm^{-2} and 2.9×10^{11} cm^{-2}, which result in abundances in the (1.4–7.0) $\times 10^{-12}$ range relative to H_2.

3.4 Deuterated ammonia: the role of hyperfine interactions

Accurate determination of molecular abundances of astrochemical species is a crucial step towards the understanding of the chemistry occurring in the ISM. Even when LTE approximation holds (or provides reasonable results), this process requires a detailed and accurate knowledge of the molecular energy levels. If electric and/or magnetic interactions take place, then this knowledge should be extended to the hyperfine splittings of the rotational energy levels. As mentioned above, the main interaction to account for is the nuclear quadrupole coupling. In this respect, it has to be noted that, even if the effect on the radioastronomical spectrum is not evident, a hyperfine structure may cause a distortion of the line profile that, if not taken into account, can lead to biased results in terms of physical parameters and molecular abundance. A significant example is provided by incorporating or not the effect of the deuterium quadrupole coupling in the analysis of astrophysical emissions of NH_2D at millimeter wavelengths in the recent observations of the starless core L1544, a low-mass star-forming core in a very early stage of evolution (Melosso et al., 2021).

The rotational spectrum of NH_2D presents a hyperfine structure which is mainly due to the nitrogen quadrupole coupling, which splits the rotational transitions in hyperfine components that are resolved also at Doppler-limited resolution (see Figure 8). However, deuterium is a quadrupolar nucleus as well, thus contributing to the overall hyperfine structure. Being the interaction due to D weaker than that due to N, the overall result is that the hyperfine components further split because of this additional coupling, as evident in Figure 8 (top panel). Though, the deuterium hyperfine structure can be resolved only by exploiting sub-Doppler resolution (see Figure 8; Melosso et al., 2020b).

In Melosso et al. (2021), the importance of including the deuterium hyperfine structure in the analysis of astrophysical emissions of NH_2D at millimeter wavelengths has been pointed out for the recent observations of the starless core L1544. This source is a prototypical cold, quiescent core on the verge of the gravitational collapse exhibiting very narrow line emissions (Tafalla et al., 2002; Keto and Caselli, 2010). The rotational features considered were the *ortho*-NH_2D $J_{K_a,K_c} = 1_{1,1} - 1_{0,1}$ transition at 85926.3 MHz and



	N	N+D
N	3.01(14)	2.44(9)
T_{ex}	3.95(3)	4.07(2)
υ_{LSR}	7.194(1)	7.195(1)
FWHM	0.389(3)	0.346(2)
τ	3.48	3.46
rms	19	11

FIGURE 8
Top panel: Lamb-dip spectrum of the $J_{K_a,K_c} = 1_{1,1} \leftarrow 1_{0,1}$ transition of *para*-NH_2D showing the nitrogen hyperfine structure already resolved at Doppler-limited resolution. The deuterium hyperfine structure is resolved only by exploiting the Lamb-dip technique and is better shown in the insets. The numbers in red refer to the hyperfine components due to nitrogen quadrupole coupling. Red sticks represent the position and the intensity of each transition as predicted by Melosso et al. (2020b). Bottom panel: Analysis of corresponding emissions in L1544. The dotted green trace plots the model computed considering the hyperfine structure due to both N and D (N + D). The solid red trace plots the model accounting only for the N quadrupole coupling. The residuals of both models are plotted using the same color code. On the right side, the parameters resulting from the two models are reported.

the *para*-NH_2D $J_{K_aK_c} = 1_{1,1} - 1_{0,1}$ transition at 110153.6 MHz. The resulting spectra are shown in Melosso et al. (2021), while in Figure 8 only the *para* transition is shown together with the Lamb-dip spectrum recorded in the laboratory (Melosso et al., 2020b). As evident in Figure 8, the astronomical spectrum has been analyzed by considering the hyperfine structure due to nitrogen only (N; solid red lines plot) and the full hyperfine structure due to both nitrogen and deuterium (N + D; dotted green trace plot). The parameters of the fit (which are reported in Figure 8 as well) are the column density (N), the excitation temperature (T_{ex}), the systemic velocity (v_{LSR}) and the line full width at half maximum (FWHM), while the total opacity of the transition (τ) is regarded as a derived quantity. While the two fits are almost indistinguishable by visual inspection, small but significant differences are evident in the fit results. In addition to a 30%–40% reduction of the residual root-mean-square deviation of the fit, the proper treatment of the hyperfine effects (i.e., incorporation of the deuterium quadrupole coupling) determines a reduction of the derived line FWHM by about 12%. This change is then reflected in the N and T_{ex} values. For the *para*-NH_2D emission shown in Figure 8, the column density decreases by about 19% and the excitation temperature increases by about 3%, while τ remains substantially unchanged.

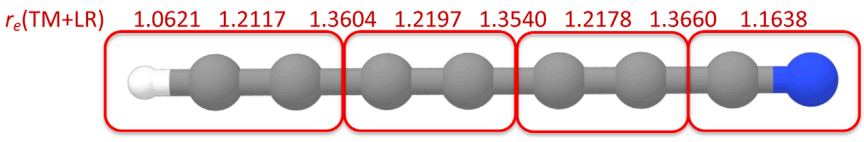
In Melosso et al. (2020b) and Melosso et al. (2021), the rotational spectrum of the singly-deuterated ammonia was recorded in the 265–1,565 GHz range, and the Lamb-dip technique exploited in the entire frequency region considered. NH_2D was produced *in situ* by flowing a small amount of NH_3 (less than 0.5 mTorr) into the absorption cell where deuterium (D_2) was previously discharged for some minutes. A frequency modulation of 1 kHz was used and the modulation-depth varied between 8 and 24 kHz according to the frequency region. To analyze the recorded spectra, the hyperfine structure was predicted using the computed hyperfine constants together with the other spectroscopic parameters (rotational and centrifugal distortion constants) from Cohen and Pickett (1982) and Fusina et al. (1988). In Melosso et al. (2020b), the hyperfine parameters were calculated at the CCSD(T) level correlating all electrons, extrapolating to the CBS limit, and incorporating the fT and fQ corrections as well as those due to vibrational effects. Such a level of theory led to parameters in quantitative agreement with experiment, thus allowing an easy assignment of the complicated hyperfine structure to the correct hyperfine components.

3.5 Toward molecular complexity

3.5.1 DC_7N: a long carbon chain

From the point of view of the terrestrial chemistry, carbon chains are definitely exotic species. In the ISM, instead, after the discovery of the first carbon-chain molecules in the 1970s (Avery et al., 1976; Broten et al., 1978; Guelin et al., 1978; Kroto et al., 1978), this type of systems has been largely detected (McGuire, 2022). In fact, they represent about 40% of all detected species (Burkhardt et al., 2017) and have been proposed as precursors of PAHs (Guzman-Ramirez et al., 2011). In Burkhardt et al. (2017), the first interstellar detection towards the dark cloud TMC-1 of DC_7N, together with six ^{13}C-bearing isotopologues of HC_7N, was reported. In that study, for DC_7N, a column density of 2.5(9) $\times 10^{11}$ cm^{-2} was derived.

In McCarthy et al. (2000), the rotational spectra of all single-substituted isotopologues of three long linear cyanopolyynes, namely, HC_7N, HC_9N, and $HC_{11}N$, have been investigated at low frequency (6–18 GHz). While, in that work, the isotopic species of HC_7N (as well as HC_9N and $HC_{11}N$) were produced by high-voltage low-current discharge, in Puzzarini et al. (2023), DC_7N was produced by FVP, co-pyrolizing fully-deuterated toluene (toluene-d_8) and trichloroacetonitrile (Cl_3CCN) in a 1:20 ratio at 1,200 °C, and measurements were extended in the millimeter-wave region (80–90 GHz). The ground-state rotational constant retrieved in McCarthy et al. (2000) and that derived in Puzzarini et al. (2023) from a global fit are collected in Figure 9, while the comparison of the full set of spectroscopic parameters can be found in Puzzarini et al. (2023). Here, we limit ourselves to note that, as pointed out in Puzzarini et al. (2023), incorporation of millimeter-wave measurements in the fit led to a relevant improvement in the accuracy of the line catalog, with uncertainties reducing from 0.9 to 1 MHz to about 10–20 kHz. Furthermore, above 50 GHz, transition frequencies were found to vary up to ~500 kHz.



	revDSD	TM	TM+LR	McCarthy et al. (2000)	Puzzarini et al. (2023)
B_e	545.67	543.92	544.90	-	-
ΔB_{vib}	+0.53	0.53	0.53	-	-
B_0	546.20	544.45	545.43	545.31523(7)	545.315200(11)
	+0.17%	-0.15%	+0.02%		

FIGURE 9
Molecular structure of HC_7N: the TM+LR equilibrium parameters are reported. In the table, the equilibrium rotational constants at different level of theory are collected, and the corresponding ground-state rotational constants are compared with experiment (McCarthy et al., 2000; Puzzarini et al., 2023).

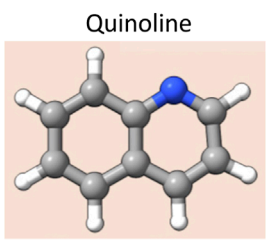
As shown in Figure 9, DC_7N is an illustrative example to demonstrate the accuracy that can be obtained with the TM+LR approach and that increasing the number of "Lego bricks" does not affect the effectiveness of the TM+LR approach. The corresponding equilibrium structure is reported in Figure 9 together with the associated equilibrium rotational constant. To compare with the experiment, the vibrational correction (evaluated at the fc-MP2/cc-pVDZ level) has been incorporated, thus providing the theoretical estimate of the ground-state rotational constant. The comparison of Figure 9 points out that the relative deviation from experiment is +0.17% at the revDSD-D3BJ/jun-cc-pVTZ level and −0.15% for the TM structure. When resorting to the TM+LR approach, the discrepancy reduces by one order of magnitude, i.e., to +0.02%. Such an accuracy can be obtained only by exploiting high-level CC composite schemes. Therefore, TM+LR is a very powerful tool for accurately predicting rotational constants for systems of increasing complexity, as indeed demonstrated in Ye et al. (2022).

3.5.2 Toward PAHs and N-bearing derivatives

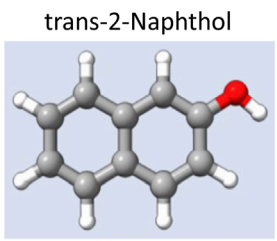
PAHs and their N-bearing derivatives are a prominent class of aromatics. Their elemental unit is benzene, which is however transparent to microwave radiation because of the lack of any permanent dipole moment. Consequently, its substituted species can be considered good proxies to trace its presence and abundance. The same applies to fused benzene rings such as naftalene and antracene. Among the possible substituents, the cyano (–CN) and ethynyl (–CCH) groups introduce good dipole moment and have led to the detection of benzonitrile (McGuire et al., 2018), cyanonaphtalenes (McGuire et al., 2021), ethynyl cyclopropenylidene (Cernicharo et al., 2021a), ethynyl cyclopentadienes (Cernicharo et al., 2021b), and cyanoindenes (Sita et al., 2022). The limiting factor to further interstellar identifications is the availability of accurate line lists obtained from high-resolution laboratory investigations. Even when production of the molecular species is not a challenge, the recorded rotational spectra are often characterized by a high density of the spectral lines. In such cases, their analysis becomes the limiting step to derive accurate catalogs. In this respect, accurate predictions is a key into simplifying and speeding up the spectral assignments. However, accuracy becomes an issue when dealing with large systems, with the TM+LR approach offering an excellent way-out.

In Ye et al. (2022), a set of small PAHs and polycyclic aromatic nitrogen heterocycles (PANHs) were considered to test the accuracy obtainable with the TM/TM+LR approach. The species considered ranged from one to four fused aromatic rings, with nitrogen being either a member of the ring or part of the substituents. A few molecules containing oxygen in the functional groups were also part of the set. The benchmark study of Ye et al. (2022) demonstrated that the TM+LR approach is able to provide, at a very limited computational cost, accurate rotational constants, with deviations from experiment as small as 0.01%–0.03% and in all cases well within 0.1%. A selection of results is shown in Figure 10: in order to compare theoretical and experimental data, the equilibrium rotational constants were augmented by vibrational corrections computed at the B3LYP-D3BJ/jun-cc-pVDZ level. In Ye et al. (2022), after the benchmark study, the TM/TM+LR approach was then applied to a set of 13 small PA(N)Hs not yet experimentally investigated (such as ethynylphenanthrene and pyrenecarbonitrile), thus providing accurate predictions for their rotational constants.

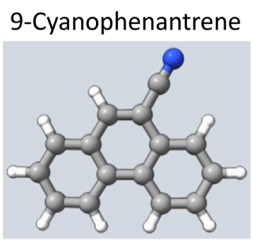
A first application of the TM+LR model to systems of astrophysical interest in order to support their experimental characterization has been reported in Spaniol et al. (2023), where the isomers of cyanoethynylbenzene (*ortho*–, *meta*– and *para* – HCC–C_6H_4–CN) have been considered. All of them are strong candidates for interstellar identification in view of the fact that benzonitrile (McGuire et al., 2018) and phenylacetylene (tentative; Cernicharo et al., 2021b) have already been detected, and both –CN and –CCH are polar and common functional groups. Indeed, the CN and CCH radicals are abundant and the gas-phase reaction of the CN and CCH radicals with benzene is the most accepted formation route for benzonitrile and phenylacetylene,



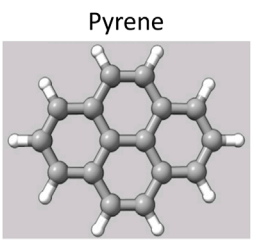
	TM	%Diff.
A_0	3144.74	-0.03
B_0	1271.62	+0.00
C_0	905.64	-0.01



	TM+LR	%Diff.
A_0	2844.76	-0.02
B_0	825.33	-0.03
C_0	639.90	-0.03



	TM+LR	%Diff.
A_0	985.34	-0.05
B_0	451.17	-0.01
C_0	309.54	-0.03



	TM	%Diff.
A_0	1010.59	-0.05
B_0	556.39	+0.02
C_0	358.93	-0.01

FIGURE 10
Selected examples from the benchmark set employed in Ye et al. (2022): for each system, the computed ground-state rotational constants and the relative deviation from experiment are reported. The "Lego bricks" used are: benzene and pyridine for quinoline, benzene and water for trans-2-naphthol, benzene and HCN for 9-cyanophenantrene, benzene for pyrene.

respectively. For *ortho*–, *meta*– and *para* – HCC–C_6H_4–CN, the same methodology mentioned above (TM+LR approach for B_e and the ΔB_{vib} correction at the B3LYP-D3BJ/jun-cc-pVDZ level) was used not only for predicting the ground-state rotational constants, but also for the rotational constants of vibrational excited states. Another crucial information provided by theory was the prediction of the quartic centrifugal distortion constants (at the fc-MP2/cc-pVTZ level). In line with Ye et al. (2022), the computed B_0 values were found to deviate from experiment, on average, by 0.026%. The discrepancy increased to 2.5% for quartic centrifugal terms (Spaniol et al., 2023). The experimental spectroscopic data obtained by Spaniol et al. (2023) will enable astronomical searches for these molecules.

Another recent example where the TM+LR predictions from Ye et al. (2022) were employed to guide the laboratory measurements, is provided by ethynylnaphthalene (Cabezas et al., 2022b). In Cabezas et al. (2022b), both 1- and 2-ethynylnaphthalene were considered, and the average deviation of TM+LR predictions was found to be 0.04%, thus in line with the outcomes of Ye et al. (2022).

4 Conclusion

In this review, a methodology to be exploited for the discovery of new molecules in the ISM has been presented. The steps (summarized in Figure 1) required to move from the laboratory spectroscopic characterization to the detection in space, are detailed. Different illustrative applications, ranging from small prebiotic molecules to radicals, to deuterated species and to larger systems such as long carbon chains and small PAHs, have been discussed.

In hunting for new interstellar molecules, the choice of the species to be investigated is driven by different reasons such as reaction mechanisms that suggest its formation in the ISM or astrochemical networks/theories that assume its presence in space. Rare isotopologues are also interesting targets because they can provide information on different physical properties such as isotopic ratios and fractionation. Then, the spectroscopic characterization (in the field of rotational spectroscopy) relies on a strong interplay of experiment and theory, which is crucial because exotic molecules are often not or poorly studied and produced using techniques, such as FVP and DC discharge, that lead to the concomitant formation of several other species, thus making the spectra crowded and difficult to be analyzed. An extensive spectral investigation combined with an accurate analysis then allows for obtaining a precise spectroscopic line catalog, which is used for astronomical searches. The detection in space of these exotic molecules, which have usually low abundances (especially when increasing molecular complexity) and hence low line intensities, requires unbiased (covering large bandwidths, to maximize the number of targeted transitions of interest) and ultra-sensitive spectral surveys (to increase sensitivity), like the ones obtained towards G+0.693−0.027 using Yebes 40 m and IRAM 30 m. As a part of the detection, the abundance of the new species is retrieved. This is usually based on assuming that the LTE approximation holds. However, a more accurate estimate is obtained by accounting for collisions competing with radiative processes. This however requires a major effort from the computational and/or experimental point of view, with the so-called COMs (complex organic molecules; Herbst and van Dishoeck, 2009) being an additional challenge for theoreticians and experimentalists because of the huge number of closely spaced rotational energy levels and the large number of internal coordinates.

While our strategy has been rationalized, it should be noted that each step is not at all straightforward but is instead hampered by challenges. The computational characterization suffers for the system dimension because of the unfavourable scaling of quantum-chemical models with the molecular size (even if an effective methodology has been proposed), but also for the possible presence of large amplitude motions, not to mention the theoretical difficulties when dealing with open-shell species. From the experimental point of view, envisioning suitable precursors and/or their synthesis are major difficulties together with finding efficient working conditions. The spectral analysis is also not devoid of difficulties because standard Hamiltonians might be not appropriate for the spectroscopic description of the system under investigation because of, for example, unexpected interactions. The detection of a new molecule is hampered not only by low line intensities (as stressed above), but also by the 'forest' of lines due to other species that makes it difficult to find unblended emissions. Furthermore, molecules with small dipole moments (and thus weak emissions) are hardly observable even if abundant, while the apolar ones require suitable proxies (Alessandrini et al., 2023). Finally, the derivation of accurate molecular abundances can be hampered by relevant deviation from the LTE model.

The last note concerns interstellar molecules adsorbed on or locked inside dust grains. While about 300 gas-phase molecules have been detected in the interstellar medium or circumstellar shells (Woon, 2004; Müller et al., 2005; McGuire, 2022), far more limited is the knowledge about the chemical composition of the ice covering dust grains. However, it is expected that, in the near future, the JWST will significantly improve our current knowledge (McClure et al., 2023).

Author contributions

CP contributed on conceptualization, project administration, methodology, data curation, validation, writing–original draft, writing–review and editing. CP and VR contributed on funding acquisition. SA, LB, MM, and VR contributed on methodology, investigation, data curation, formal analysis, validation, and writing–review and editing. All authors contributed to the article and approved the submitted version.

Funding

This work has been supported by MUR (PRIN Grant Number 202082CE3T) and by the University of Bologna (RFO funds). VR

has received support from the project RYC2020-029387-I funded by MCIN/AEI/10.13039/501100011033.

Acknowledgments

The COST Action CA21101 "COSY–Confined molecular systems: from a new generation of materials to the stars" is acknowledged.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alberton, D., Bizzocchi, L., Jiang, N., Melosso, M., Rivilla, V. M., Charmet, A. P., et al. (2023a). Laboratory spectroscopy of allylimine and tentative detection towards the G+0.693-0.027 molecular cloud. *Astron. Astrophys.* 669, A93. doi:10.1051/0004-6361/202244618

Alberton, D., Lattanzi, V., Endres, C., Rivilla, V. M., Guillemin, J. C., Caselli, P., et al. (2023b). A high-resolution spectroscopic analysis of aminoacrylonitrile and an interstellar search toward G+0.693. *Astrophys. J.* 951, 108. doi:10.3847/1538-4357/acd8bd

Alessandrini, S., Bizzocchi, L., Melosso, M., and Puzzarini, C. (2023). Protonation of apolar species: from Cl_2H^+ to (E)-NCCHCHCNH$^+$ through computational investigations. *Front. Astron. Space Sci.* 10, 1128896. doi:10.3389/fspas.2023.1128896

Alessandrini, S., Gauss, J., and Puzzarini, C. (2018). Accuracy of rotational parameters predicted by high-level quantum-chemical calculations: case study of sulfur-containing molecules of astrochemical interest. *J. Chem. Theory Comput.* 14, 5360–5371. doi:10.1021/acs.jctc.8b00695

Aliev, M. R., and Watson, J. K. (1976). Calculated sextic centrifugal distortion constants of polyatomic molecules. *J. Mol. Spectrosc.* 61, 29–52. doi:10.1016/0022-2852(76)90379-9

Aliev, M. R., and Watson, J. K. G. (1985). "Chapter 1 - higher-order effects in the vibration–rotation spectra of semirigid molecules," in *Molecular spectroscopy: Modern research*. Editor K. N. Rao (New York: Academic Press), 1–67. doi:10.1016/B978-0-12-580643-5.50006-3

Avery, L. W., Broten, N. W., MacLeod, J. M., Oka, T., and Kroto, H. W. (1976). Detection of the heavy interstellar molecule cyanodiacetylene. *Astrophys. J. Lett.* 205, L173–L175. doi:10.1086/182117

Barone, V., Alessandrini, S., Biczysko, M., Cheeseman, J. R., Clary, D. C., McCoy, A. B., et al. (2021). Computational molecular spectroscopy. *Nat. Rev. Methods Prim.* 1, 38. doi:10.1038/s43586-021-00034-1

Belloche, A., Meshcheryakov, A. A., Garrod, R. T., Ilyushin, V. V., Alekseev, E. A., Motiyenko, R. A., et al. (2017). Rotational spectroscopy, tentative interstellar detection, and chemical modeling of N-methylformamide. *Astron. Astrophys.* 601, A49. doi:10.1051/0004-6361/201629724

Bermúdez, C., Tercero, B., Motiyenko, R. A., Margulès, L., Cernicharo, J., Ellinger, Y., et al. (2018). The millimeter-wave spectrum of methyl ketene and the astronomical search for it. *Astron. Astrophys.* 619, A92. doi:10.1051/0004-6361/201833267

Bizzocchi, L., Alessandrini, S., Melosso, M., Rivilla, V. M., and Puzzarini, C. (2022). *Ab initio* study of fine and hyperfine interactions in triplet POH. *Molecules* 27, 302. doi:10.3390/molecules27010302

Bizzocchi, L., Prudenzano, D., Rivilla, V. M., Pietropolli-Charmet, A., Giuliano, B. M., Caselli, P., et al. (2020). Propargylimine in the laboratory and in space: millimetre-wave spectroscopy and its first detection in the ism. *Astron. Astrophys.* 640, A98. doi:10.1051/0004-6361/202038083

Broten, N. W., Oka, T., Avery, L. W., MacLeod, J. M., and Kroto, H. W. (1978). The detection of HC_9N in interstellar space. *Astrophys. J. Lett.* 223, L105–L107. doi:10.1086/182739

Brown, R. D., Godfrey, P. D., and Winkler, D. A. (1981). Hyperfine interactions in the microwave spectrum of 2-propen-1-Imine (vinylimine). *Chem. Phys.* 59, 243–247. doi:10.1016/0301-0104(81)85167-1

Burkhardt, A. M., Herbst, E., Kalenskii, S. V., McCarthy, M. C., Remijan, A. J., and McGuire, B. A. (2017). Detection of HC_5N and HC_7N isotopologues in TMC-1 with the green bank telescope. *Mon. Not. R. Astron. Soc.* 474, 5068–5075. doi:10.1093/mnras/stx2972

Cabezas, C., Agúndez, M., Marcelino, N., Tercero, B., Endo, Y., Fuentetaja, R., et al. (2022a). Discovery of the elusive thioketenylium, $HCCS^+$, in TMC-1. *Astron. Astrophys.* 657, L4. doi:10.1051/0004-6361/202142815

Cabezas, C., Peña, I., and Cernicharo, J. (2022b). Laboratory rotational spectroscopy and astronomical search of ethynyl substituted naphthalene. *Mon. Not. R. Astron. Soc.* 519, 2590–2597. doi:10.1093/mnras/stac3698

Cassone, G., Saija, F., Sponer, J., Sponer, J. E., Ferus, M., Krus, M., et al. (2018). Dust motions in magnetized turbulence: source of chemical complexity. *Astrophys. J. Lett.* 866, L23. doi:10.3847/2041-8213/aae529

Cazzoli, G., Cludi, L., Buffa, G., and Puzzarini, C. (2012). Precise thz measurements of HCO^+, N_2H^+, and CF^+ for astrophysical observations. *Astrophys. J. Suppl. Ser.* 203, 11. doi:10.1088/0067-0049/203/1/11

Cazzoli, G., Lattanzi, V., Alonso, J. L., Gauss, J., and Puzzarini, C. (2015). The hyperfine structure of the rotational spectrum of HDO and its extension to the THz region: accurate rest frequencies and spectroscopic parameters for astrophysical observations. *Astrophys. J.* 806, 100. doi:10.1088/0004-637X/806/1/100

Cazzoli, G., Lattanzi, V., Kirsch, T., Gauss, J., Tercero, B., Cernicharo, J., et al. (2016). Laboratory measurements and astronomical search for the HSO radical. *Astron. Astrophys.* 591, A126. doi:10.1051/0004-6361/201628745

Cazzoli, G., Puzzarini, C., Harding, M. E., and Gauss, J. (2009). The hyperfine structure in the rotational spectrum of water: lamb-dip technique and quantum-chemical calculations. *Chem. Phys. Lett.* 473, 21–25. doi:10.1016/j.cplett.2009.03.045

Cazzoli, G., Puzzarini, C., and Lapinov, A. V. (2004). Precise laboratory frequencies for the $J \leftarrow J-1$ (J=1, 2, 3, 4) rotational transitions of ^{13}CO. *Astrophys. J.* 611, 615–620. doi:10.1086/421992

Cernicharo, J., Agúndez, M., Cabezas, C., Tercero, B., Marcelino, N., Pardo, J. R., et al. (2021a). Pure hydrocarbon cycles in TMC-1: discovery of ethynyl cyclopropenylidene, cyclopentadiene, and indene. *Astron. Astrophys.* 649, L15. doi:10.1051/0004-6361/202141156

Cernicharo, J., Agúndez, M., Kaiser, R. I., Cabezas, C., Tercero, B., Marcelino, N., et al. (2021b). Discovery of two isomers of ethynyl cyclopentadiene in TMC-1: abundances of CCH and CN derivatives of hydrocarbon cycles. *Astron. Astrophys.* 655, L1. doi:10.1051/0004-6361/202142226

Cernicharo, J., Cabezas, C., Pardo, J. R., Agúndez, M., Roncero, O., Tercero, B., et al. (2023). The magnesium paradigm in IRC+10216: discovery of MgC_4H^+, MgC_3N^+, MgC_6H^+, and MgC_5N^+. *Astron. Astrophys.* 672, L13. doi:10.1051/0004-6361/202346467

Ceselin, G., Barone, V., and Tasinato, N. (2021). Accurate biomolecular structures by the nano-LEGO approach: pick the bricks and build your geometry. *J. Chem. Theory Comput.* 17, 7290–7311. doi:10.1021/acs.jctc.1c00788

Changala, P. B., Gupta, H., Cernicharo, J., Pardo, J. R., Agúndez, M., Cabezas, C., et al. (2022). Laboratory and astronomical discovery of magnesium dicarbide, MgC_2. *Astrophys. J. Lett.* 940, L42. doi:10.3847/2041-8213/aca144

Cheng, Q., Fortenberry, R. C., and DeYonker, N. J. (2017). Towards a quantum chemical protocol for the prediction of rovibrational spectroscopic data for transition metal molecules: exploration of CuCN, CuOH, and CuCCH. *J. Chem. Phys.* 147, 234303. doi:10.1063/1.5006931

Christen, D., Coudert, L., Suenram, R., and Lovas, F. (1995). The rotational/concerted torsional spectrum of the *g′Ga* conformer of ethylene glycol. *J. Mol. Spectrosc.* 172, 57–77. doi:10.1006/jmsp.1995.1155

Cohen, E., and Pickett, H. (1982). The rotation-inversion spectra and vibration-rotation interaction in NH_2D. *J. Mol. Spectrosc.* 93, 83–100. doi:10.1016/0022-2852(82)90276-4

Császár, A. G., Allen, W. D., and Schaefer, H. F. (1998). In pursuit of the *ab initio* limit for conformational energy prototypes. *J. Chem. Phys.* 108, 9751–9764. doi:10.1063/1.476449

Demaison, J., Boggs, J. E., and Császár, A. G. (2016). *Equilibrium molecular structures: From spectroscopy to quantum chemistry*. Boca Raton: CRC Press. doi:10.1201/b10374

Dunning, T. H. (1989). Gaussian basis sets for use in correlated molecular calculations. I. the atoms boron through neon and hydrogen. *J. Chem. Phys.* 90, 1007–1023. doi:10.1063/1.456153

Endo, Y., Saito, S., and Hirota, E. (1981). Microwave spectra of the HSO and DSO radicals. *J. Chem. Phys.* 75, 4379–4384. doi:10.1063/1.442600

Feller, D. (1993). The use of systematic sequences of wave functions for estimating the complete basis set, full configuration interaction limit in water. *J. Chem. Phys.* 98, 7059–7071. doi:10.1063/1.464749

Fusina, L., Di Lonardo, G., Johns, J., and Halonen, L. (1988). Far-infrared spectra and spectroscopic parameters of NH_2D and ND_2H in the ground state. *J. Mol. Spectrosc.* 127, 240–254. doi:10.1016/0022-2852(88)90023-9

García de la Concepción, J., Puzzarini, C., Barone, V., Jiménez-Serra, I., and Roncero, O. (2021). Formation of phosphorus monoxide (PO) in the interstellar medium: insights from quantum-chemical and kinetic calculations. *Astrophys. J.* 922, 169. doi:10.3847/1538-4357/ac1e94

Goldsmith, P. F., and Langer, W. D. (1999). Population diagram analysis of molecular line emission. *Astrophys. J.* 517, 209–225. doi:10.1086/307195

Gordy, W., and Cook, R. L. (1984). *Microwave molecular spectra*. New York: Wiley.

Grimme, S., Antony, J., Ehrlich, S., and Krieg, H. (2010). A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* 132, 154104. doi:10.1063/1.3382344

Grimme, S., Ehrlich, S., and Goerigk, L. (2011). Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* 32, 1456–1465. doi:10.1002/jcc.21759

Grimme, S. (2006). Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* 124, 034108. doi:10.1063/1.2148954

Guelin, M., Green, S., and Thaddeus, P. (1978). Detection of the C_4H radical toward IRC +10216. *Astrophys. J. Lett.* 224, L27–L30. doi:10.1086/182751

Guzmán, V., Roueff, E., Gauss, J., Pety, J., Gratier, P., Goicoechea, J. R., et al. (2012). The hyperfine structure in the rotational spectrum of CF^+. *Astron. Astrophys.* 548, A94. doi:10.1051/0004-6361/201220174

Guzman-Ramirez, L., Zijlstra, A. A., NíChuimín, R., Gesicki, K., Lagadec, E., Millar, T. J., et al. (2011). Carbon chemistry in Galactic bulge planetary nebulae. *Mon. Not. R. Astron. Soc.* 414, 1667–1678. doi:10.1111/j.1365-2966.2011.18502.x

Hays, B. M., Gupta, D., Guillaume, T., Abdelkader Khedaoui, O., Cooke, I. R., Thibault, F., et al. (2022). Collisional excitation of HNC by He found to be stronger than for structural isomer HCN in experiments at the low temperatures of interstellar space. *Nat. Chem.* 14, 811–815. doi:10.1038/s41557-022-00936-x

Heckert, M., Kállay, M., and Gauss, J. (2005). Molecular equilibrium geometries based on coupled-cluster calculations including quadruple excitations. *Mol. Phys.* 103, 2109–2115. doi:10.1080/00268970500083416

Heckert, M., Kállay, M., Tew, D. P., Klopper, W., and Gauss, J. (2006). Basis-set extrapolation techniques for the accurate calculation of molecular equilibrium geometries using coupled-cluster theory. *J. Chem. Phys.* 125, 044108. doi:10.1063/1.2217732

Helgaker, T., Klopper, W., Koch, H., and Noga, J. (1997). Basis-set convergence of correlated calculations on water. *J. Chem. Phys.* 106, 9639–9646. doi:10.1063/1.473863

Herbst, E. (2021). Unusual chemical processes in interstellar chemistry: past and present. *Front. Astron. Space Sci.* 8, 776942. doi:10.3389/fspas.2021.776942

Herbst, E., and van Dishoeck, E. F. (2009). Complex organic interstellar molecules. *Annu. Rev. Astron. Astrophys.* 47, 427–480. doi:10.1146/annurev-astro-082708-101654

Huang, X., and Lee, T. J. (2009). Accurate *ab initio* quartic force fields for and CCH^- and rovibrational spectroscopic constants for their isotopologs. *J. Chem. Phys.* 131, 104301. doi:10.1063/1.3212560

James, P. L., Sims, I. R., Smith, I. W. M., Alexander, M. H., and Yang, M. (1998). A combined experimental and theoretical study of rotational energy transfer in collisions between NO ($X^2\Pi_{1/2}$, v=3,*J*) and He, Ar and N_2 at temperatures down to 7 K. *J. Chem. Phys.* 109, 3882–3897. doi:10.1063/1.476517

Jiang, N., Melosso, M., Tamassia, F., Bizzocchi, L., Dore, L., Canè, E., et al. (2021). High-resolution infrared spectroscopy of DC_3N in the stretching region. *Front. Astron. Space Sci.* 8, 29. doi:10.3389/fspas.2021.656295

Karton, A., and Martin, J. M. L. (2012). Explicitly correlated wn theory: W1-F12 and W2-F12. *J. Chem. Phys.* 136, 124114. doi:10.1063/1.3697678

Karton, A., Rabinovich, E., Martin, J. M. L., and Ruscic, B. (2006). W4 theory for computational thermochemistry: in pursuit of confident sub-kJ/mol predictions. *J. Chem. Phys.* 125, 144108. doi:10.1063/1.2348881

Karton, A., and Talbi, D. (2014). Pinning the most stable $H_xC_yO_z$ isomers in space by means of high-level theoretical procedures. *Chem. Phys.* 436-437, 22–28. doi:10.1016/j.chemphys.2014.03.010

Keto, E., and Caselli, P. (2010). Dynamics and depletion in thermally supercritical starless cores. *Mon. Not. R. Astron. Soc.* 402, 1625–1634. doi:10.1111/j.1365-2966.2009.16033.x

Kłos, J., and Lique, F. (2017). "Cold molecular collisions: quantum scattering calculations and their relevance in astrophysical applications," in *Cold chemistry: Molecular scattering and reactivity near absolute zero* (RSC Theoretical and Computational Chemistry Series). doi:10.1039/9781782626800-00046

Koelemay, L. A., Burton, M. A., Singh, A. P., Sheridan, P. M., Bernal, J. J., and Ziurys, L. M. (2022). Laboratory and astronomical detection of the SiP radical ($X^2\Pi_i$): more circumstellar phosphorus. *Astrophys. J. Lett.* 940, L11. doi:10.3847/2041-8213/ac9d9b

Kroto, H. W., Kirby, C., Walton, D. R. M., Avery, L. W., Broten, N. W., MacLeod, J. M., et al. (1978). The detection of cyanohexatriyne, $H(C{\equiv}C)_3CN$, in Heile's Cloud 2. *Astrophys. J. Lett.* 219, L133–L137. doi:10.1086/182623

Lee, K. L. K., Loomis, R. A., Burkhardt, A. M., Cooke, I. R., Xue, C., Siebert, M. A., et al. (2021). Discovery of interstellar trans-cyanovinylacetylene (HC ≡ CCH = CHC ≡ N) and vinylcyanoacetylene ($H_2C = CHC_3N$) in GOTHAM observations of TMC-1. *Astrophys. J. Lett.* 908, L11. doi:10.3847/2041-8213/abdbb9

Lupi, J., Alessandrini, S., Barone, V., and Puzzarini, C. (2021). junChS and junChS-F12 models: parameter-free efficient yet accurate composite schemes for energies and structures of noncovalent complexes. *J. Chem. Theory Comput.* 17, 6974–6992. doi:10.1021/acs.jctc.1c00869

Marcelino, N., Puzzarini, C., Agúndez, M., Fuentetaja, R., Tercero, B., de Vicente, P., et al. (2023). First detection of the HSO radical in space. *Astron. Astrophys.* 674, L13. doi:10.1051/0004-6361/202346935

Martín, S., Martín-Pintado, J., Blanco-Sánchez, C., Rivilla, V. M., Rodríguez-Franco, A., and Rico-Villas, F. (2019). Spectral line identification and modelling (SLIM) in the MAdrid data CUBe analysis (MADCUBA) package. Interactive software for data cube analysis. *Astron. Astrophys.* 631, A159. doi:10.1051/0004-6361/201936144

Martín, S., Requena-Torres, M. A., Martín-Pintado, J., and Mauersberger, R. (2008). Tracing shocks and photodissociation in the galactic center region. *Astrophys. J.* 678, 245–254. doi:10.1086/533409

Matthews, D. A., Cheng, L., Harding, M. E., Lipparini, F., Stopkowicz, S., Jagau, T. C., et al. (2020). Coupled-cluster techniques for computational chemistry: the CFOUR program package. *J. Chem. Phys.* 152, 214108. doi:10.1063/5.0004837

McCarthy, M. C., Lee, K. L. K., Loomis, R. A., Burkhardt, A. M., Shingledecker, C. N., Charnley, S. B., et al. (2021). Interstellar detection of the highly polar five-membered ring cyanocyclopentadiene. *Nat. Astron* 5, 176–180. doi:10.1038/s41550-020-01213-y

McCarthy, M. C., Levine, E. S., Apponi, A. J., and Thaddeus, P. (2000). Experimental structures of the carbon chains HC_7N, HC_9N, and $HC_{11}N$ by isotopic substitution. *J. Mol. Spectrosc.* 203, 75–81. doi:10.1006/jmsp.2000.8149

McClure, M. K., Rocha, W. R. M., Pontoppidan, K. M., Crouzet, N., Chu, L. E. U., Dartois, E., et al. (2023). An Ice Age JWST inventory of dense molecular cloud ices. *Nat. Astron.* 7, 431–443. doi:10.1038/s41550-022-01875-w

McGuire, B. A. (2018). 2018 census of interstellar, circumstellar, extragalactic, protoplanetary disk, and exoplanetary molecules. *Astrophys. J. Suppl. Ser.* 239, 17. doi:10.3847/1538-4365/aae5d2

McGuire, B. A. (2022). 2021 census of interstellar, circumstellar, extragalactic, protoplanetary disk, and exoplanetary molecules. *Astrophys. J. Suppl. Ser.* 259, 30. doi:10.3847/1538-4365/ac2a48

McGuire, B. A., Burkhardt, A. M., Kalenskii, S., Shingledecker, C. N., Remijan, A. J., Herbst, E., et al. (2018). Detection of the aromatic molecule benzonitrile (C_6H_5CN) in the interstellar medium. *Science* 359, 202–205. doi:10.1126/science.aao4890

McGuire, B. A., Loomis, R. A., Burkhardt, A. M., Lee, K. L. K., Shingledecker, C. N., Charnley, S. B., et al. (2021). Detection of two interstellar polycyclic aromatic hydrocarbons via spectral matched filtering. *Science* 371, 1265–1269. doi:10.1126/science.abb7535

Melli, A., Melosso, M., Bizzocchi, L., Alessandrini, S., Jiang, N., Tonolo, F., et al. (2022). Rotational spectra of unsaturated carbon chains produced by pyrolysis: the case of propadienone, cyanovinylacetylene, and allenylacetylene. *J. Phys. Chem. A* 126, 6210–6220. doi:10.1021/acs.jpca.2c05018

Melli, A., Tonolo, F., Barone, V., and Puzzarini, C. (2021). Extending the applicability of the semi-experimental approach by means of "Template Molecule" and "Linear Regression" models on top of DFT computations. *J. Phys. Chem. A* 125, 9904–9916. doi:10.1021/acs.jpca.1c07828

Melosso, M., Belloche, A., Martin-Drumel, M. A., Pirali, O., Tamassia, F., Bizzocchi, L., et al. (2020a). Far-infrared laboratory spectroscopy of aminoacetonitrile and first

interstellar detection of its vibrationally excited transitions. *Astron. Astrophys.* 641, A160. doi:10.1051/0004-6361/202038466

Melosso, M., Bizzocchi, L., Dore, L., Kisiel, Z., Jiang, N., Spezzano, S., et al. (2021). Improved centrifugal and hyperfine analysis of ND_2H and NH_2D and its application to the spectral line survey of L1544. *J. Mol. Spectrosc.* 377, 111431. doi:10.1016/j.jms.2021.111431

Melosso, M., Bizzocchi, L., Gazzeh, H., Tonolo, F., Guillemin, J. C., Alessandrini, S., et al. (2022). Gas-phase identification of (Z)-1,2-ethenediol, a key prebiotic intermediate in the formose reaction. *Chem. Commun.* 58, 2750–2753. doi:10.1039/D1CC06919E

Melosso, M., Bizzocchi, L., Tamassia, F., Degli Esposti, C., Canè, E., and Dore, L. (2019a). The rotational spectrum of ^{15}ND. Isotopic-independent Dunham-type analysis of the imidogen radical. *Phys. Chem. Chem. Phys.* 21, 3564–3573. doi:10.1039/c8cp04498h

Melosso, M., Conversazioni, B., Degli Esposti, C., Dore, L., Cané, E., Tamassia, F., et al. (2019b). The pure rotational spectrum of $^{15}ND_2$ observed by millimetre and submillimetre-wave spectroscopy. *J. Quant. Spectrosc. Radiat. Transf.* 222, 186–189. doi:10.1016/j.jqsrt.2018.10.028

Melosso, M., Dore, L., Gauss, J., and Puzzarini, C. (2020b). Deuterium hyperfine splittings in the rotational spectrum of NH_2D as revealed by lamb-dip spectroscopy. *J. Mol. Spectrosc.* 370, 111291. doi:10.1016/j.jms.2020.111291

Melosso, M., Melli, A., Puzzarini, C., Codella, C., Spada, L., Dore, L., et al. (2018). Laboratory measurements and astronomical search for cyanomethanimine. *Astron. Astrophys.* 609, A121. doi:10.1051/0004-6361/201731972

Mills, I. M. (1972). "Vibration-rotation structure in asymmetric- and symmetric-top molecules," in *Molecular spectroscopy: Modern research*. Editors K. N. Rao, and C. W. Matthews (New York: Academic Press), 115–140.

Møller, C., and Plesset, M. S. (1934). Note on an approximation treatment for many-electron systems. *Phys. Rev.* 46, 618–622. doi:10.1103/physrev.46.618

Morgan, W. J., Matthews, D. A., Ringholm, M., Agarwal, J., Gong, J. Z., Ruud, K., et al. (2018). Geometric energy derivatives at the complete basis set limit: application to the equilibrium structure and molecular force field of formaldehyde. *J. Chem. Theory Comput.* 14, 1333–1350. doi:10.1021/acs.jctc.7b01138

Müller, H. S. P., Schlöder, F., Stutzki, J., and Winnewisser, G. (2005). The cologne database for molecular spectroscopy, CDMS: A useful tool for astronomers and spectroscopists. *J. Mol. Struct.* 742, 215–227. doi:10.1016/j.molstruc.2005.01.027

Negron-Mendoza, A., and Ramos-Bernal, S. (2001). "Interstellar molecules: an exotic chemistry," in *First steps in the origin of life in the universe*. Editors J. Chela-Flores, T. Owen, and F. Raulin (Dordrecht: Springer Netherlands), 145–150.

Orr, B. J. (1995). Collision-induced state-to-state energy transfer in perturbed rovibrational manifolds of small polyatomic molecules: mechanistic insights and observations. *Chem. Phys.* 190, 261–278. doi:10.1016/0301-0104(94)00279-J

Papajak, E., Zheng, J., Xu, X., Leverentz, H. R., and Truhlar, D. G. (2011). Perspectives on basis sets beautiful: seasonal plantings of diffuse basis functions. *J. Chem. Theory Comput.* 7, 3027–3034. doi:10.1021/ct200106a

Penn, R. E. (1978). Microwave spectrum of 2-propene-1-imine, CH_2=CHCH=NH. *J. Mol. Spectrosc.* 69, 373–382. doi:10.1016/0022-2852(78)90230-8

Penocchio, E., Piccardo, M., and Barone, V. (2015). Semiexperimental equilibrium structures for building blocks of organic and biological molecules: the B2PLYP route. *J. Chem. Theory Comput.* 11, 4689–4707. doi:10.1021/acs.jctc.5b00622

Peterson, K. A., and Dunning, T. H. (2002). Accurate correlation consistent basis sets for molecular core-valence correlation effects: the second row atoms Al-Ar, and the first row atoms B-Ne revisited. *J. Chem. Phys.* 117, 10548–10560. doi:10.1063/1.1520138

Peterson, K. A., Feller, D., and Dixon, D. (2012). Chemical accuracy in *ab initio* thermochemistry and spectroscopy: current strategies and future challenges. *Theor. Chem. Acc.* 113, 1079. doi:10.1007/s00214-011-1079-5

Piccardo, M., Penocchio, E., Puzzarini, C., Biczysko, M., and Barone, V. (2015). Semi-experimental equilibrium structure determinations by employing B3LYP/SNSD anharmonic force fields: validation and application to semirigid organic molecules. *J. Phys. Chem. A* 119, 2058–2082. doi:10.1021/jp511432m

Pickett, H. M., Poynter, R. L., Cohen, E. A., Delitsky, M. L., Pearson, J. C., and Müller, H. S. P. (1998). Submillimeter, millimeter and microwave spectral line catalog. *J. Quant. Spectrosc. Radiat. Transf.* 60, 883–890. doi:10.1016/S0022-4073(98)00091-0

Pickett, H. M. (1991). The fitting and prediction of vibration-rotation spectra with spin interactions. *J. Mol. Spectrosc.* 148, 371–377. doi:10.1016/0022-2852(91)90393-o

Prestage, R. M., Constantikes, K. T., Hunter, T. R., King, L. J., Lacasse, R. J., Lockman, F. J., et al. (2009). The green bank telescope. *Proc. IEEE* 97, 1382–1390. doi:10.1109/jproc.2009.2015467

Pulay, P., Meyer, W., and Boggs, J. E. (1978). Cubic force constants and equilibrium geometry of methane from Hartree–Fock and correlated wavefunctions. *J. Chem. Phys.* 68, 5077–5085. doi:10.1063/1.435626

Puzzarini, C., Alessandrini, S., Bizzocchi, L., and Melosso, M. (2023). Hunting for interstellar molecules: rotational spectra of reactive species. *Faraday Discuss.* doi:10.1039/D3FD00052D

Puzzarini, C., and Barone, V. (2011). Extending the molecular size in accurate quantum-chemical calculations: the equilibrium structure and spectroscopic properties of uracil. *Phys. Chem. Chem. Phys.* 13, 7189–7197. doi:10.1039/C0CP02636K

Puzzarini, C., Bloino, J., Tasinato, N., and Barone, V. (2019). Accuracy and interpretability: the devil and the holy grail. new routes across old boundaries in computational spectroscopy. *Chem. Rev.* 119, 8131–8191. doi:10.1021/acs.chemrev.9b00007

Puzzarini, C., Cazzoli, G., Harding, M. E., Vázquez, J., and Gauss, J. (2009). A new experimental absolute nuclear magnetic shielding scale for oxygen based on the rotational hyperfine structure of $H_2^{17}O$. *J. Chem. Phys.* 131, 234304. doi:10.1063/1.3274062

Puzzarini, C., Cazzoli, G., Harding, M. E., Vázquez, J., and Gauss, J. (2015). The hyperfine structure in the rotational spectra of $D_2^{17}O$ and $HD^{17}O$: confirmation of the absolute nuclear magnetic shielding scale for oxygen. *J. Chem. Phys.* 142, 124308. doi:10.1063/1.4916068

Puzzarini, C. (2020). Grand challenges in astrochemistry. *Front. Astron. Space Sci.* 7, 19. doi:10.3389/fspas.2020.00019

Puzzarini, C., Heckert, M., and Gauss, J. (2008). The accuracy of rotational constants predicted by high-level quantum-chemical calculations. I. molecules containing first-row atoms. *J. Chem. Phys.* 128, 194108. doi:10.1063/1.2912941

Puzzarini, C., and Stanton, J. F. (2023). Connections between the accuracy of rotational constants and equilibrium molecular structures. *Phys. Chem. Chem. Phys.* 25, 1421–1429. doi:10.1039/D2CP04706C

Puzzarini, C., Stanton, J. F., and Gauss, J. (2010). Quantum-chemical calculation of spectroscopic parameters for rotational spectroscopy. *Int. Rev. Phys. Chem.* 29, 273–367. doi:10.1080/01442351003643401

Raghavachari, K., Trucks, G. W., Pople, J. A., and Head-Gordon, M. (1989). A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* 157, 479–483. doi:10.1016/s0009-2614(89)87395-6

Rivilla, V. M., Colzi, L., Jiménez-Serra, I., Martín-Pintado, J., Megías, A., Melosso, M., et al. (2022). Precursors of the rna world in space: detection of (Z)-1,2-ethenediol in the interstellar medium, a key intermediate in sugar formation. *Astrophys. J. Lett.* 929, L11. doi:10.3847/2041-8213/ac6186

Rivilla, V. M., Martín-Pintado, J., Jiménez-Serra, I., Zeng, S., Martín, S., Armijos-Abendaño, J., et al. (2018). Abundant Z-cyanomethanimine in the interstellar medium: paving the way to the synthesis of adenine. *Mon. Not. R. Astron. Soc.* 483, L114–L119. doi:10.1093/mnrasl/sly228

Roueff, E., and Lique, F. (2013). Molecular excitation in the interstellar medium: recent advances in collisional, radiative, and chemical processes. *Chem. Rev.* 113, 8906–8938. doi:10.1021/cr400145a

Santra, G., Sylvetsky, N., and Martin, J. M. L. (2019). Minimally empirical double-hybrid functionals trained against the GMTKN55 database: revDSD-PBEP86-D4, revDOD-PBE-D4, and DOD-SCAN-D4. *J. Phys. Chem. A* 123, 5129–5143. doi:10.1021/acs.jpca.9b03157

Silva, W. G. D. P., Cernicharo, J., Schlemmer, S., Marcelino, N., Loison, J. C., Agúndez, M., et al. (2023). Discovery of H_2CCCH^+ in TMC-1. *Astron. Astrophys.* 676, L1. doi:10.1051/0004-6361/202347174

Sita, M. L., Changala, P. B., Xue, C., Burkhardt, A. M., Shingledecker, C. N., Lee, K. L. K., et al. (2022). Discovery of interstellar 2-cyanoindene (2-C_9H_7CN) in GOTHAM observations of TMC-1. *Astrophys. J. Lett.* 938, L12. doi:10.3847/2041-8213/ac92f4

Spaniol, J. T., Lee, K. L. K., Pirali, O., Puzzarini, C., and Martin-Drumel, M. A. (2023). A rotational investigation of the three isomeric forms of cyanoethynylbenzene (HCC–C_6H_4–CN): benchmarking experiments and calculations using the "lego brick" approach. *Phys. Chem. Chem. Phys.* 25, 6397–6405. doi:10.1039/d2cp04825f

Tafalla, M., Myers, P., Caselli, P., Walmsley, C., and Comito, C. (2002). Systematic molecular differentiation in starless cores. *Astrophys. J.* 569, 815–835. doi:10.1086/339321

Tajti, A., Szalay, P. G., Császár, A. G., Kállay, M., Gauss, J., Valeev, E. F., et al. (2004). Heat: high accuracy extrapolated *ab initio* thermochemistry. *J. Chem. Phys.* 121, 11599–11613. doi:10.1063/1.1811608

Tennyson, J. (2019). *Astronomical spectroscopy. An introduction to the atomic and molecular physics of astronomical spectroscopy (WSPC)*. Europe: World Scientific. doi:10.1142/q0207

Tercero, F., López-Pérez, J. A., Gallego, J. D., Beltrán, F., García, O., Patino-Esteban, M., et al. (2021). Yebes 40 m radio telescope and the broad band Nanocosmos receivers at 7 mm and 3 mm for line surveys. *Astron. Astrophys.* 645, A37. doi:10.1051/0004-6361/202038701

Tielens, A. (2021). *Molecular astrophysics*. Cambrindge: Cambrindge Unversity Press. doi:10.1017/9781316718490

Tonolo, F., Bizzocchi, L., Melosso, M., Lique, F., Dore, L., Barone, V., et al. (2021). An improved study of HCO^+ and He system: interaction potential, collisional relaxation, and pressure broadening. *J. Chem. Phys.* 155, 234306. doi:10.1063/5.0075929

Tonolo, F., Lique, F., Melosso, M., Puzzarini, C., and Bizzocchi, L. (2022). Hyperfine resolved rate coefficients of $HC^{17}O^+$ with H_2 (j = 0). *Mon. Not. R. Astron. Soc.* 516, 2653–2661. doi:10.1093/mnras/stac2394

Watson, J. K. G. (1977). "Aspects of quartic and sextic centrifugal effects on rotational energy levels," in *Vibrational spectra and structure*. Editor J. Durig (Amsterdam: Elsevier), 1–89.

Watson, J. K. G. (1968). Determination of centrifugal distortion coefficients of asymmetric-top molecules. III. sextic coefficients. *J. Chem. Phys.* 48, 4517–4524. doi:10.1063/1.1668020

Wilson, A. K., van Mourik, T., and Dunning, T. H. (1996). Gaussian basis sets for use in correlated molecular calculations. VI. sextuple zeta correlation consistent basis sets for boron through neon. *J. Mol. Struct. THEOCHEM* 388, 339–349. doi:10.1016/S0166-1280(96)80048-0

Woon, D. E. (2004). The astrochymist: an internet resource for astrochemists and interested bystanders. Avaliable At: https://www.astrochymist.org/.

Woon, D. E., and Dunning, T. H. (1995). Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon. *J. Chem. Phys.* 103, 4572–4585. doi:10.1063/1.470645

Wootten, A., and Thompson, A. R. (2009). The atacama large millimeter/submillimeter array. *IEEE Proc.* 97, 1463–1471. doi:10.1109/jproc.2009.2020572

Yamamoto, S. (2017). *Introduction to astrochemistry: Chemical evolution from interstellar clouds to star and planet formation*. Tokyo: Springer Tokyo. doi:10.1007/978-4-431-54171-4

Ye, H., Alessandrini, S., Melosso, M., and Puzzarini, C. (2022). Exploiting the "Lego brick" approach to predict accurate molecular structures of PAHs and PANHs. *Phys. Chem. Chem. Phys.* 24, 23254–23264. doi:10.1039/D2CP03294E

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 15 September 2023
DOI 10.3389/fspas.2023.1186172

Check for updates

OPEN ACCESS

EDITED BY
Ryan C. Fortenberry,
University of Mississippi, United States

REVIEWED BY
Vladislav Chernov,
Voronezh State University, Russia
Heather Abbott-Lyon,
Kennesaw State University, United States

*CORRESPONDENCE
Martin Ferus,
martinferus@email.cz

RECEIVED 14 March 2023
ACCEPTED 28 August 2023
PUBLISHED 15 September 2023

CITATION
Ferus M, Knížek A, Cassone G, Rimmer PB, Changela H, Chatzitheodoridis E, Uwarova I, Žabka J, Kabáth P, Saija F, Saeidfirozeh H, Lenža L, Krůs M, Petera L, Nejdl L, Kubelík P, Křivková A, Černý D, Divoký M, Pisařík M, Kohout T, Palamakumbure L, Drtinová B, Hlouchová K, Schmidt N, Martins Z, Yáñez J, Civiš S, Pořízka P, Mocek T, Petri J and Klinkner S (2023), Simulating asteroid impacts and meteor events by high-power lasers: from the laboratory to spaceborne missions.
Front. Astron. Space Sci. 10:1186172.
doi: 10.3389/fspas.2023.1186172

COPYRIGHT
© 2023 Ferus, Knížek, Cassone, Rimmer, Changela, Chatzitheodoridis, Uwarova, Žabka, Kabáth, Saija, Saeidfirozeh, Lenža, Krůs, Petera, Nejdl, Kubelík, Křivková, Černý, Divoký, Pisařík, Kohout, Palamakumbure, Drtinová, Hlouchová, Schmidt, Martins, Yáñez, Civiš, Pořízka, Mocek, Petri and Klinkner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Simulating asteroid impacts and meteor events by high-power lasers: from the laboratory to spaceborne missions

Martin Ferus[1]*, Antonín Knížek[1,2], Giuseppe Cassone[3], Paul B. Rimmer[4], Hitesh Changela[1], Elias Chatzitheodoridis[5], Inna Uwarova[6], Ján Žabka[1], Petr Kabáth[7], Franz Saija[3], Homa Saeidfirozeh[1], Libor Lenža[1], Miroslav Krůs[8], Lukáš Petera[1,2], Lukáš Nejdl[9], Petr Kubelík[1], Anna Křivková[1,10], David Černý[1], Martin Divoký[11], Michael Pisařík[11], Tomáš Kohout[12,13], Lakshika Palamakumbure[12], Barbora Drtinová[10], Klára Hlouchová[14], Nikola Schmidt[15], Zita Martins[16], Jorge Yáñez[17], Svatopoluk Civiš[1], Pavel Pořízka[18], Tomáš Mocek[11], Jona Petri[19] and Sabine Klinkner[19]

[1]J. Heyrovský Institute of Physical Chemistry, Czech Academy of Sciences, Prague, Czechia, [2]Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University, Prague, Czechia, [3]Institute for Physical-Chemical Processes, National Research Council of Italy (IPCF-CNR), Messina, Italy, [4]University of Cambridge, Cavendish Astrophysics, Cambridge, United Kingdom, [5]School of Mining and Metallurgical Engineering, National Technical University of Athens, Athens, Greece, [6]S A B Aerospace s r o, Brno, Czechia, [7]Astronomical Institute CAS, Ondřejov, Czechia, [8]Institute of Plasma Physics CAS, Prague, Czechia, [9]Department of Chemistry and Biochemistry, Mendel University in Brno, Brno, Czechia, [10]Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, Czechia, [11]HiLASE Centre, Institute of Physics of the Czech Academy of Sciences, Prague, Czechia, [12]Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland, [13]Institute of Geology of the Czech Academy of Sciences, Prague, Czechia, [14]Department of Biochemistry, Faculty of Science, Charles University, Prague, Czechia, [15]Institute of International Relations Prague, Prague, Czechia, [16]Departamento de Engenharia Química, Universidade de Lisboa, Lisboa, Portugal, [17]Departamento de Química Analítica e Inorgánica, University of Concepción, Concepción, Chile, [18]Central European Institute of Technology, Brno University of Technology, Brno, Czechia, [19]Institute of Space Systems, University of Stuttgart, Stuttgart, Germany

Meteor plasmas and impact events are complex, dynamic natural phenomena. Simulating these processes in the laboratory is, however, a challenge. The technique of laser induced dielectric breakdown was first used for this purpose almost 50 years ago. Since then, laser-based experiments have helped to simulate high energy processes in the Tunguska and Chicxulub impact events, heavy bombardment on the early Earth, prebiotic chemical evolution, space weathering of celestial bodies and meteor plasma. This review summarizes the current level of knowledge and outlines possible paths of future development.

KEYWORDS

astrochemistry, laser, ablation, origin of life, exoplanet

1 Introduction

Currently, there are two main approaches to studying meteors and impacts—laboratory experiments, and *in silico* theoretical calculations. Both require understanding of the structure, dynamics, chemistry and spectroscopy of the meteor plasma, and disintegration of meteoroids after atmospheric entry (Madiedo et al., 2014; Kuwahara and Sugita, 2015; Oppenheim and Dimant, 2015; Vojacek et al., 2015; Borovička and Berezhnoy, 2016; Rotelli et al., 2016; Berezhnoy et al., 2018; Ferus et al., 2018; Silber et al., 2018; Bizzarri et al., 2020; Ferus et al., 2020; Zahnle et al., 2020; Ferus et al., 2021; Popov et al., 2021; Kubelík et al., 2022; Zakuskin et al., 2023). Meteor science is also important for future cosmic and engineering applications, such as the mitigation of potential catastrophic impacts (Collins et al., 2005), asteroid mining (Sonter, 1997; Elvis, 2012; Mueller and Van Susante, 2012), future spaceborne observation of meteors (Jenniskens et al., 2000; Carbary et al., 2003; Ishimaru et al., 2014; Rambaux et al., 2014; Petri et al., 2019; Chen et al., 2020; Czech Ministry of Transportation, 2021), or the detection of impacts on (exo)planets by space telescopes (PLATO, Ariel, JWST (Rimmer et al., 2019).

In the first part, this review summarizes contemporary experimental approaches to studying meteor and impact plasma physics and chemistry. The second part of this paper focuses on recent developments in the application of high power lasers to spaceborne observations of meteors in the Earth's atmosphere (Ferus et al., 2019; Czech Ministry of Transportation, 2021), detection of impacts and atmospheric reprocessing on exoplanets (Ferus et al., 2019; Navarro-González et al., 2019; Navarro et al., 2020; Ferus et al., 2021), space weathering and meteoroid ablative disintegration during atmospheric entry (Kaluna et al., 2017; Křivková et al., 2022).

2 State of the art experimental approaches

A large number of computational and observational studies on meteor science has been published (Ceplecha et al., 1998; Rogers et al., 2005; Johnston et al., 2018; Silber et al., 2018) and references therein for details). On the other hand, experimental investigations are relatively sparse, particularly because of the complexity of this phenomena. There exist four main experimental approaches: (a) hypervelocity gun experiments, (b) shock tube experiments, (c) air plasma flows and (d) high power lasers.

2.1 Hypervelocity gun experiments

Hypervelocity gun target experiments are a known tool in impact physics (Libourel et al., 2019). However, the typical speed of such projectiles reaches 3–7 km s^{-1} (Hibbert et al., 2017), which is a major drawback for simulating meteor plasma during atmospheric entry, where the meteoroids reach significantly higher velocities. This projectile velocity is also less than the minimum velocity of a meteoroid, which is ~11.2 km s^{-1} given by the gravity of the Earth (Ceplecha et al., 1998). Furthermore, meteor ablation begins during atmospheric entry at altitudes typically near 120 ± 10 km (depending on the meteoroid mechanical strength and composition (Adolfsson and Gustafson, 1994; Ceplecha et al., 1998). During hypervelocity gun experiments, the plasma forms only after collision with the target. The projectiles are made of synthetic materials, such as steel, or hard natural materials, such as basalt or dunite. The composition of these materials is naturally not representative of chondritic materials. Chondritic materials, on the other hand, cannot be used as projectiles, because they cannot survive the compression of the gun (Libourel et al., 2019).

However, gun experiments remain an important tool for understanding, e.g., impact-induced formation of biologically relevant materials (Blank et al., 2001; Blank et al., 2001; Furukawa et al., 2009; Martins et al., 2013), observed, e.g., in sterilized stainless steel containers (Martins et al., 2013). Several other studies have performed impact shock synthesis of organic molecules important for the origin of life (Blank et al., 2001; Blank et al., 2001; Furukawa et al., 2009). Some of those were supported by *ab initio* molecular dynamical simulations (Goldman et al., 2010).

2.2 Shock tube experiments

Shock tube experiments are of great importance to impact plasma chemistry (Dremin, 1989; Roy et al., 2022). These experiments are useful for the assessment of the survival potential, decomposition lifetime, and *de novo* synthesis of molecules in a cometary (Rubin et al., 2019) or asteroid impact event (Singh et al., 2022). Recently, piston driven shock tube experiments were used to heat Ar and N_2 gas mixture (Jayaram et al., 2013) to significant temperatures approaching conditions of impacts (9,100–12 300 K) with reflected shock pressure of about 59–70 bar for about 2–4 ms. Shock tube experiments remain an important source of data for impact plasma simulation, especially regarding its chemical consequences.

2.3 Air plasma flows

Another approach to simulating meteor plasmas is by exposing meteorite samples to powerful heat sources such as high-enthalpy subsonic air plasma flows in a wind tunnel (Loehle et al., 2017a; Agrawal et al., 2018; Drouard et al., 2018) or (Helber et al., 2019; Loehle et al., 2022). These experiments are useful for simulating the atmospheric entry of, e.g., spacecraft materials. They have been used for simple spectral line identification from emission spectra of chondrites, achondrites, and iron meteorites (Drouard et al., 2018).

One drawback of these experiments is that the simulated meteor plasma in wind tunnels can attain temperatures up to 2,400 K (Loehle et al., 2017a). Real meteor plasmas exhibit a complicated structure with high temperature regions as hot as 4,000–100,000 K (Silber et al., 2018). Another disadvantage is that the high-enthalpy experiments involve plasma treatment of the specimen material performed in large chambers, which are not fully vacuum sealed, chemically clean and filled with different gas mixtures.

2.4 High power lasers

A fourth possible approach is to use high power lasers. The laser radiation can be well characterized, and they are versatile enough to allow experimental simulation of the chemistry and physics of supernovae explosion shocks, the interiors of gas giants, explosions and plasma jets (Villagrán-Muniz et al., 2003), accretion processes, nuclear fusion, hypervelocity atmospheric entries, impacts and meteors (Ferus et al., 2014b; Ferus et al., 2017b; Rimmer et al., 2019), and lightning discharges (Wooding, 1972; Borucki, 1985; Borucki and McKay, 1987; Borucki et al., 1988; Sobral et al., 2000; Navarro-González et al., 2001a; Sobral et al., 2002; Villagrán-Muniz et al., 2003).

Lasers have multiple advantages for studying consequential chemistry of meteor plasma events: chemical contamination from an electrode material is excluded (Borucki et al., 1988; Sobral et al., 2000), gas volumes are small, the system is generally chemically clean, isolated and well-defined (Sobral et al., 2000), and they can be applied in the meso-to microscale on mineral surfaces to allow experimentation on single phases. The properties of laser induced dielectric breakdown are characterized by pulse duration, energy density, laser wavelength, photon fluency, the chemical nature of the irradiated material or gas density, and chemical composition of the surrounding atmosphere (McKay and Borucki, 1997; Villagrán-Muniz et al., 2003; Saeidfirozeh et al., 2022; Zakuskin et al., 2023). This creates a noticeable range of parameters which are adjustable on demand and based on the required properties suitable for the simulation of the selected phenomena.

The use of lasers for simulation of impact events was theoretically suggested as early as 1964 by Rae and Hertzberg (Rae and Hertzberg, 1964) and experimentally explored by (Hapke et al., 1975), who simulated impact evaporation processes. A pioneering study on the use of small lasers in the exploration of meteorites and meteors was provided by (Pirri, 1977). In parallel, laser-based experiments were also suggested to serve as a simulation of the ball (Wooding, 1972) or lightning discharges (Borucki, 1985; Borucki and McKay, 1987; Jebens et al., 1992). Since then, lasers are used by experimentalists for both simulations of impacts and lightning ever since, as for instance in the case of laser-based experiments using very similar experimental set-ups focused on nitrogen oxide synthesis by impact plasma (Navarro-González et al., 2019; Heays et al., 2022), or lightning plasma (Navarro-González et al., 2001b).

The first connection between laser and lightning physics was discussed in (Borucki, 1985; Borucki and McKay, 1987; Jebens et al., 1992) by identifying the similarity in electron density (7×10^{17} to 9×10^{16} cm^{-3}), temperature (over 16,000 K), energy dissipation per unit spark length (10^4 J m^{-1}), and chemical freeze-out temperature. However, the hydrodynamic evolution of laser-induced dielectric breakdown (LIDB) leads to a faster cooling compared to lightning (Stark et al., 1996; Navarro-González et al., 2001a; Sobral et al., 2002). For example, the high power laser facility Prague Asterix Laser System (PALS) is able to deliver energy of up to 600 J creating a plasma fireball with the electron density 1.5×10^{17}–10^{20} cm^{-3} and the temperature decreasing from 10^5 K to 9,300 K (Šmíd et al., 2019) after the pulse. These conditions are consistent with the findings of Sobral for lightning (Sobral et al., 2000), (Sponer et al., 2020).

However, the suitability of laser-based experiments for meteor and impact plasma simulations is also supported by multiple arguments (Pirri, 1977; Ferus et al., 2017a; Rimmer et al., 2019). For example (McKay and Borucki, 1997), found that the expansion of the high temperature gas in LIDB takes between 1 and 2 μs. These values are consistent with the laser breakdown duration of 1.5 μs, as indicated by the combined experimental and theoretical simulations of laser plasma chemistry (Ferus et al., 2014a). These times are shorter than the typical shocks from an atmospheric entry (from 0.5 s for small meteoroids and up to 10 s for large impacts). A real meteor plasma is quite a complex phenomenon. Two temperature components of 4,400 K and 10,000 K have been confirmed by spectral observations (Borovička, 1994). However the meteor plasma consists in fact of envelopes expected to involve not only plasmas exhibiting relatively low temperatures ranging from 2,900 K to 5,500 K (Jenniskens et al., 2004; Jenniskens and Stenbaek-Nielsen, 2004), but also deeper bow shock regions, as hot as 95, 000 K (Silber et al., 2018). The laser plasma reaches a peak temperature up to 100, 000 K. Then, the plasma cools down quickly to temperatures below 10,000 K as the gas expands from the initial dielectric breakdown location. Spectroscopic examination of the plasma generated by high-power lasers, such as PALS, shows a decrease in temperatures from 100,000 K to 9,300 K (Šmíd et al., 2019; Křivková et al., 2021). Recent development in high repetition high power lasers (Haefner et al., 2017; Mason et al., 2017; Röcker et al., 2020) allows to study phenomena that are connected not only with plasma creation, but also repetitive processes connected with heating of the target, its ablation, etc. Such processes can be experienced by meteoroids during atmospheric entry. The lasers offer peak power from hundreds of megawatts and nanosecond pulses, through tens of gigawatts and picosecond pulses to petawatts and femtosecond pulses, all with hundreds of watts to kilowatt average power.

The major weakness of laser experiments is a mismatch in electron densities between observed meteors and the experiments (Křivková et al., 2021). Spectroscopic (Borovička and Betlem, 1997; Kasuga et al., 2005; Jenniskens, 2007; Ferus et al., 2018; Kubelík et al., 2022) and radar data attest to electron densities 10^{11}–10^{15} cm^{-3}, while laser induced plasma usually reaches electron densities ≈10^{16}–10^{20} cm^{-3}, which fulfils the McWhirter equilibrium criterion (De Giacomo, 2011; Ferus et al., 2018; Ferus et al., 2019; Ferus et al., 2020; Křivková et al., 2021).

Suitability of laser-based experiments for estimation of chemical yields is also of great importance, especially for evaluation of environmental changes triggered by large impacts. Impact events deliver sufficient energy able to chemically transform an atmosphere or a surface and produce new chemicals and aerosols, or enrich the surface of a meteoroid with heavy elements providing environment supporting chemical catalysis (Ferus et al., 2021). This is especially important for early planets exposed to early and late heavy bombardment (Koeberl, 2006; Reimold and Gibson, 2006; Ferus et al., 2021).

Studies using lasers for the simulation of meteor and impact plasma include for instance the synthesis of silicon oligomers and simple hydrocarbons (Managadze et al., 2003), hydrogen cyanide and acetylene formation (Scattergood et al., 1989; McKay and Borucki, 1997), or origin of nitrates on early Mars (Navarro-González et al., 2019). During the last

two decades, our team has also conducted several experiments on the PALS laser facility focusing mostly on the prebiotic synthesis and the impact transformation of planetary atmospheres. Subsequent studies successfully demonstrated the crucial role of impact plasma in the origin of canonical nucleobases and amino acids (Ferus et al., 2014a; Ferus et al., 2014b; Ferus et al., 2017b), sugars (Civis et al., 2016). Studies with this and other lasers also showed the transformation of simple molecules in early terrestrial planet environments, such as hydrogen cyanide (Ferus et al., 2017a), acetylene, methane (Civiš et al., 2017), carbon monoxide (Civiš et al., 2008) and aromatic compounds (Petera et al., 2023).

Moreover, laser plasma has been used to perform ablation experiments (Ferus et al., 2018; Ferus et al., 2019; Ferus et al., 2020; Křivková et al., 2021). For instance, in the first comprehensive work published by (Hawkes et al., 2008), it has been proposed that laboratory-based laser ablation techniques can be used to study the size of the luminous region, predict spectral features in meteors, estimate the luminous efficiency factor, and assess the role of chemically differentiated thermal ablation of meteoroids. Then (Milley et al., 2007), simulated the meteor luminosity through laser ablation of meteorites and later (Ebert et al., 2017) studied the virtually instantaneous melting of target rocks during meteorite impacts. They discovered that the entropy changes when laser-melting sandstone and iron meteorites corresponds to a minimum impact velocity which is approximately 6 km s^{-1}, inducing peak shock pressures at around 100 GPa on the target (Zakharov, 2003) also concluded that laser experiments can reproduce several phenomena such as barium release in Earth's magnetosphere, collision less deceleration of supernova remnants and related shockwave generation in the interstellar medium or near-Earth asteroid laser defense system.

Alongside the above-mentioned studies, space weathering simulations and asteroid deflection by lasers remain the most studied research topics connected with the application of lasers. Space weathering is a common process affecting the surfaces of airless bodies across our Solar System. It is caused by a combination of the solar wind, micrometeorite bombardment, and galactic radiation, and it alters the physical, chemical, and crystallographic properties of surface materials (Hapke, 2001; Bennett et al., 2013; Domingue et al., 2014; Pieters and Noble, 2016). The micrometeorite bombardment component caused by high-velocity sub-millimeter dust grains is often simulated by pulse laser irradiation. The idea behind use of laser pulse is that its short duration and small footprint rather realistically mimics the kinetic energy deposition of an impacting dust particle and its conversion to thermal energy causing localized melting and vaporization with subsequent melt sputtering, and vapor redeposition. First attempts (Moroz et al., 1996) used rather long ~ 1 μs pulse durations which was about 1,000 times longer than expected timescale of micrometeorite impact (Sasaki et al., 2001). Thus shorter, nanosecond, or later femtosecond laser pulses were later applied and reproduced well the space weathering morphology of natural microimpacts (Fulvio et al., 2021) including microstructure of amorphous rims often with reduced nanophase iron (Sasaki et al., 2001; Kurahashi et al., 2002; Sasaki et al., 2002; Sasaki et al., 2003; Worms, 2003; Brunetto et al., 2006; Lazzarin et al., 2006; Markley et al., 2013), and redeposition of impact ejecta on mineral surfaces (Loeffler et al., 2008). The laser-irradiated reflectance spectrum of minerals and meteorites match these of asteroids or the Moon (Yamada et al., 1999; Sasaki et al., 2001; Sasaki et al., 2002; Jiang et al., 2019; Matsuoka et al., 2020). While nanosecond laser pulse causes mainly heating and melting with subsonic and sonic evaporation from the surface (Gusarov and Smurov, 2005) the femtosecond laser pulse with high peak irradiance allows the propagation of a shock wave of several tens of GPa and confined melting (Boustie et al., 2008; Berthe et al., 2011) being better proxy to shock processes associated with natural impacts and prevents undesired interaction of the laser with the generated plasma plume as can happen in nanosecond laser pulse. Ablation of the material proceeds through spallation, fragmentation, homogeneous nucleation, and vaporization (Perez and Lewis, 2003), resulting in structure of the microcraters that is remarkably similar not only to natural microimpact craters, but also to that of large craters on the Moon or asteroid (25,143) Itokawa (Fazio et al., 2018).

3 Future applications of high-power lasers

Laboratory-based astrochemistry and astrophysics, directly connected to the upcoming space missions, will determine the field of high-power laser applications in the future. We are currently in an era of unprecedented opportunities for proposing a flight of a spacecrafts exploring nearly any object in the Solar System. At the same time, space telescopes such as CoRoT, Kepler/K2 and TES along with ground-based observatories confirmed over 5,300 exoplanets. The known sample of confirmed exoplanets has shown an astonishing diversity in properties and orbital parameters. In the upcoming decade the space mission PLATO is expected to observe a sample of about a million bright stars for planetary transits. The first detection of exo-atmospheres is dated back to 2002 (Charbonneau et al., 2002). We possess now more details about exo-atmospheres of gas planets (Redfield et al., 2008; Wyttenbach et al., 2015; Kabáth et al., 2019; Kabáth et al., 2019). Characterization of exo-atmospheres is still a rapidly developing discipline. The telescopes such as JWST or Ariel and beyond (Tinetti et al., 2018) will fill this gap mainly by observation of atmospheric spectral imprints (Guilluy et al., 2022). In spite of the wealth of information to be returned by these spacecrafts, ground based experimental support will be needed in order to create a spectral and chemical knowledge for supporting the characterization of exoplanets as well as to learn more about geochemistry of interplanetary space, more details about primordial highly dynamic nebular accretion processes, planetary evolution, origin of life and its evolution, and finally the potential of mineral resources for metals and rare elements.

In the following section on future applications, we specifically discuss application potential of laser-based laboratory experiments with regard to upcoming space telescopes (Ferus et al., 2017b), orbiters (Borucki et al., 1996; Ferus et al., 2019), landing probes and rovers (Lasue et al., 2012; Wiens et al., 2012; Wiens et al., 2012; Clegg et al., 2017), for the analysis of propulsion systems (Hudson and Lemmer, 2017), plasma created by natural (Cipriano et al., 2018) or artificial (Lee et al., 2017) kinetic impactors as planned for Moon

or asteroid missions (Cheng et al., 2018), or remote sensing from a small flyby or orbiting satellites (Arnold et al., 2019). The specific future perspectives of laser-based regarding all these specific fields is provided below. We also discuss future uses of laser experiments in the analysis of meteor plasmas.

3.1 Spaceborne observation of meteors

Spaceborne spectral survey of meteors or impact events evades many limitations of ground-based observations and therefore, represents very promising future application field for simulating meteors and impacts using laser-based experiments. Observations provided by ground based spectroscopy are limited by the absorption of atmospheric molecules (Kaltenegger et al., 2020), aerosols, molecular scattering (Yan et al., 2015), or by weather conditions (Jenniskens, 2007; Milley et al., 2007; Vojacek et al., 2015; Ferus et al., 2018; Ferus et al., 2019). This situation can be easily demonstrated on a broad band spectral data compilation of the Leonid meteors (Carbary et al., 2003; Ferus et al., 2018; Kubelík et al., 2022), depicted in Figure 1, panel A, with indication of particular regions accessible by the naked eye, ground based video camera, spaceborne optical-UV emission hyperspectral camera on a CubeSat, and space telescopes. Laser-based experiments have been so far used for mapping the spectral features in visible region, as depicted in Figure 1, panel B. Future applications will support to identify and describe meteor spectral features in other regions. The ground-based observations are not limited only by capability to record spectra mostly in ViS range, but also by restricted number of convenient observation places, which must not be located only in accessible places on land, but also in politically stable and cooperating countries with institutions open to international collaboration. Observation of meteors by a spacecraft avoids most of these problems.

So far, extraterrestrial impacts or meteor events have been observed on the Moon (NASA et al., 2022), Mars (Selsis et al., 2005; Christou et al., 2012; Brown et al., 2014), and Jupiter (Levy, 1998). Considering the outer space, occasional asteroids have been explored, such as 'Oumuamua, see (Bannister et al., 2019) and references therein, or sporadic meteors suspected to be having an extrasolar origin (Afanasiev et al., 2007). It has been proposed that spectral surveys of interstellar meteors offer a unique opportunity to probe extrasolar systems. The first and only analyzed spaceborne meteor UV spectrum has been captured by the MSX military satellite (Carbary et al., 2003). The only running instrument is currently on ISS in the WORF facility. That project aims at the detection of major elements (Fe, Ca, Mg, and Na) in the spectral range between 304–700 nm (Kramer, 1994). On the other hand, several satellite mission concepts have been already proposed: A proposal by Nuth et al. (Nuth et al., 2008) suggesting the observation of meteors in the UV range, between 125–300 nm, has been followed by Rambaux et al. (Rambaux et al., 2014), proposing a 3U-Nanosatellite equipped with a UV spectrometer and ViS camera for meteors and space debris that re-enter the terrestrial atmosphere (Chen et al., 2020). Similar CubeSats, such as S-CUBE, have been developed in parallel by Ishimaru et al. (Ishimaru et al., 2014), while recently Petri et al. (Petri et al., 2019) introduced the FACIS mission for stereoscopic meteor observation. A combined orbital spectroscopic-stereoscopic survey of meteors has been recently proposed in the framework of the Czech CubeSat mission SLAVIA (Czech Ministry of Transportation, 2021).

Notably, spectral records of plasma sources emitting radiation in the atmosphere do not include only natural meteors, but also fireball or airglow surrounding artificial objects entering the atmosphere (Loehle et al., 2021), atmospheric electricity (Pérez-Invernón et al., 2022), rocket launches (Qian et al., 2005) or explosions of manmade objects (Zhang et al., 2018) and interplanetary bodies (Svettsov, 1998; Ferus et al., 2019). The application potential of spaceborne spectroscopy of plasma provided by a hyperspectral camera and supported by laboratory laser-based experiments includes:

(a) Detection, identification, classification and triangulation from at least two satellites which simultaneously observe an individual event (Chen et al., 2020; Petri, 2020). This is important for warning systems, hazard mitigation regarding, e.g., distinguishing atmospheric entries of natural objects from spacecraft or space debris (Coradini et al., 1994; Leiser et al., 2022), space debris surveillance and statistics (Loehle et al., 2017b; Leiser et al., 2022), explosions, rocket launchers (Qian et al., 2005; Langford et al., 2007), and to establish a statistical determination of the mass flux and the distribution of interplanetary material reaching out to the Earth (Rambaux et al., 2021).

(b) Measurement and analysis for scientific or engineering studies, which include physical characterization of a plasma surrounding a spacecraft or meteoroids (Martin and Boyd, 2012), exploration of plasma dynamics, or the evaporation and ablation of surfaces during the atmospheric entry (Silber et al., 2018).

3.2 Detection of impacts on exoplanets

Laser experiments will be crucial for the future identification of markers of impacts, such as in the case of the Shoemaker-Levy comet impact on Jupiter (Zahnle et al., 1995), where S_2, CS_2, OCS, NH_3, HCN, H_2O, and CO have been observed (Zahnle et al., 1995). Recent studies are also associated the tentative detection of PH_3 in the atmosphere of Venus (Bains et al., 2021; Schulze-Makuch, 2021; Truong and Lunine, 2021; Bains et al., 2022) to a stochastic atmospheric corrosion of a single giant impactor or to a random elevation in the impact flux (Omran et al., 2021). A possible future milestone in astronomy is to detect very large meteors and impacts in exoplanetary atmospheres. This will be provided by upcoming spectroscopic studies of exoplanets (Schneider, 2018). However, it has been estimated that an impact of the size of the Chicxulub event, which occurred 65 Mya ago, viewed from 10 light-years away would exhibit an absolute magnitude of 20.9 and be barely detectable even assuming 100% emissivity (Paine, 2006). We can expect that regular meteors will be too faint to be observed with the ARIEL or JWST telescopes. On the other hand, we may assume that in other exoplanet systems, large impact events, explosions or collisions can be occasionally observed such as an analogue of large impact event recently created different twins in the Kepler-107 exoplanet system (Bonomo et al., 2019). For young planetary systems, mathematical models (De Niem et al.,

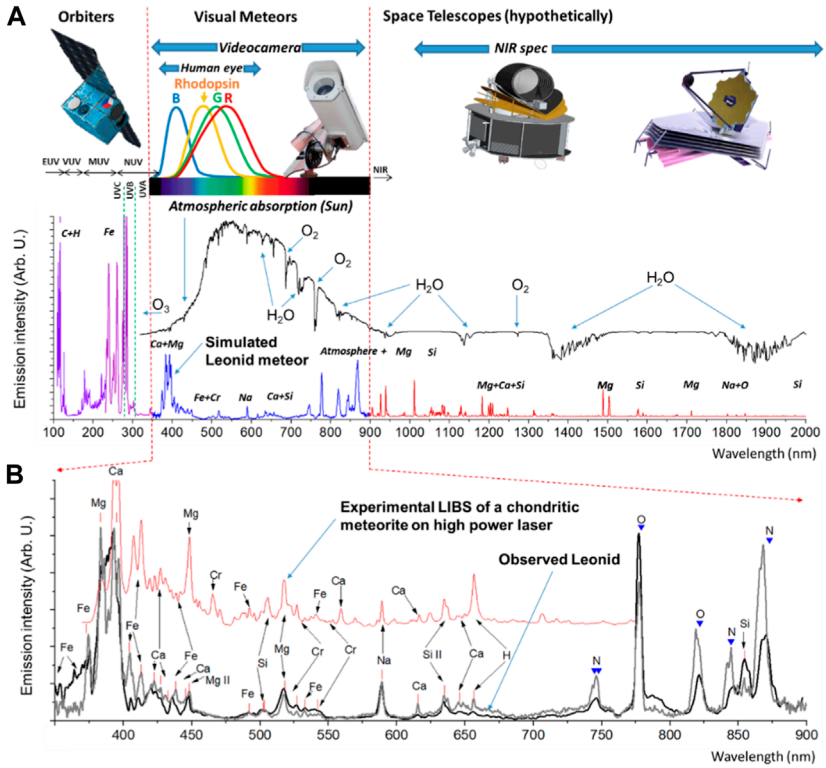


FIGURE 1
Panel **(A)** shows a computer simulation of Leonid meteor emission spectra in the ranges from UV to NIR (violet, blue and red) with indicated regions typically observable by naked eye or ground based video camera. However, these spectral regions are limited by atmospheric transmission embedded in the figure (black). Orbiters allow to overcome this limitation to both UV + visible (likely accessible by a CubeSat, such as the future SLAVIA mission (Czech Ministry of Transportation, 2021), and NIR (operated by space telescopes) regions (Tinetti et al., 2018). Panel **(B)** shows two example spectra of Leonid meteors (grey and black) compared to LIBS spectra of a chondritic meteorite recorded using high power TW-kJ-class laser PALS (red). The data are compiled from the paper (Ferus et al., 2019).

2012) suggest that oligarchic growth, late accretion and early stages of their evolution (i.e., first a few hundred million years) are dominated by giant impacts (Koeberl et al., 2005; Reimold and Gibson, 2006; Quintana et al., 2016). This is supported by looking at the Moon's cratering history (Geiss and Rossi, 2013; Morbidelli et al., 2018), and at the evidence for late veneer on Earth (Brasser et al., 2016). The history of impact mass delivery and size distribution is not entirely resolved, and it is unknown whether other planetary systems would have had the same impact history, because different planet formation models make diverging predictions about the typical early impact frequency (Morbidelli et al., 2018). Such predictions are wedded to planet formation history and planetary system architectures (Genda et al., 2017; Sinclair et al., 2020). Impact history is also relevant to the question of habitability, because the extraplanetary delivery of chemical compounds is important to the evolution of environments that can support life (Todd and Öberg, 2020), the origin of life's building blocks in the impact plasmas, (Ferus et al., 2014b; Ferus et al., 2017b), as well as in the post-impact (Ferus et al., 2021), and origin of feedstock molecules for prebiotic synthesis (Martin et al., 2007) as the result of impact-induced atmospheric and surface chemistry (Quintana et al., 2016).

Experiments with laser treatment of gas mixtures, and solid or liquid targets simulating the chemical effect of an impact plasma on the atmosphere, the surface, or an impactor already indicate the synthesis of species to be connected to the plasma chemistry, such as in the case of C_2H_2 and HCN (Ferus et al., 2017a; Rimmer et al., 2019) (shown in Figure 2), as well as of N_2O (Heays et al., 2022), HNCO and $HCONH_2$. This might be crucial to recognizing impact processes on young planets (Ferus et al., 2021). It should be noted that the results of laser-based experiments in our laboratory lead us to the prediction that the recognition of impacts directly on early Earth-like planets will be a very complicated task due to very complex interplay among the geochemical, the physical and the geological processes, because other possible sources must be ruled out. It is likely that the evidence of elevation in impact frequency might have been provided by the detection of the otherwise unexplainable disequilibria in the mixing ratios of molecular species that are associated with the impact chemistry, or their time variation is associated with their formation and decomposition.

3.3 Simulation of meteoroid disintegration

There is also one relatively new potential application for high-power lasers, especially regarding suitability of high power lasers to create laboratory plasmas characterized by high temperature

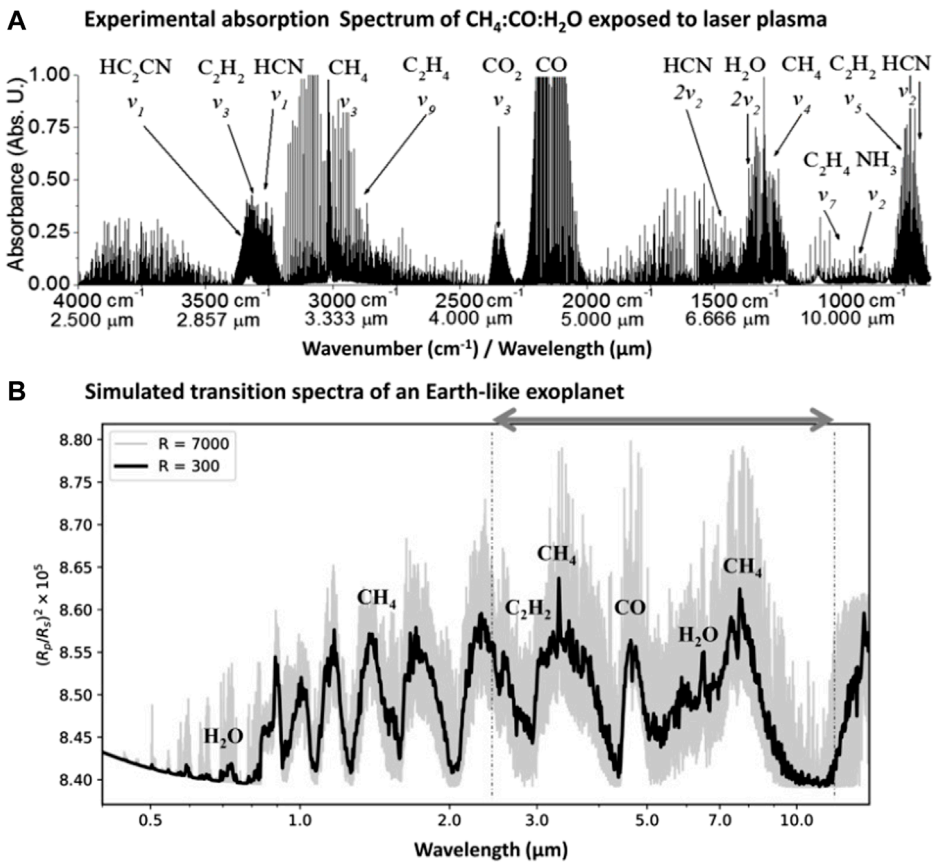


FIGURE 2
Laser-based experiments have been crucial for recent identification of C_2H_2/HCN disequilibrium as indirect evidence of heavy bombardment on early exoplanets with reducing atmosphere (Rimmer et al., 2019). Without a combination of such experiments with sophisticated atmospheric and spectral models, such a prediction is not possible. Chemistry of rather extreme events is not well described and can be understood in close collaboration with *in silico* investigation of reaction networks (Ferus et al., 2017b). The figure is newly compiled from the author´s original dataset published in (Rimmer et al., 2019). The grey arrow in Panel **(B)** shows the wavelength range corresponding to panel **(A)**.

shock expansion of material released from the solid meteorite target (Křivková et al., 2022). The atmospheric entry of a meteoroid does not only result into its sputtering, its fragmentation or ablation (Rogers et al., 2005; Popova, 2006; Vondrak et al., 2008; Silber et al., 2018), or even to meteorite or asteroid impact, if the body is big enough or cohesive (Collins et al., 2005), but also to the random explosion event that has been observed many times for smaller bodies and which are well-known from the case of the Tunguska event in 1908 (Napier and Asher, 2009). Explosive events during atmospheric entry have been recently investigated and explained by (Tabetah and Melosh, 2017; Tabetah and Melosh, 2018); when a meteoroid hurtles through the Earth's atmosphere, high-pressure air in the front of the object infiltrates cracks and pores in the rock, generating a great deal of internal pressure. This pressure is sufficient to cause the object to effectively explode from its core, even when the meteoroid material is strong enough to resist the intense external atmospheric pressures. There's an extreme pressure gradient between the high-pressure air in the front of the descending body and the low-pressure region behind it. If the air can move through the fissures and cracks in the body of the meteorite, it can easily get further inside it and cause fragmentation into pieces. The eventual break-up of a 17–20 m asteroid can release the energy that is equivalent to approximately 500 kt of TNT. As mentioned above, this can be hardly simulated in the laboratory, especially when we are also interested in its chemical evolution. On the other hand, advanced supercomputing techniques can help understanding the fundamental mechanisms at a molecular level by pushing further the limits of laboratory exploration.

4 Conclusion

Laser-based experiments are an excellent tool for generating laboratory plasma for the simulation of specific phenomena which are inaccessible by standard experimental approaches. The laser radiation is a powerful, precisely characterized, and versatile energy source. Such experiments can provide a realistic simulation of extreme states of matter in supernovae explosion shocks, the interiors of gas giants, explosions and plasma jets (Villagrán-Muniz et al., 2003), astrophysical accretion processes, nuclear fusion, hypervelocity atmospheric entries, space weathering,

impacts and meteors (Ferus et al., 2014b; Ferus et al., 2017b; Rimmer et al., 2019) and partly also the simulation of lightning (Wooding, 1972; Borucki, 1985; Borucki and McKay, 1987; Borucki et al., 1988; Sobral et al., 2000; Sobral et al., 2002; Navarro-González et al., 2001a; Villagrán-Muniz et al., 2003). Meteor science has recently broadened to include spaceborne observation of meteor events. Laser simulations play irreplaceable role in ground based support of future space missions (Czech Ministry of Transportation, 2021) by providing knowledge for evaluation of the meteor spectral data for detection, identification, classification and triangulation of meteors (Coradini et al., 1994; Martin and Boyd, 2012; Chen et al., 2020; Petri, 2020; Leiser et al., 2022), as well as space debris surveillance and statistics (Loehle et al., 2017b; Leiser et al., 2022), the detection of explosions, rocket launches (Qian et al., 2005; Langford et al., 2007), or for impacting mass flux and distribution statistics (Silber et al., 2018; Rambaux et al., 2021). Laser experiments also perfectly fit into the landscape of contemporary exoplanetary science. Exoplanets serve as natural laboratories, but for the rigorous interpretation of spectral observations, systematic experimental and theoretical data are needed to support models of planetary environments and spectral retrieval. Only a combination of experiments with models and observations will provide deep insights into exoplanet chemistry and will serve for an extrapolation of chemical and physical parameters broadening our knowledge about evolution and future of our Solar System (Rimmer et al., 2019; Ferus et al., 2021).

Author contributions

MF came up with the idea and prepared initial material for the paper. All authors contributed to the article and approved the submitted version.

Acknowledgments

We acknowledge financial support provided by the Czech Science Foundation (grant reg. no. 21-11366S) and by the Czech Academy of Sciences Program of Regional Cooperation, Reg. No. R200402101. Authors also acknowledge financial support from the Joint Bilateral Project CNR-16964/CNR2211 between the Italian National Research Council (CNR) and the Czech Academy of Sciences (CAS). This paper is a part of a research series originating from activities associated with instrument VESNA within the SLAVIA CubeSat consortium. The feasibility study of the CubeSat has been supported by project SLAVIA in the framework of Ambitious Projects (Mission Proposals) for the Czech Republic: Phase 0/A/B1 Studies, ESA Call ITT3—CZ3_2. Centro de Química Estrutural acknowledges the financial support of Fundação para a Ciência e Tecnologia (FCT) through projects UIDB/00100/2020 and UIDP/00100/2020. Institute of Molecular Sciences acknowledges the financial support of FCT through project LA/P/0056/2020. TK is supported by Academy of Finland project no. 335595. and institutional support RVO 67985831 of the Institute of Geology of the Czech Academy of Sciences. LP is supported by the Doctoral Programme in Geosciences (GeoDoc) of the University of Helsinki. PP gratefully acknowledge the support of the Technology Agency of the Czech Republic within the TREND project (no. FW06010042). Support of the Fondecyt project No. 1231470, ANID-Chile, is also greatly acknowledged. Manufacturing of mirrors for the Ariel telescope is supported by ESA Prodex project under PEA 4000129979. PK acknowledges MSMT grant LTT-20015.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adolfsson, L. G., and Gustafson, B. Å. S. (1994). Effect of meteoroid rotation on atmospheric entry heating and meteor beginning height. *Planet. Space Sci.* 42, 593–598. doi:10.1016/0032-0633(9490034-5)

Afanasiev, V. L., Kalenichenko, V. V., and Karachentsev, I. D. (2007). Detection of an intergalactic meteor particle with the 6-m telescope. *Astrophys. Bull.* 62, 301–310. doi:10.1134/s1990341307040013

Agrawal, P., Jenniskens, P. M., Stern, E. C., Arnold, J. O., and Chen, Y. K. (2018). Arcjet ablation of stony and iron meteorites. *Aerodynamic Meas. Technol. Ground Test. Conf.*, 1–17. doi:10.2514/6.2018-4284

Arnold, G. E., Kappel, D., Moroz, L., Markus, K., and Helbert, J. (2019). "Spaceborne VIR spectroscopy of small planetary bodies and inherent clues to their composition: A review and discussions of future requirements, M. Strojnik, and G. E. Arnold Infrared Remote Sensing and Instrumentation XXVII Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 21 August 2023, USA, (IEEE), 1. doi:10.1117/12.2528287

Bains, W., Petkowski, J. J., Seager, S., Ranjan, S., Sousa-Silva, C., Rimmer, P. B., et al. (2021). Phosphine on Venus cannot Be explained by conventional processes. *Astrobiology* 21, 1277–1304. doi:10.1089/ast.2020.2352

Bains, W., Petkowski, J. J., Seager, S., Ranjan, S., Sousa-Silva, C., Rimmer, P. B., et al. (2022). Venusian phosphine: A 'wow!' signal in chemistry? *Phosphorus, Sulfur Silicon Relat. Elem.* 197, 438–443. doi:10.1080/10426507.2021.1998051

Bannister, M. T., Bhandare, A., Dybczyński, P. A., Fitzsimmons, A., Guilbert-Lepoutre, A., Jedicke, R., et al. (2019). The natural history of 'Oumuamua. *Nat. Astron.* 3, 594–602. doi:10.1038/s41550-019-0816-x

Bennett, C. J., Pirim, C., and Orlando, T. M. (2013). Space-weathering of solar system bodies: A laboratory perspective. *Chem. Rev.* 113, 9086–9150. doi:10.1021/cr400153k

Berezhnoy, A. A., Borovička, J., Santos, J., Rivas-Silva, J. F., Sandoval, L., Stolyarov, A. V., et al. (2018). The CaO orange system in meteor spectra. *Planet. Space Sci.* 151, 27–32. doi:10.1016/j.pss.2017.10.007

Berthe, L., Bezaeva, N. S., Gattaceca, J., Boustie, M., de Resseguier, T., and Rochette, P. (2011). Behavior of basalt under laser-induced shock-wave application to the planetary hypervelocity impact effect. *J. Laser Appl.* 23. doi:10.2351/1.3556591

Bizzarri, B. M., Šponer, J. E., Šponer, J., Cassone, G., Kapralov, M., Timoshenko, G. N., et al. (2020). Meteorite-assisted phosphorylation of adenosine under proton irradiation conditions. *ChemSystemsChem* 2, e1900039. doi:10.1002/syst.201900039

Blank, J. G., Miller, G. H., Ahrens, M. J., and Winans, R. E. (2001). Experimental shock chemistry of aqueous amino acid solutions and the cometary delivery of prebiotic compounds. *Orig. Life Evol. Biosph.* 31, 15–51. doi:10.1023/A:1006758803255

Bonomo, A. S., Zeng, L., Damasso, M., Leinhardt, Z. M., Justesen, A. B., Lopez, E., et al. (2019). A giant impact as the likely origin of different twins in the Kepler-107 exoplanet system. *Nat. Astron.* 3, 416–423. doi:10.1038/s41550-018-0684-9

Borovička, J., and Berezhnoy, A. A. (2016). Radiation of molecules in Benešov bolide spectra. *Icarus* 278, 248–265. doi:10.1016/j.icarus.2016.06.022

Borovička, J., and Betlem, H. (1997). Spectral analysis of two Perseid meteors. *Planet. Space Sci.* 45, 563–575. doi:10.1016/s0032-0633(97)00024-x

Borovička, J. (1994). Two components in meteor spectra. *Planet. Space Sci.* 42, 145–150. doi:10.1016/0032-0633(94)90025-6

Borucki, W. J., Giver, L. P., McKay, C. P., Scattergood, T., and Parris, J. E. (1988). Lightning production of hydrocarbons and HCN on titan: laboratory measurements. *Icarus* 76, 125–134. doi:10.1016/0019-1035(88)90145-5

Borucki, W. J., McKay, C. P., Jebens, D., Lakkaraju, H. S., and Vanajakshi, C. T. (1996). Spectral irradiance measurements of simulated lightning in planetary atmospheres. *Icarus* 123, 336–344. doi:10.1006/icar.1996.0162

Borucki, W. J., and McKay, C. P. (1987). Optical efficiencies of lightning in planetary atmospheres. *Nature* 328, 509–510. doi:10.1038/328509a0

Borucki, W., Mc Kenzie, R., McKay, C., Duong, N., and Boac, D. (1985). Spectra of simulated lightning on Venus, jupiter, and titan. *Icarus* 64, 221–232. doi:10.1016/0019-1035(85)90087-9

Boustie, M., Berthe, L., Resseguier, T. de, and Arrigoni, M. (2008). Laser shock waves: fundamentals and applications, 1st International Symposium on Laser Ultrasonics: Science, Technology and Applications, July 16-18 2008, Montreal. IEEE, 1–6.

Brasser, R., Mojzsis, S. J., Werner, S. C., Matsumura, S., and Ida, S. (2016). Late veneer and late accretion to the terrestrial planets. *Earth Planet. Sci. Lett.* 455, 85–93. doi:10.1016/j.epsl.2016.09.013

Brown, D., Webster, G., Zubritsky, E. A., and Jones, N. N. (2014). Mars spacecraft reveal comet flyby effects on martian atmosphere. Nasa press release. Available at: 2https://www.nasa.gov/press/2014/november/mars-spacecraft-reveal-comet-flyby-effects-on-martian-atmosphere/#:~:text=Dustfromthecometimpacted,andpossiblelonger-termperturbations.

Brunetto, R., Romano, F., Blanco, A., Fonti, S., Martino, M., Orofino, V., et al. (2006). Space weathering of silicates simulated by nanosecond pulse UV excimer laser. *Icarus* 180, 546–554. doi:10.1016/j.icarus.2005.10.016

Carbary, J. F., Morrison, D., Romick, G. J., and Yee, J. H. (2003). Leonid meteor spectrum from 110 to 860 nm. *Icarus* 161, 223–234. doi:10.1016/S0019-1035(02)00064-7

Ceplecha, Z., Borovička, J., Elford, W. G., Revelle, D. O., Hawkes, R. L., Porubčan, V., et al. (1998). Meteor phenomena and bodies. *Space Sci. Rev.* 84, 327–471. doi:10.1023/a:1005069928850

Charbonneau, D., Brown, T. M., Noyes, R. W., and Gilliland, R. L. (2002). Detection of an extrasolar planet atmosphere. *Astrophys. J.* 568, 377–384. doi:10.1086/338770

Chen, H., Rambaux, N., and Vaubaillon, J. (2020). Accuracy of meteor positioning from space- and ground-based observations. *Astron. Astrophys.* 642, L11. doi:10.1051/0004-6361/202039014

Cheng, A. F., Rivkin, A. S., Michel, P., Atchison, J., Barnouin, O., Benner, L., et al. (2018). AIDA DART asteroid deflection test: planetary defense and science objectives. *Planet. Space Sci.* 157, 104–115. doi:10.1016/j.pss.2018.02.015

Christou, A. A., Oberst, J., Elgner, S., Flohrer, J., Margonis, A., McAuliffe, J. P., et al. (2012). Orbital observations of meteors in the Martian atmosphere using the SPOSH camera. *Planet. Space Sci.* 60, 229–235. doi:10.1016/j.pss.2011.09.002

Cipriano, A. M., Dei Tos, D. A., and Topputo, F. (2018). Orbit design for LUMIO: the lunar meteoroid impacts observer. *Front. Astron. Sp. Sci.* 5. doi:10.3389/fspas.2018.00029

Civiš, S., Babánková, D., Cihelka, J., Sazama, P., and Juha, L. (2008). Spectroscopic investigations of high-power laser-induced dielectric breakdown in gas mixtures containing carbon monoxide. *J. Phys. Chem. A* 112, 7162–7169. doi:10.1021/jp712011t

Civiš, S., Knížek, A., Ivanek, O., Kubelík, P., Zukalová, M., Kavan, L., et al. (2017). The origin of methane and biomolecules from a CO2 cycle on terrestrial planets. *Nat. Astron.* 1, 721–726. doi:10.1038/s41550-017-0260-8

Civis, S., Szabla, R., Szyja, B. M., Smykowski, D., Ivanek, O., Knizek, A., et al. (2016). Correction: corrigendum: tiO2-catalyzed synthesis of sugars from formaldehyde in extraterrestrial impacts on the early earth. *Sci. Rep.* 2016, 27962. doi:10.1038/srep27962

Clegg, S. M., Wiens, R. C., Anderson, R., Forni, O., Frydenvang, J., Lasue, J., et al. (2017). Recalibration of the mars science laboratory ChemCam instrument with an expanded geochemical database. *Spectrochim. Acta - Part B At. Spectrosc.* 129, 64–85. doi:10.1016/j.sab.2016.12.003

Collins, G. S., Melosh, H. J., and Marcus, R. A. (2005). Earth impact effects Program: A web-based computer program for calculating the regional environmental consequences of a meteoroid impact on earth. *Meteorit. Planet. Sci.* 40, 817–840. doi:10.1111/j.1945-5100.2005.tb00157.x

Coradini, M., Giblin, I., Martelli, G., Mottola, S., Smith, P. N., and Woodward, A. J. (1994). Spectral signature of satellite fragments re-entering the Earth's atmosphere: A laboratory simulation. *Planet. Space Sci.* 42, 441–446. doi:10.1016/0032-0633(94)00125-1

Czech Ministry of Transportation, C. (2021). *The first Czech space mission will explore mining possibilities on asteroids*. Prague: Czech Ministry of Transportation. Available at: https://www.mdcr.cz/Media/Media-a-tiskove-zpravy/Prvni-ceska-kosmicka-mise-bude-zkoumat-moznosti-te?lang=en-GB.

De Giacomo, A. (2011). A novel approach to elemental analysis by Laser Induced Breakdown Spectroscopy based on direct correlation between the electron impact excitation cross section and the optical emission intensity. *Spectrochim. Acta - Part B At. Spectrosc.* 66, 661–670. doi:10.1016/j.sab.2011.09.003

De Niem, D., Kührt, E., Morbidelli, A., and Motschmann, U. (2012). Atmospheric erosion and replenishment induced by impacts upon the earth and mars during a heavy bombardment. *Icarus* 221, 495–507. doi:10.1016/j.icarus.2012.07.032

Domingue, D. L., Chapman, C. R., Killen, R. M., Zurbuchen, T. H., Gilbert, J. A., Sarantos, M., et al. (2014). Mercury's weather-beaten surface: understanding mercury in the context of lunar and asteroidal space weathering studies. *Space Sci. Rev.* 181, 121–214. doi:10.1007/s11214-014-0039-5

Dremin, A. N. (1989). Shock wave chemistry. *High. Press. Res.* 1, 361–364. doi:10.1080/08957958908202498

Drouard, A., Vernazza, P., Loehle, S., Gattacceca, J., Vaubaillon, J., Zanda, B., et al. (2018). Probing the use of spectroscopy to determine the meteoritic analogues of meteors. *Astron. Astrophys.* 613, A54. doi:10.1051/0004-6361/201732225

Ebert, M., Hecht, L., Hamann, C., and Luther, R. (2017). Laser-induced melting experiments: simulation of short-term high-temperature impact processes. *Meteorit. Planet. Sci.* 52, 1475–1494. doi:10.1111/maps.12809

Elvis, M. (2012). Let's mine asteroids - for science and profit. *Nature* 485, 549. doi:10.1038/485549a

Fazio, A., Harries, D., Matthaeus, G., Mutschke, H., Nolte, S., and Langenhorst, F. (2018). Femtosecond laser irradiation of olivine single crystals: experimental simulation of space weathering. *Icarus* 299, 240–252. doi:10.1016/j.icarus.2017.07.025

Ferus, M., Heays, A. N., and Knížek, A. (2021). "Chapter 12: consequences of heavy bombardment on prebiotic synthesis," in *Comprehensive series in photochemical and photobiological sciences* (China: The Royal Society of Chemistry), 239–264. doi:10.1039/9781839164354-00239

Ferus, M., Koukal, J., Lenža, L., Srba, J., Kubelík, P., Laitl, V., et al. (2018). Calibration-free quantitative elemental analysis of meteor plasma using reference laser-induced breakdown spectroscopy of meteorite samples. *Astron. Astrophys.* 610, A73. doi:10.1051/0004-6361/201629950

Ferus, M., Kubelík, P., Knížek, A., Pastorek, A., Sutherland, J., and Civiš, S. (2017a). High energy radical chemistry formation of HCN-rich atmospheres on early earth. *Sci. Rep.* 7, 6275. doi:10.1038/s41598-017-06489-1

Ferus, M., Kubelík, P., Petera, L., Lenža, L., Koukal, J., Křivková, A., et al. (2019). Main spectral features of meteors studied using a terawatt-class high-power laser. *Astron. Astrophys.* 630, A127. doi:10.1051/0004-6361/201935816

Ferus, M., Michalčíková, R., Shestivská, V., Šponer, J., Šponer, J. E., and Civiš, S. (2014a). High-energy chemistry of formamide: A simpler way for nucleobase formation. *J. Phys. Chem. A* 118, 719–736. doi:10.1021/jp411415p

Ferus, M., Nesvorný, D., Šponer, J., Kubelík, P., Michalčíková, R., Shestivská, V., et al. (2014b). High-energy chemistry of formamide: A unified mechanism of nucleobase formation. *Proc. Natl. Acad. Sci.* 112, 657–662. doi:10.1073/pnas.1412072111

Ferus, M., Petera, L., Koukal, J., Lenža, L., Drtinová, B., Haloda, J., et al. (2020). Elemental composition, mineralogy and orbital parameters of the Porangaba meteorite. *Icarus* 341, 113670. doi:10.1016/j.icarus.2020.113670

Ferus, M., Pietrucci, F., Saitta, A. M., Knížek, A., Kubelík, P., Ivanek, O., et al. (2017b). Formation of nucleobases in a Miller-Urey reducing atmosphere. *Proc. Natl. Acad. Sci. U. S. A.* 114, 4306–4311. doi:10.1073/pnas.1700010114

Fulvio, D., Fuks Maron, L., Cires Perez, Y., Tahir, H., and Del Rosso, T. (2021). Micrometeorite bombardment simulated by ns-pulsed laser ablation: morphological characterization of the impact craters. *Icarus* 366, 114532. doi:10.1016/j.icarus.2021.114532

Furukawa, Y., Sekine, T., Oba, M., Kakegawa, T., and Nakazawa, H. (2009). Biomolecule formation by oceanic impacts on early Earth. *Nat. Geosci.* 2, 62–66. doi:10.1038/ngeo383

Geiss, J., and Rossi, A. P. (2013). On the chronology of lunar origin and evolution: implications for earth, mars and the solar system as a whole. *Astron. Astrophys. Rev.* 21, 68–54. doi:10.1007/s00159-013-0068-1

Genda, H., Brasser, R., and Mojzsis, S. J. (2017). The terrestrial late veneer from core disruption of a lunar-sized impactor. *Earth Planet. Sci. Lett.* 480, 25–32. doi:10.1016/j.epsl.2017.09.041

Goldman, N., Reed, E. J., Fried, L. E., William Kuo, I. F., and Maiti, A. (2010). Synthesis of glycine-containing complexes in impacts of comets on early Earth. *Nat. Chem.* 2, 949–954. doi:10.1038/nchem.827

Guilluy, G., Sozzetti, A., Giacobbe, P., Bonomo, A. S., and Micela, G. (2022). On the synergy between Ariel and ground-based high-resolution spectroscopy. *Exp. Astron.*53, 655–677. doi:10.1007/s10686-021-09824-7

Gusarov, A. V., and Smurov, I. (2005). Thermal model of nanosecond pulsed laser ablation: analysis of energy and mass transfer. *J. Appl. Phys.* 97, 14307–1430713. doi:10.1063/1.1827321

Haefner, C. L., Bayramian, A., Betts, S., Bopp, R., Buck, S., Cupal, J., et al. (2017). High average power, diode pumped petawatt laser systems: A new generation of lasers enabling precision science and commercial applications. *Res. Using Extreme Light Enter. New Front. Petawatt-Class Lasers* III, 1024102. doi:10.1117/12.2281050

Hapke, B., Cassidy, W., and Wells, E. (1975). Effects of vapor-phase deposition processes on the optical, chemical, and magnetic properties OE the lunar regolith. *Moon* 13, 339–353. doi:10.1007/BF00567525

Hapke, B. (2001). Space weathering from Mercury to the asteroid belt. *J. Geophys. Res. Planets* 106, 10039–10073. doi:10.1029/2000JE001338

Hawkes, R. L., Milley, E. P., Ehrman, J. M., Woods, R. M., Hoyland, J. D., Pettipas, C. L., et al. (2008). "What can we learn about atmospheric meteor ablation and light production from laser ablation?," in *Earth, Moon and planets (VAN GODEWIJCKSTRAAT 30, 3311 GZ* (DORDRECHT, NETHERLANDS: Springer), 331–336. doi:10.1007/s11038-007-9186-y

Heays, A. N., Kaiserová, T., Rimmer, P. B., Knížek, A., Petera, L., Civiš, S., et al. (2022). Nitrogen oxide production in laser-induced breakdown simulating impacts on the hadean atmosphere. *J. Geophys. Res. Planets* 127, e2021JE006842. doi:10.1029/2021JE006842

Helber, B., Dias, B., Bariselli, F., Zavalan, L. F., Pittarello, L., Goderis, S., et al. (2019). Analysis of meteoroid ablation based on plasma wind-tunnel experiments, surface characterization, and numerical simulations. *Astrophys. J.* 876, 120. doi:10.3847/1538-4357/ab16f0

Hibbert, R., Cole, M. J., Price, M. C., and Burchell, M. J. (2017). The hypervelocity impact facility at the university of kent: recent upgrades and specialized capabilities. *Procedia Eng.* 204, 208–214. doi:10.1016/j.proeng.2017.09.775

Hudson, J., and Lemmer, K. (2017). "Plasma spectroscopy CubeSat: A demonstration of on-orbit electric propulsion system diagnostics,". Editor S. Ryan Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference, September 30, 2022, Germany, (IEEE), 28.

Ishimaru, R., Sakamoto, Y., Kobayashi, M., Fujita, S., Gonai, T., Senshu, H., et al. (2014). CubeSat mission for UV-visible observations of meteors from space: s-CUBE (S3: shootingstar sensing satellite). *45th Lunar Planet. Sci. Conf. Contrib.* 1846, 1.

Jayaram, V., Singh, P., and Reddy, K. P. J. (2013). Study of anatase TiO2 in the presence of N2 under shock dynamic loading in a free piston driven shock tube. *Adv. Ceram. Sci. Eng.* 2, 40–46. Available at: www.acse-journal.org.

Jebens, D. S., Lakkaraju, H. S., McKay, C. P., and Borucki, W. J. (1992). Time resolved simulation of lightning by LIP. *Geophys. Res. Lett.* 19, 273–276. doi:10.1029/91GL02937

Jenniskens, P., Laux, C. O., Wilson, M. A., and Schaller, E. L. (2004). The mass and speed dependence of meteor air plasma temperatures. *Astrobiology* 4, 81–94. doi:10.1089/153110704773600258

Jenniskens, P., Nugent, D., Tedesco, E., and Murthy, J. (2000). "Leonid shower from space," in *Leonid storm research*. Editors P. Jenniskens, F. Rietmeijer, N. Brosch, and M. Fonda (Dordrecht: Springer Netherlands), 305–312. doi:10.1007/978-94-017-2071-7_23

Jenniskens, P. (2007). Quantitative meteor spectroscopy: elemental abundances. *Adv. Sp. Res.* 39, 491–512. doi:10.1016/j.asr.2007.03.040

Jenniskens, P., and Stenbaek-Nielsen, H. C. (2004). Meteor wake in high frame-rate images - implications for the chemistry of ablated organic compounds. *Astrobiology* 4, 95–108. doi:10.1089/153110704773600267

Jiang, T., Zhang, H., Yang, Y., Hu, X., Ma, P., Sun, Y., et al. (2019). Bi-directional reflectance and polarization measurements of pulse-laser irradiated airless body analog materials. *Icarus* 331, 127–147. doi:10.1016/j.icarus.2019.05.022

Johnston, C. O., Stern, E. C., and Wheeler, L. F. (2018). Radiative heating of large meteoroids during atmospheric entry. *Icarus* 309, 25–44. doi:10.1016/j.icarus.2018.02.026

Kabáth, P., Žák, J., Boffin, H. M. J., Ivanov, V. D., Jones, D., and Skarka, M. (2019). Detection limits of exoplanetary atmospheres with 2-m class telescopes. *Publ. Astron. Soc. Pac.* 131, 085001. doi:10.1088/1538-3873/ab2143

Kaltenegger, L., Lin, Z., and Madden, J. (2020). High-resolution transmission spectra of earth through geological time. *Astrophys. J.* 892, L17. doi:10.3847/2041-8213/ab789f

Kaluna, H. M., Ishii, H. A., Bradley, J. P., Gillis-Davis, J. J., and Lucey, P. G. (2017). Simulated space weathering of Fe- and Mg-rich aqueously altered minerals using pulsed laser irradiation. *Icarus* 292, 245–258. doi:10.1016/j.icarus.2016.12.028

Kasuga, T., Yamamoto, T., Watanabe, J., Ebizuka, N., Kawakita, H., and Yano, H. (2005). Metallic abundances of the 2002 Leonid meteor deduced from high-definition TV spectra. *Astron. Astrophys.* 435, 341–351. doi:10.1051/0004-6361:20041428

Koeberl, C., Reimold, W. U., McDonald, I., and Rosing, M. (2005). "Search for petrographic and geochemical evidence for the late heavy bombardment on earth in early archean rocks from Isua, Greenland," in *Impacts and the early Earth lecture notes in Earth sciences*. Editors I. Gilmour, and C. Koeberl (Berlin, Germany: Springer-Verlag Berlin), 73–97. doi:10.1007/bfb0027757

Koeberl, C. (2006). The record of impact processes on the early earth: A review of the first 2.5 billion years. *Spec. Pap. Geol. Soc. Am.* 405, 1–22. doi:10.1130/2006.2405(01

Kramer, H. J. (1994). *Observation of the Earth and its environment*. doi:10.1007/978-3-662-09038-1

Křivková, A., Laitl, V., Chatzitheodoridis, E., Petera, L., Kubelík, P., Knížek, A., et al. (2022). Morphology of meteorite surfaces ablated by high-power lasers: review and applications. *Appl. Sci.* 12, 4869. doi:10.3390/app12104869

Křivková, A., Petera, L., Laitl, V., Kubelík, P., Chatzitheodoridis, E., Lenža, L., et al. (2021). Application of a dielectric breakdown induced by high-power lasers for a laboratory simulation of meteor plasma. *Exp. Astron.* 180, 425–451. doi:10.1007/s10686-020-09688-3

Kubelík, P., Koukal, J., Lenža, L., Srba, J., Laitl, V., Křížová, R., et al. (2022). Probing plasma physics and elemental composition of a leonid meteor by fitting complex plasma radiation model parameters. *Mon. Not. R. Astron. Soc.* 514, 5266–5275. doi:10.1093/mnras/stac1600

Kurahashi, E., Yamanaka, C., Nakamura, K., and Sasaki, S. (2002). Laboratory simulation of space weathering: ESR measurements of nanophase metallic iron in laser-irradiated materials. *Earth, Planets Sp.* 54, E5–e7. doi:10.1186/BF03352448

Kuwahara, H., and Sugita, S. (2015). The molecular composition of impact-generated atmospheres on terrestrial planets during the post-accretion stage. *Icarus* 257, 290–301. doi:10.1016/j.icarus.2015.05.007

Langford, L., Allgood, D., and Junell, J. (2007). Determination of combustion product radicals in a hydrocarbon fueled rocket exhaust plume. *Stennis Space Cent.*, 1–6.Available at: https://ntrs.nasa.gov/citations/20070010432.

Lasue, J., Wiens, R. C., Clegg, S. M., Vaniman, D. T., Joy, K. H., Humphries, S., et al. (2012). Remote laser-induced breakdown spectroscopy (LIBS) for lunar exploration. *J. Geophys. Res. Planets* 117. doi:10.1029/2011JE003898

Lazzarin, M., Marchi, S., Moroz, L. V., Brunetto, R., Magrin, S., Paolicchi, P., et al. (2006). Space weathering in the main asteroid belt: the big picture. *Astrophys. J.* 647, L179–L182. doi:10.1086/507448

Lee, J. A., Park, S. Y., Kim, Y., Bae, J., Lee, D., and Ju, G. (2017). Mission orbit design of CubeSat impactor measuring lunar local magnetic field. *J. Astron. Sp. Sci.* 34, 127–138. doi:10.5140/JASS.2017.34.2.127

Leiser, D., Loehle, S., and Fasoulas, S. (2022). Spectral features for reentry breakup event identification. *J. Spacecr. Rockets* 59, 1496–1506. doi:10.2514/1.A35258

Levy, D. H. (1998). The collision of comet shoemaker-levy 9 with jupiter. *Space Sci. Rev.* 85, 523–545. doi:10.1023/a:1005079807445

Libourel, G., Nakamura, A. M., Beck, P., Potin, S., Ganino, C., Jacomet, S., et al. (2019). Hypervelocity impacts as a source of deceiving surface signatures on iron-rich asteroids. *Sci. Adv.* 5, eaav3971. doi:10.1126/sciadv.aav3971

Loeffler, M. J., Baragiola, R. A., and Murayama, M. (2008). Laboratory simulations of redeposition of impact ejecta on mineral surfaces. *Icarus* 196, 285–292. doi:10.1016/j.icarus.2008.02.021

Loehle, S., Eberhart, M., Zander, F., Meindl, A., Rudawska, R., Koschny, D., et al. (2021). Extension of the plasma radiation database PARADE for the analysis of meteor spectra. *Meteorit. Planet. Sci.* 56, 352–361. doi:10.1111/maps.13622

Loehle, S., Zander, F., Eberhart, M., Hermann, T., Meindl, A., Massuti-Ballester, B., et al. (2022). Assessment of high enthalpy flow conditions for re-entry aerothermodynamics in the plasma wind tunnel facilities at IRS. *CEAS Sp. J.* 14, 395–406. doi:10.1007/s12567-021-00396-y

Loehle, S., Zander, F., Hermann, T., Eberhart, M., Meindl, A., Oefele, R., et al. (2017a). Experimental simulation of meteorite ablation during earth entry using a plasma wind tunnel. *Astrophys. J.* 837, 112. doi:10.3847/1538-4357/aa5cb5

Loehle, S., Zander, F., Lemmens, S., and Krag, H. (2017b). Airborne observations of re-entering space debris - results and prospects, 7th European Conference on Space Debris 18–21 April Germany, 1–4. IEEE.

Madiedo, J. M., Ortiz, J. L., Trigo-Rodríguez, J. M., Dergham, J., Castro-Tirado, A. J., Cabrera-Caño, J., et al. (2014). Analysis of bright Taurid fireballs and their ability to produce meteorites. *Icarus* 231, 356–364. doi:10.1016/j.icarus.2013.12.025

Managadze, G. G., Brinckerhoff, W. B., and Chumikov, A. E. (2003). Molecular synthesis in hypervelocity impact plasmas on the primitive Earth and in interstellar clouds. *Geophys. Res. Lett.* 30. doi:10.1029/2002gl016422

Markley, M. ~M., Fuller, M. ~D., and Kletetschka, G. (2013). Magnetic scanning of iron blebs in laser irradiated olivine grains. *Meteorit. Planet. Sci. Suppl.* 76, 5330.

Martin, A., and Boyd, I. D. (2012). Modeling of heat transfer attenuation by ablative gases during the Stardust re-entry. in 50th *AIAA aerospace sciences meeting including the new horizons forum and aerospace exposition* doi:10.2514/6.2012-814

Martin, R. S., Mather, T. A., and Pyle, D. M. (2007). Volcanic emissions and the early Earth atmosphere. *Geochim. Cosmochim. Acta* 71, 3673–3685. doi:10.1016/j.gca.2007.04.035

Martins, Z., Price, M. C., Goldman, N., Sephton, M. A., and Burchell, M. J. (2013). Shock synthesis of amino acids from impacting cometary and icy planet surface analogues. *Nat. Geosci.* 6, 1045–1049. doi:10.1038/ngeo1930

Mason, P., Divoký, M., Ertel, K., Pilař, J., Butcher, T., Hanuš, M., et al. (2017). Kilowatt average power 100 J-level diode pumped solid state laser. *Optica* 4, 438. doi:10.1364/optica.4.000438

Matsuoka, M., Nakamura, T., Hiroi, T., Okumura, S., and Sasaki, S. (2020). Space weathering simulation with low-energy laser irradiation of murchison CM chondrite for reproducing micrometeoroid bombardments on C-type asteroids. *Astrophys. J. Lett.* 890, L23. doi:10.3847/2041-8213/ab72a4

McKay, C. P., and Borucki, W. J. (1997). Organic synthesis in experimental impact shocks. *Sci. (80-* 276, 390–392. doi:10.1126/science.276.5311.390

Milley, E. P., Hawkes, R. L., and Ehrman, J. M. (2007). Meteor luminosity simulation through laser ablation of meteorites. *Mon. Not. R. Astron. Soc. Lett.* 382, 67–71. doi:10.1111/j.1745-3933.2007.00390.x

Morbidelli, A., Nesvorny, D., Laurenz, V., Marchi, S., Rubie, D. C., Elkins-Tanton, L., et al. (2018). The timeline of the lunar bombardment: revisited. *Icarus* 305, 262–276. doi:10.1016/j.icarus.2017.12.046

Moroz, L. V., Fisenko, A. V., Semjonova, L. F., Pieters, C. M., and Korotaeva, N. N. (1996). Optical effects of regolith processes on S-asteroids as simulated by laser shots on ordinary chondrite and other mafic materials. *Icarus* 122, 366–382. doi:10.1006/icar.1996.0130

Mueller, R. P., and Van Susante, P. J. (2012).A review of extra-terrestrial mining robot concepts, Earth and Space 2012 - Proceedings of the 13th ASCE Aerospace Division Conference and the 5th NASA/ASCE Workshop on Granular Materials in Space Exploration, APRIL 15-18, USA. IEEE, 295–314. doi:10.1061/9780784412190.034

Napier, B., and Asher, D. (2009). The tunguska impact event and beyond. *Astron. Geophys.* 50, 1.18–1.26. doi:10.1111/j.1468-4004.2009.50118.x

NASA, Mohon, L., and Dunbar, B. (2022). About lunar impact monitoring. *Lunar Impacts*. Available at: http://www.nasa.gov/centers/marshall/news/lunar/overview.html%0A (Accessed May 15, 2020).

Navarro, K. F., Urrutia-Fucugauchi, J., Villagran-Muniz, M., Sánchez-Aké, C., Pi-Puig, T., Pérez-Cruz, L., et al. (2020). Emission spectra of a simulated Chicxulub impact-vapor plume at the Cretaceous–Paleogene boundary. *Icarus* 346, 113813. doi:10.1016/j.icarus.2020.113813

Navarro-González, R., McKay, C. P., and Mvondo, D. N. (2001a). A possible nitrogen crisis for Archaean life due to reduced nitrogen fixation by lightning. *Nature* 412, 61–64. doi:10.1038/35083537

Navarro-González, R., Navarro, K. F., Coll, P., McKay, C. P., Stern, J. C., Sutter, B., et al. (2019). Abiotic input of fixed nitrogen by bolide impacts to gale crater during the hesperian: insights from the mars science laboratory. *J. Geophys. Res. Planets* 124, 2018JE005852–113. doi:10.1029/2018JE005852

Navarro-González, R., Villagrán-Muniz, M., Sobral, H., Molina, L. T., and Molina, M. J. (2001b). The physical mechanism of nitric oxide formation in simulated lightning. *Geophys. Res. Lett.* 28, 3867–3870. doi:10.1029/2001GL013170

Nuth, J. A., Lowrance, J. L., and Carruthers, G. R. (2008). Neocam: the near earth object chemical analysis mission. *Earth, Moon Planets* 102, 495–504. doi:10.1007/s11038-007-9178-y

Omran, A., Oze, C., Jackson, B., Mehta, C., Barge, L. M., Bada, J., et al. (2021). Phosphine generation pathways on rocky planets. *Astrobiology* 21, 1264–1276. doi:10.1089/ast.2021.0034

Oppenheim, M. M., and Dimant, Y. S. (2015). First 3-D simulations of meteor plasma dynamics and turbulence. *Geophys. Res. Lett.* 42, 681–687. doi:10.1002/2014GL062411

Paine, M. (2006). Can we detect asteroid impacts with rocky extrasolar planets? *Sp. Rev.* Available at: https://www.thespacereview.com/article/761/1.

Perez, D., and Lewis, L. J. (2003). Molecular-dynamics study of ablation of solids under femtosecond laser pulses. *Phys. Rev. B* 67, 184102. doi:10.1103/PhysRevB.67.184102

Pérez-Invernón, F. J., Gordillo-Vázquez, F. J., Passas-Varo, M., Neubert, T., Chanrion, O., Reglero, V., et al. (2022). Multispectral optical diagnostics of lightning from space. *Remote Sens.* 14, 2057. doi:10.3390/rs14092057

Petera, L., Knížek, A., Laitl, V., and Ferus, M. (2023). Decomposition of benzene during impacts in N 2 -dominated atmospheres. *Astrophys. J.* 945, 149. doi:10.3847/1538-4357/acbd48

Petri, J. (2020). "Optimizing the scientific output of satellite formation for a stereoscopic meteor observation,". Editors U. Pajer, J. Rendtel, M. Gyssens, and C. Verbeeck Proceedings of the International Meteor Conference, 29-October 2, China, (IEEE), 119–125.

Petri, J., Schmidt, A., Zink, J., and Klinker, S. (2019). Design and test of a COTS based imaging system for stereo-scopic meteor observations. *IAA Symp.*, 1.

Pieters, C. M., and Noble, S. K. (2016). Space weathering on airless bodies. *J. Geophys. Res. Planets* 121, 1865–1884. doi:10.1002/2016JE005128

Pirri, A. N. (1977). Theory for laser simulation of hypervelocity impact. *Phys. Fluids* 20, 221–228. doi:10.1063/1.861859

Popov, A. M., Berezhnoy, A. A., Borovička, J., Labutin, T. A., Zaytsev, S. M., and Stolyarov, A. V. (2021). Tackling the FeO orange band puzzle in meteor and airglow spectra through combined astronomical and laboratory studies. *Mon. Not. R. Astron. Soc.* 500, 4296–4306. doi:10.1093/mnras/staa3487

Popova, O. (2006). "Meteoroid ablation models," in *Modern meteor science an interdisciplinary view*. Editors R. Hawkes, I. Mann, and P. Brown (Dordrecht: Springer Netherlands), 303–319. doi:10.1007/1-4020-5075-5_32

Qian, L., Zaidi, S. H., and Miles, R. B. (2005). Narrow linewidth potassium imaging filter for near infrared detection of missile plumes. in 43rd *AIAA Aerosp. Sci. Meet. Exhib. - Meet. Pap.*, 14209–14221. doi:10.2514/6.2005-825

Quintana, E. V., Barclay, T., Borucki, W. J., Rowe, J. F., and Chambers, J. E. (2016). The frequency of giant impacts on earth-like worlds. *Astrophys. J.* 821, 126. doi:10.3847/0004-637x/821/2/126

Rae, J. W. J., and Hertzberg, A. (1964). *On the possibility of simulating meteoroid impact by the user of lasers*. Buffalo, NY, USA: Cornell Aeronautical Lab, Inc. Available at: https://scholar.google.com/scholar_lookup?title=On+the+Possibility+of+Simulating+Meteoroid+Impact+by+the+User+of+Lasers&author=Rae,+J.W.J.&author=Hertzberg.

Rambaux, N., Galayko, D., Mariscal, J.-F., Breton, M.-A., Vaubaillon, J., Birlan, M., et al. (2014). "Detection of spectral UV of meteor from a nanosatellite," in *Proceedings of the IMC, giron, 3*. Available at: http://www.imo.net/imcs/imc2014/2014-51-rambaux-final.pdf.

Rambaux, N., Vaubaillon, J., Derelle, S., Jacquart, M., Millet, M., Lacassagne, L., et al. (2021). Meteorix camera tests for space-based meteor observations. *WGN, J. Int. Meteor Organ.* 49, 142–144.

Redfield, S., Endl, M., Cochran, W. D., and Koesterke, L. (2008). Sodium absorption from the exoplanetary atmosphere of HD 189733b detected in the optical transmission spectrum. *Astrophys. J.* 673, L87–L90. doi:10.1086/527475

Reimold, W. U., and Gibson, R. L. (2006). "Processes on the early earth," in *Gibson 3635 CONCORDE PKWY STE 500*. Editor W. U. Reimold (CHANTILLY, VA 20151-1125 USA: Geological Society of America). doi:10.1130/spe405

Rimmer, P. B., Ferus, M., Waldmann, I. P., Knížek, A., Kalvaitis, D., Ivanek, O., et al. (2019). Identifiable acetylene features predicted for young earth-like exoplanets with reducing atmospheres undergoing heavy bombardment. *Astrophys. J.* 888, 21. doi:10.3847/1538-4357/ab55e8

Röcker, C., Loescher, A., Bienert, F., Villeval, P., Lupinski, D., Bauer, D., et al. (2020). Ultrafast green thin-disk laser exceeding 1.4 kW of average power. *Opt. Lett.* 45, 5522. doi:10.1364/ol.403781

Rogers, L. A., Hill, K. A., and Hawkes, R. L. (2005). Mass loss due to sputtering and thermal processes in meteoroid ablation. *Planet. Space Sci.* 53, 1341–1354. doi:10.1016/j.pss.2005.07.002

Rotelli, L., Trigo-Rodríguez, J. M., Moyano-Cambero, C. E., Carota, E., Botta, L., Di Mauro, E., et al. (2016). The key role of meteorites in the formation of relevant prebiotic molecules in a formamide/water environment. *Sci. Rep.* 6, 38888. doi:10.1038/srep38888

Roy, A., Singh, S. V., Ambresh, M., Sahu, D., Meka, J. K., Ramachandran, R., et al. (2022). Shock processing of amorphous carbon nanodust. *Adv. Sp. Res.* 70, 2571–2581. doi:10.1016/j.asr.2022.06.068

Rubin, M., Bekaert, D. V., Broadley, M. W., Drozdovskaya, M. N., and Wampfler, S. F. (2019). Volatile species in comet 67P/Churyumov-Gerasimenko: investigating the link from the ISM to the terrestrial planets. *ACS Earth Sp. Chem.* 3, 1792–1811. doi:10.1021/acsearthspacechem.9b00096

Saeidfirozeh, H., Myakalwar, A. K., Kubelík, P., Ghaderi, A., Laitl, V., Petera, L., et al. (2022). ANN-LIBS analysis of mixture plasmas: detection of xenon. *J. Anal. At. Spectrom.* 37, 1815–1823. doi:10.1039/d2ja00132b

Sasaki, S., Hiroi, T., Nakamura, K., Hamabe, Y., Kurahashi, E., and Yamada, M. (2002). Simulation of space weathering by nanosecond pulse laser heating: dependence on mineral composition, weathering trend of asteroids and discovery of nanophase iron particles. *Adv. Sp. Res.* 29, 783–788. doi:10.1016/S0273-1177(02)00012-1

Sasaki, S., Kurahashi, E., Yamanaka, C., and Nakamura, K. (2003). Laboratory simulation of space weathering: changes of optical properties and TEM/ESR confirmation of nanophase metallic iron. *Adv. Sp. Res.* 31, 2537–2542. doi:10.1016/S0273-1177(03)00575-1

Sasaki, S., Nakamura, K., Hamabe, Y., Kurahashi, E., and Hiroi, T. (2001). Production of iron nanoparticles by laser irradiation in a simulation of lunar-like space weathering. *Nature* 410, 555–557. doi:10.1038/35069013

Scattergood, T. W., McKay, C. P., Borucki, W. J., Giver, L. P., van Ghyseghem, H., Parris, J. E., et al. (1989). Production of organic compounds in plasmas: A comparison among electric sparks, laser-induced plasmas, and UV light. *Icarus* 81, 413–428. doi:10.1016/0019-1035(89)90061-4

Schneider, J. (2018). "Future exoplanet research: science questions and how to address them," in *Handbook of exoplanets*. Editors H. J. Deeg, and J. A. Belmonte (Cham: Springer International Publishing), 3245–3267. doi:10.1007/978-3-319-55333-7_163

Schulze-Makuch, D. (2021). The case (Or not) for life in the venusian clouds. *Life* 11, 255. doi:10.3390/life11030255

Selsis, F., Lemmon, M. T., Vaubaillon, J., and Bell, J. F. (2005). Extraterrestrial meteors: A martian meteor and its parent comet. *Nature* 435, 581. doi:10.1038/435581a

Silber, E. A., Boslough, M., Hocking, W. K., Gritsevich, M., and Whitaker, R. W. (2018). Physics of meteor generated shock waves in the Earth's atmosphere – a review. *Adv. Sp. Res.* 62, 489–532. doi:10.1016/j.asr.2018.05.010

Sinclair, C. A., Wyatt, M. C., Morbidelli, A., and Nesvorný, D. (2020). Evolution of the Earth's atmosphere during late veneer accretion. *Mon. Not. R. Astron. Soc.* 499, 5334–5362. doi:10.1093/mnras/staa3210

Singh, S. V., Vishakantaiah, J., Meka, J. K., Muruganantham, M., Thiruvenkatam, V., Sivaprahasam, V., et al. (2022). Three-dimensional complex architectures observed in shock processed amino acid mixtures. *Exp. Results* 3, e8–e14. doi:10.1017/exp.2021.17

Šmíd, M., Renner, O., Colaitis, A., Tikhonchuk, V. T., Schlegel, T., and Rosmej, F. B. (2019). Characterization of suprathermal electrons inside a laser accelerated plasma via highly-resolved Kα-emission. *Nat. Commun.* 10, 4212. doi:10.1038/s41467-019-12008-9

Sobral, H., Villagrán-Muniz, M., Navarro-González, R., and Camps, E. (2002). Experimental simulation of a double return-stroke lightning flash by lasers. *Geophys. Res. Lett.* 29, 1-1–1-4. doi:10.1029/2002GL015715

Sobral, H., Villagrán-Muniz, M., Navarro-González, R., and Raga, A. C. (2000). Temporal evolution of the shock wave and hot core air in laser induced plasma. *Appl. Phys. Lett.* 77, 3158–3160. doi:10.1063/1.1324986

Sonter, M. J. (1997). The technical and economic feasibility of mining the near-earth asteroids. *Acta Astronaut.* 41, 637–647. doi:10.1016/S0094-5765(98)00087-3

Sponer, J. E., Mohammadi, E., Petera, L., Saeidfirozeh, H., Knížek, A., Kubelík, P., et al. (2020). Formic acid, a ubiquitous but overlooked component of the early Earth atmosphere. *Chem.-A Eur. J.* doi:10.1002/chem.202000323

Stark, M. S., Harrison, J. T. H., and Anastasi, C. (1996). Formation of nitrogen oxides by electrical discharges and implications for atmospheric lightning. *J. Geophys. Res. Atmos.* 101, 6963–6969. doi:10.1029/95JD03008

Svettsov, V. V. (1998). Explosions of meteoroids and estimating their parameters from light emission. *Combust. Explos. Shock Waves* 34, 474–484. doi:10.1007/BF02675619

Tabetah, M. E., and Melosh, H. J. (2018). Air penetration enhances fragmentation of entering meteoroids. *Meteorit. Planet. Sci.* 53, 493–504. doi:10.1111/maps.13034

Tabetah, M. E., and Melosh, H. J. (2017). The role of air penetration in the break-up of entering meteoroids. *Lunar Planet. Sci. 48th Lunar Planet. Sci. Conf. Contrib. No.* 1964, 1267.

Tinetti, G., Drossart, P., Eccleston, P., Hartogh, P., Heske, A., Leconte, J., et al. (2018). A chemical survey of exoplanets with ARIEL. *Exp. Astron.* 46, 135–209. doi:10.1007/s10686-018-9598-x

Todd, Z. R., and Öberg, K. I. (2020). Cometary delivery of hydrogen cyanide to the early earth. *Astrobiology* 20, 1109–1120. doi:10.1089/ast.2019.2187

Truong, N., and Lunine, J. I. (2021). Volcanically extruded phosphides as an abiotic source of Venusian phosphine. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2021689118. doi:10.1073/pnas.2021689118

Villagrán-Muniz, M., Sobral, H., Navarro-González, R., Velázquez, P. F., and Raga, A. C. (2003). Experimental simulation of lightning, interacting explosions and astrophysical jets with pulsed lasers. *Plasma Phys. control. Fusion* 45, 571–584. doi:10.1088/0741-3335/45/5/305

Vojacek, V., Borovička, J., Koten, P., Spurny, P., and Štork, R. (2015). Catalogue of representative meteor spectra. *Astron. Astrophys.* 580, A67. doi:10.1051/0004-6361/201425047

Vondrak, T., Plane, J. M. C., Broadley, S., and Janches, D. (2008). A chemical model of meteoric ablation. *Atmos. Chem. Phys.* 8, 7015–7031. doi:10.5194/acp-8-7015-2008

Wiens, R. C., Maurice, S., Barraclough, B., Saccoccio, M., Barkley, W. C., Bell, J. F., et al. (2012). The ChemCam instrument suite on the mars science laboratory (MSL) rover: body unit and combined system tests. *Space Sci. Rev.* 170, 167–227. doi:10.1007/s11214-012-9902-4

Wooding, E. R. (1972). Laser analogue to ball lightning. *Nature* 239, 394–395. doi:10.1038/239394a0

Worms, J. C. (2003). *Interpretation of the remote and in-situ observations of small bodies*. Committee on Space Research.

Wyttenbach, A., Ehrenreich, D., Lovis, C., Udry, S., and Pepe, F. (2015). Spectrally resolved detection of sodium in the atmosphere of HD 189733b with the HARPS spectrograph. *Astron. Astrophys.* 577, A62. doi:10.1051/0004-6361/201525729

Yamada, M., Sasaki, S., Nagahara, H., Fujiwara, A., Hasegawa, S., Yano, H., et al. (1999). Simulation of space weathering of planet-forming materials: nanosecond pulse laser irradiation and proton implantation on olivine and pyroxene samples. *EARTH PLANETS Sp.* 51, 1255–1265. doi:10.1186/BF03351599

Yan, F., Fosbury, R. A. E., Petr-Gotzens, M. G., Zhao, G., Wang, W., Wang, L., et al. (2015). High-resolution transmission spectrum of the Earth's atmosphere-seeing Earth as an exoplanet using a lunar eclipse. *Int. J. Astrobiol.* 14, 255–266. doi:10.1017/S1473550414000172

Zahnle, K. J., Lupu, R., Catling, D. C., and Wogan, N. (2020). Creation and evolution of impact-generated reduced atmospheres of early earth. *Planet. Sci. J.* 1, 11. doi:10.3847/PSJ/ab7e2c

Zahnle, K., Mac Low, M. -M., Lodders, K., and Fegley, B. (1995). Sulfur chemistry in the wake of comet Shoemaker-Levy 9. *Geophys. Res. Lett.* 22, 1593–1596. doi:10.1029/95GL01190

Zakharov, Y. P. (2003). Collisionless laboratory astrophysics with lasers. *IEEE Trans. Plasma Sci.* 31, 1243–1251. doi:10.1109/TPS.2003.820957

Zakuskin, A. S., Beglaryan, B. G., and Labutin, T. A. (2023). Laboratory modeling in laser-induced plasma to estimate the pressure in bolide wake. *Astron. Astrophys.* 670, L13. doi:10.1051/0004-6361/202245462

Zhang, W., Tang, Y., Shi, A., Bao, L., Shen, Y., Shen, R., et al. (2018). Recent developments in spectroscopic techniques for the detection of explosives. *Mater. (Basel)* 11, 1364. doi:10.3390/ma11081364

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 09 October 2023
DOI 10.3389/fspas.2023.1158213

Check for updates

OPEN ACCESS

EDITED BY
Daniel Hestroffer,
Université Paris Sciences and Lettres PSL, France

REVIEWED BY
Ferréol Soulez,
UMR5574 Centre de recherche astrophysique de Lyon (CRAL), France
Arun Kannawadi,
Princeton University, United States

*CORRESPONDENCE
Tobías I. Liaudat,
tobiasliaudat@gmail.com

RECEIVED 03 February 2023
ACCEPTED 29 August 2023
PUBLISHED 09 October 2023

CITATION
Liaudat TI, Starck J-L and Kilbinger M (2023), Point spread function modelling for astronomical telescopes:a review focused on weak gravitational lensing studies.
Front. Astron. Space Sci. 10:1158213.
doi: 10.3389/fspas.2023.1158213

COPYRIGHT
© 2023 Liaudat, Starck and Kilbinger. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Point spread function modelling for astronomical telescopes: a review focused on weak gravitational lensing studies

Tobías I. Liaudat[1,2,3]*, Jean-Luc Starck[1,4] and Martin Kilbinger[1]

[1]Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, Gif-sur-Yvette, France, [2]Department of Computer Science, University College London (UCL), London, United Kingdom, [3]Mullard Space Science Laboratory, University College London (UCL), Dorking, United Kingdom, [4]Institute of Astrophysics, Foundation for Research and Technology-Hellas (FORTH), Heraklion, Greece

The accurate modelling of the point spread function (PSF) is of paramount importance in astronomical observations, as it allows for the correction of distortions and blurring caused by the telescope and atmosphere. PSF modelling is crucial for accurately measuring celestial objects' properties. The last decades have brought us a steady increase in the power and complexity of astronomical telescopes and instruments. Upcoming galaxy surveys like *Euclid* and Legacy Survey of Space and Time (LSST) will observe an unprecedented amount and quality of data. Modelling the PSF for these new facilities and surveys requires novel modelling techniques that can cope with the ever-tightening error requirements. The purpose of this review is threefold. Firstly, we introduce the optical background required for a more physically motivated PSF modelling and propose an observational model that can be reused for future developments. Secondly, we provide an overview of the different physical contributors of the PSF, which includes the optic- and detector-level contributors and atmosphere. We expect that the overview will help better understand the modelled effects. Thirdly, we discuss the different methods for PSF modelling from the parametric and non-parametric families for ground- and space-based telescopes, with their advantages and limitations. Validation methods for PSF models are then addressed, with several metrics related to weak-lensing studies discussed in detail. Finally, we explore current challenges and future directions in PSF modelling for astronomical telescopes.

KEYWORDS

point spread function, inverse problems, weak gravitational lensing, image processing, super-resolution

1 Introduction

Any astronomical image is observed through an optical system that introduces deformations and distortions. Even the most powerful imaging system introduces distortions to the observed object. How to characterise these distortions is a subject of study known as PSF modelling. Specific science applications, like weak gravitational lensing (WL) in cosmology (for a review, see Kilbinger, 2015; Mandelbaum, 2018), require very accurate and precise measurements of galaxy shapes. A crucial step of any weak-lensing mission is to estimate the PSF at any position of the observed images. If the PSF is not considered when

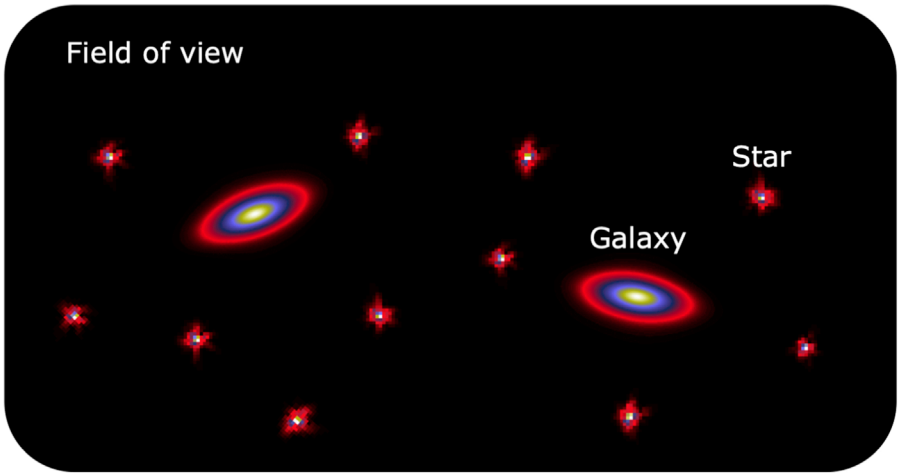


FIGURE 1
Illustration of a field of view showing the PSF modelling problem. Firstly, the PSF model should be estimated from the stars. The model should then be used to estimate the PSF at the target positions, e.g., galaxy positions.

measuring galaxy shapes, the measurement will be biased, resulting in unacceptably biased WL studies. Furthermore, the PSF can be the predominant source of systematic errors and biases in WL studies. This fact makes PSF modelling a vital task. Forthcoming astronomical telescopes, such as the *Euclid* space telescope (Laureijs et al., 2011), the *Roman* Space Telescope (Spergel et al., 2015; Akeson et al., 2019), and the Vera C. Rubin Observatory (LSST Science Collaboration et al., 2009; Ivezić et al., 2019), raise many challenges for PSF models as the instruments get more complex and the imposed scientific requirements get tighter. These factors have triggered and continue to trigger developments in the PSF modelling literature.

PSF modelling is an interdisciplinary problem that requires knowledge of optics, inverse problems, and the target science application—in our case, weak gravitational lensing studies. The objective is to estimate the PSF at target positions, e.g., galaxy positions, from degraded star observations and complementary sources of information. Figure 1 shows an illustration of the problem. The PSF modelling problem is challenging as the model should account for the different variations of the PSF in the field of view, i.e., spatial, spectral, and temporal. This review is related to these three scientific fields, discusses in detail the PSF, and aims to help understand the different PSF modelling choices. We start by introducing optical concepts required to analyse optical imaging systems that are required to understand the more physically based PSF models in Section 2. Then, motivated by the optical introduction, we describe the adopted general observational model in Section 3. Section 4 introduces the different contributors to the PSF at the optical and detector levels. Section 5 gives an overview of state-of-the-art PSF modelling techniques and leads to Section 6, which includes comments on the desirable properties of a PSF model. We end the review by describing different techniques for validating PSF models in Section 7 and concluding in Section 8. In addition, we include Table 1, which summarizes the notation and the different coordinates used throughout this article.

TABLE 1 Coordinates and notation used throughout this article.

Variable	Description
Coordinates	
(x, y)	Pupil plane or output aperture plane coordinates
(u, v)	Image or focal plane coordinates
(ξ, η)	Object plane coordinates
$(\bar{u}, \bar{v})$	Pixel coordinates, the discrete counterpart of the image plane
$\mathbf{p}_i$	3D spatial coordinate
λ	Wavelength
t	Time
Notation	
$\mathcal{I}, \mathcal{H}, \ldots$	Calligraphic uppercase variables are continuous functions
$I, H, \ldots$	Uppercase variables are matrices
$c_m, b_1^k, \ldots$	Lowercase variables are scalar
$I_{\text{img}}(\bar{u}, \bar{v}; t \vert u_i, v_i) \in \mathbb{R}$	Pixel value at position $(\bar{u}, \bar{v})$ for the image I_{img} with its centroid at position (u_i, v_i) observed at time t
$I_{\text{img},(\vert u_i, v_i)} \in \mathbb{R}^{p \times p}$	Observed image with its centroid at position (u_i, v_i)

2 Gentle introduction to optics

A rigorous treatment of the optics involved in the formation of the PSF on complex optical systems could be the sole topic of a review article. In this section, we introduce simplified optical concepts to motivate a more physical understanding of the PSF, how to model it, and certain implicit assumptions that are usually adopted. This review follows the optic formalism of Goodman (2005). For a profound and rigorous description of optical theory, we refer the readers to the seminal book of Born and Wolf (1999) or more concise works (Gaskill, 1978; Gross, 2005; Hecht, 2017);

for more information on practical wave propagation, we refer the readers to Schmidt (2010); and if the readers are familiar with the Fourier optics literature, we recommend continuing to Section 3.

This introduction is based on the scalar formulation of diffraction. It starts by presenting diffraction equations from a general perspective with the Huygens–Fresnel principle to the more simplified formulations of Fresnel and Fraunhofer. The introduction continues with the diffraction analysis of the effects of a thin single-lens optical system. The results motivate the analysis of more general optical systems that are treated with the black box concept from Goodman (2005). The section proceeds by introducing the modelling of aberrations in the optical system, and then extending the monochromatic to the polychromatic analysis briefly studying the coherent and incoherent cases. The optical introduction ends by mentioning several assumptions usually adopted in the PSF modelling literature.

2.1 Scalar diffraction theory

2.1.1 Huygens–Fresnel principle

When studying the PSF, we are examining how an optical system with a specific instrument contributes to and modifies our observations. To understand how the optical system interacts with the propagation of light, we have to dig into the nature of light, an electromagnetic (EM) wave. To make a fundamental analysis, one would have to use Maxwell's equations, solve them with the optical system under study, and obtain the electric and magnetic fields. Solving a set of coupled partial differential equations is an arduous task. Several approximations can be made, if some conditions are met, to alleviate the mathematical burden of solving Maxwell's equations without introducing much error into the analysis.

The diffraction theory provides a fundamental framework for analysing light propagation through an optical system. This is especially the case when working with EM waves in the optical range when the optical image is close to the focus region. The *Huygens–Fresnel principle* (Huygens, 1690; Fresnel, 1819; Crew et al., 1900) states that every point of a wavefront may be considered a secondary disturbance giving rise to spherical wavelets. At any later instant, the wavefront may be regarded as the envelope of all the disturbances. Fresnel's contribution to the principle is that the secondary wavelets mutually interfere. This principle provides a powerful method of analysis of luminous wave propagation. In Figure 2A, the propagation of an incident plane wavefront through an obstacle, a single slit, is shown. The secondary wavelets constitute the plane wavefront before the obstacle. Then, the wavefront shape is modified due to the obstacle, following the Huygens–Fresnel principle.

The secondary waves mutually interfere constructively or destructively, according to their phases. The analysis of the light propagation in a homogeneous medium is simple as the spherical wavelets interfere without obstacles, and the total wavefront propagates spherically in the medium. However, suppose the wave encounters an obstacle. In that case, the secondary waves in the vicinity of the boundaries of the obstacle will interfere in ways that are not obvious from the incident wavefront.

Let us study the Huygens–Fresnel principle and consider a diffractive aperture in a plane (x,y) illuminated in the positive z direction. We analyse the diffracted wave in a parallel plane (u,v) at a normal distance z from the first plane. The z-axis is orthogonal to both planes and intersects them at their origins. Figure 3 illustrates the coordinate system described previously. The diffracted wave, which can be intuitively understood as the superposition of spherical waves, is written as

$$\mathcal{U}(\mathbf{p}_1) = \frac{z}{\mathrm{j}\lambda}\iint_{\Sigma}\mathcal{U}(x,y;0)\,\frac{\exp\left[\mathrm{j}\,k\,r_{01}\right]}{r_{01}^2}\,\mathrm{d}x\mathrm{d}y, \tag{1}$$

where j denotes the imaginary unit, λ is the wavelength, $k = 2\pi/\lambda$, $\mathbf{p}_0 = (x_0, y_0;\ 0)$, $\mathbf{p}_1 = (u_1, v_1;\ z)$, $r_{01} = \|\mathbf{p}_1 - \mathbf{p}_0\|_2$, Σ is the aperture in the (x,y) plane, and $\mathcal{U}$ is the electric field. The incident wave is $\mathcal{U}(\mathbf{p}_0)$, and the diffracted wave is $\mathcal{U}(\mathbf{p}_1)$.

There are two main approximations in the derivation of Eq. 1. The first approximation is that we are considering a *scalar theory of diffraction*, a scalar electric and magnetic field, and not the fields in their complete vectorial form. The scalar theory provides a full description of the EM fields in a dielectric medium that is linear, isotropic, homogeneous, and non-dispersive. However, even if the medium verifies these properties, when some boundary conditions are imposed on a wave, e.g., an aperture, some coupling is introduced between the EM field components and the scalar theory becomes no longer exact. Nevertheless, the EM fields are modified only at the edges of the aperture, and the effects extend over only a few wavelengths into the aperture. Therefore, if the aperture is large when compared to the wavelength, the error introduced by the scalar theory is negligible. Refractive optical elements can also induce polarisation of the EM field. The level of accuracy desired will determine if the bias introduced can be neglected or has to be taken into account.

Although the current formulation is powerful in representing the diffraction phenomena, it is still challenging to work with the integral from Eq. 1. As a consequence, we will explore further approximations that will give origin to the *Fresnel diffraction* and *Fraunhofer diffraction*.

2.1.2 Fresnel diffraction

The Fresnel approximation is based on the binomial expansion of the square root in the expression $\sqrt{1+b}$ for some b[1]. The distance r_{01} can be expressed as

$$r_{01} = z\sqrt{1 + \left(\frac{u_1 - x_0}{z}\right)^2 + \left(\frac{v_1 - y_0}{z}\right)^2}, \tag{2}$$

which can be approximated using the first two terms of the binomial expansion, as

$$r_{01} \approx z\left(1 + \frac{1}{2}\left(\frac{u_1 - x_0}{z}\right)^2 + \frac{1}{2}\left(\frac{v_1 - y_0}{z}\right)^2\right). \tag{3}$$

The r_{01} appearing in the exponential of Eq. 1 has much more influence on the result than the r_{01}^2 in the divisor. Therefore, we use Eq. 3 to approximate r_{01} in the exponential, and for the divisor, we approximate $r_{01}^2 \approx z^2$. Then, we can express the diffracted field as

$$\mathcal{U}(u_1, v_1; z) = \frac{\exp\left[\mathrm{j}kz\right]}{\mathrm{j}\lambda z}\iint_{\Sigma}\mathcal{U}(x,y;0) \times \exp\left[\mathrm{j}\,\frac{k}{2z}\left[(u_1 - x)^2 + (v_1 - y)^2\right]\right]\mathrm{d}x\mathrm{d}y, \tag{4}$$

1 The binomial expansion is given by $\sqrt{1+b} = 1 + \frac{1}{2}b - \frac{1}{8}b^2 + \cdots$.

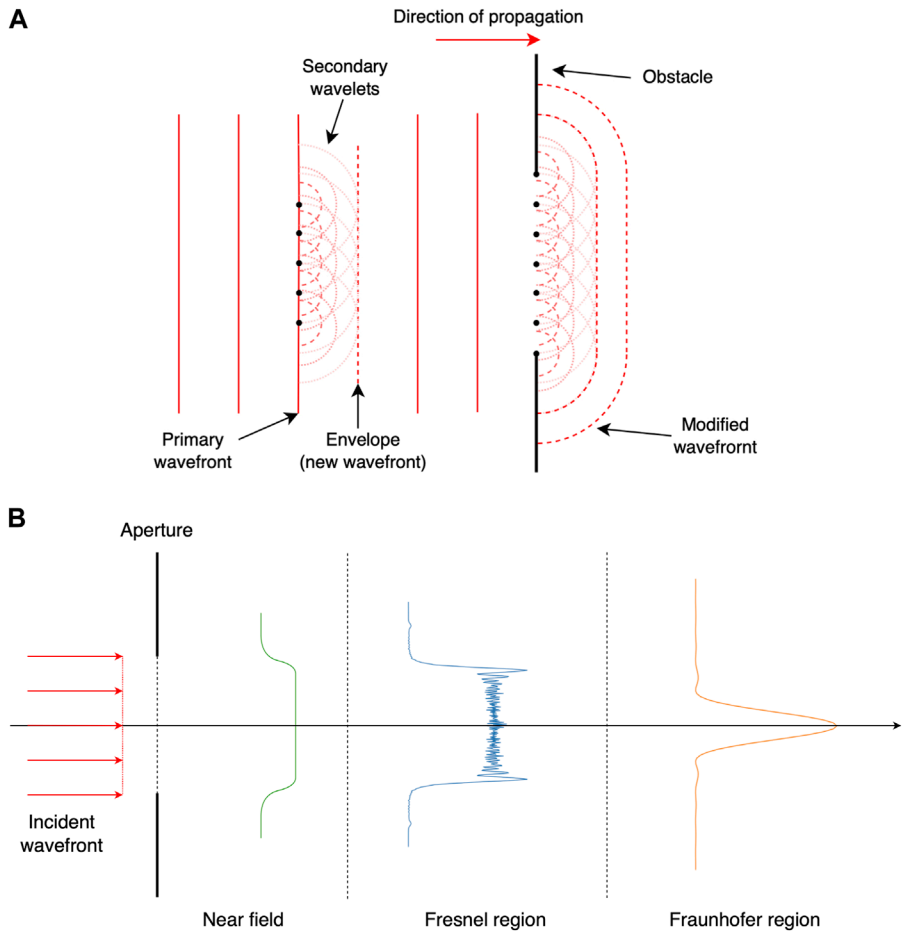


FIGURE 2
(A) Illustration of the Huygens–Fresnel principle and the modification of a wavefront due to an obstacle. Reproduced from Liaudat (2022). **(B)** Illustration of the different diffraction regions behind an aperture.

and if we expand the terms in the exponential, we get

$$\mathcal{U}(u_1,v_1;z) = \frac{\exp[\mathrm{j}kz]}{\mathrm{j}\lambda z}\exp\left[\mathrm{j}\,\frac{k}{2z}\left(u_1^2+v_1^2\right)\right] \times \iint_{\Sigma}\left\{\mathcal{U}(x,y;0)\exp\left[\mathrm{j}\,\frac{k}{2z}\left(x^2+y^2\right)\right]\right\} \times \exp\left[-\mathrm{j}\,\frac{2\pi}{\lambda z}\left(u_1x+v_1y\right)\right]\mathrm{d}x\mathrm{d}y. \quad (5)$$

The Fourier transform (FT) expression can be recognised with some multiplicative factors in Eq. 5. The diffracted wave is the FT of the product of the incident wave and a quadratic phase exponential. In this case, we have approximated the spherical secondary waves of the Huygens–Fresnel principle by parabolic wavefronts. The approximation in the Fresnel diffraction formula is equivalent to the paraxial approximation Goodman (2005, §4.2.3). This last approximation consists of a *small-angle approximation* as it restricts the rays to be close to the optical axis. This restriction also allows us to approximate Eq. 2 with Eq. 3. The region where the approximation is valid is known as the *region of Fresnel diffraction*. In this region, the major contributions to the integral come from points (x,y) for which $x \approx u$ and $y \approx v$, i.e., the higher-order terms in the expansion that we are not considering are unimportant. The *region of Fresnel diffraction* can be seen as the coordinates (u,v,z) that verify

$$z^3 \gg \frac{\pi}{4\lambda}\left((u-x)^2+(v-y)^2\right)^2, \quad \forall (x,y) \in \Sigma. \quad (6)$$

A more practical and widely used condition is the *Fresnel number* (Hecht, 2017, §10.3.3) which can be written as follows:

$$N_F = \frac{r^2}{\lambda z}, \quad (7)$$

where r is the radius of a circular aperture and z is the distance from the aperture. If N_F is close to unity, the Fresnel diffraction is a good approximation. However, if $N_F \ll 1$, then Fraunhofer's approximation, which we will introduce in the following section, is valid. For more information on the validity of the Fresnel approximation, we refer the readers to Southwell (1981) and Rees (1987).

2.1.3 Fraunhofer diffraction

We continue to present a further approximation that, if valid, can significantly simplify the calculations. The Fraunhofer approximation assumes that the exponential term with a quadratic dependence of (x,y) is approximately unity over the aperture. The region where the approximation is valid is the *far field* or *Fraunhofer*

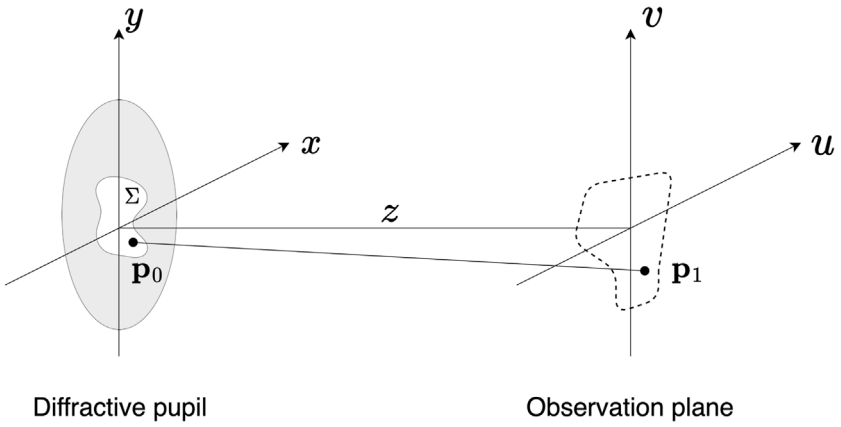


FIGURE 3
Illustration of the coordinate system for the diffraction equations. Figure adapted from Liaudat (2022).

region. Figure 2B illustrates the different diffraction approximations as a function of the aperture's distance. The required condition to be in this region reads

$$z \gg \frac{k(x^2+y^2)}{2}, \quad \forall (x,y) \in \Sigma. \tag{8}$$

The Fraunhofer diffraction formula is given by

$$\mathcal{U}(u_1,v_1;z) = \frac{\exp[\mathrm{j}kz]}{\mathrm{j}\lambda z}\exp\left[\mathrm{j}\frac{k}{2z}(u_1^2+v_1^2)\right]\iint_\Sigma \{\mathcal{U}(x,y;0)\} \times \exp\left[-\mathrm{j}\frac{2\pi}{\lambda z}(u_1 x + y v_1)\right]\mathrm{d}x\mathrm{d}y, \tag{9}$$

where we can reformulate the previous equation using the FT as follows:

$$\mathcal{U}(u_1,v_1;z) = \frac{\exp[\mathrm{j}kz]}{\mathrm{j}\lambda z}\exp\left[\mathrm{j}\frac{k}{2z}(u_1^2+v_1^2)\right] \times \mathrm{FT}\{\iota_\Sigma(x,y;0)\mathcal{U}(x,y;0)\}\left(\frac{u_1}{\lambda z},\frac{v_1}{\lambda z}\right), \tag{10}$$

where ι_Σ is an indicator function over the aperture taking values in $\{0,1\}$. It is also possible to consider image vignetting and multiply the indicator with a weight function so that the resulting function takes values in $[0,1]$. Cameras are sensitive to the light's intensity reaching their detectors. The instantaneous intensity of an EM wave is equal to its squared absolute value. Therefore, we can write the intensity of the diffracted wave as

$$\mathcal{I}(u_1,v_1;z) = |\mathcal{U}(u_1,v_1;z)|^2 = \frac{1}{\lambda^2 z^2}\left|\mathrm{FT}\{\iota_\Sigma(x,y;0)\mathcal{U}(x,y;0)\}\left(\frac{u_1}{\lambda z},\frac{v_1}{\lambda z}\right)\right|^2, \tag{11}$$

which is significantly simpler than the original Rayleigh–Sommerfeld expression from Eq. 1.

2.2 Modelling diffraction in a simple optical system

The study of the diffraction phenomena is necessary but not sufficient to describe the effects of an optical system. Optical imaging systems are generally based on lenses or mirrors, which have the ability to form images. To simplify the analysis, we studied the effect of a single positive (converging) thin lens illuminated with monochromatic illumination and computed the impulse response of such a system. The coordinate system used for this analysis is shown in Figure 4.

Let us write the output wave as a function of the input wave using the superposition integral as follows:

$$\mathcal{U}_{\mathrm{img}}(u,v) = \iint_{-\infty}^{+\infty}\mathcal{H}(u,v;\xi,\eta)\,\mathcal{U}_{\mathrm{obj}}(\xi,\eta)\,\mathrm{d}\xi\mathrm{d}\eta, \tag{12}$$

where $\mathcal{H}$ is the field's value at image coordinates (u,v) due to a unitary point-source object at position (ξ,η). We describe a monochromatic input wave reaching the entrance pupil of the lens coming from a point source located at (ξ,η) at the object plane, which is located at a distance z_o from the lens. Following the *paraxial approximation*, we can write the waves at the entrance pupil as follows:

$$\mathcal{U}_l(x,y) = \frac{1}{\mathrm{j}\lambda z_o}\exp\left[\mathrm{j}\frac{k}{2z_o}((x-\xi)^2+(y-\eta)^2)\right], \tag{13}$$

and the wave at the output pupil is as follows:

$$\mathcal{U}_{l'}(x,y) = \mathcal{U}_l(x,y)\,\mathcal{P}(x,y)\exp\left[-\mathrm{j}\frac{k}{2f}(x^2+y^2)\right], \tag{14}$$

where f is the focal length of the lens and $\mathcal{P}$ is the pupil function of the lens which accounts for the finite dimension of the lens, i.e., the obscured and unobscured areas. We have implicitly assumed that the pupil function is constant for any (u,v) position considered. This assumption does not hold for wide-field imagers where there are obscurations involved in the pupil function. We continue by using the Fresnel diffraction formula from Section 2.1.2 to compute the diffraction effect from the lens' exit pupil to the image plane. Replacing $\mathcal{U}$ in Eq. 5 with the output lens wave $\mathcal{U}_{l'}$ to compute the

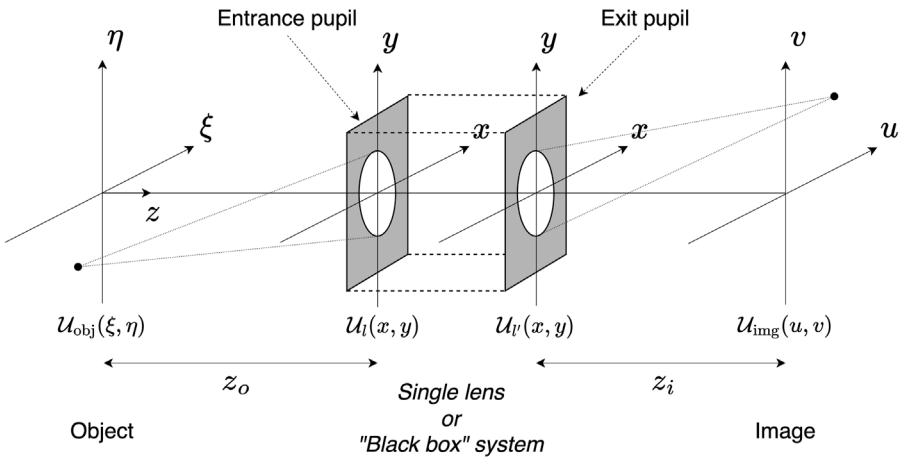


FIGURE 4
Illustration of the coordinate system of the imaging systems we are studying. The central imaging system can be a single positive lens or the generalised *black box* concept of an imaging system. The image plane coordinates are (u,v), the input and output aperture plane coordinates are (x,y), and the object plane coordinates are (ξ,η). This figure has been adapted from Goodman (2005).

impulse response, we obtain

$$\mathcal{H}(u,v;\xi,\eta) = \overbrace{\frac{1}{\lambda^2 z_o z_i}\exp\left[\mathrm{j}\frac{k}{2z_i}(u^2+v^2)\right]}^{\text{(II)}}\overbrace{\exp\left[\mathrm{j}\frac{k}{2z_o}(\xi^2+\eta^2)\right]}^{\text{(III)}} \times \iint_{-\infty}^{\infty}\mathcal{P}(x,y)\underbrace{\exp\left[\mathrm{j}\frac{k}{2}\left(\frac{1}{z_o}+\frac{1}{z_i}-\frac{1}{f}\right)(x^2+y^2)\right]}_{\text{(I)}} \times \exp\left[-\mathrm{j}k\left(\left(\frac{\xi}{z_o}+\frac{u}{z_i}\right)x+\left(\frac{\eta}{z_o}+\frac{v}{z_i}\right)y\right)\right]\mathrm{d}x\mathrm{d}y. \tag{15}$$

The previous formula of the impulse response of a positive lens is hard to exploit in a practical sense due to the quadratic phase terms. However, several approximations can be exploited to remove them.

- We start studying the term (I) inside the integrand. We consider the image plane to coincide with the focal plane, i.e., $z_i = f$, and the imaged object to be very far away from the entrance pupil. Consequently, the term (I) is approximately one. The part of the exponent which is close to zero is

$$\frac{1}{z_o}+\frac{1}{z_i}-\frac{1}{f}\approx 0, \tag{16}$$

which, in the case of equality, is known as the *lens law* of geometrical optics.

- The term (II) only depends on the image coordinates (u,v). The term can be ignored as we are interested in the intensity distribution of the image, and it is not being integrated in Eq. 12.
- The term (III) depends on the object coordinates, is integrated into the convolution operation in Eq. 12, and therefore might significantly change the imaged object. We can neglect the influence of this term if its phase changes by a small amount, i.e., a small fraction of a radian, within the region of the object that mostly contributes to the image position (u,v). A deeper discussion about the validity of the term (III) approximation is found in Goodman (2005, §5.3.2) and references therein.

We can now apply the previous approximations to the calculation of the impulse response of an optical system with a positive lens. Under Fresnel diffraction, we simplify Eq. 15 to obtain

$$\mathcal{H}\left(u,v;\tilde{\xi},\tilde{\eta}\right) \approx \frac{1}{\lambda^2 z_o f}\iint_{-\infty}^{+\infty}\mathcal{P}(x,y) \times \exp\left[-\mathrm{j}\frac{2\pi}{\lambda f}\left(\left(u-\tilde{\xi}\right)x+\left(v-\tilde{\eta}\right)y\right)\right]\mathrm{d}x\mathrm{d}y, \tag{17}$$

where $m = -f/z_o$ is the magnification of the system, which can be positive or negative depending on whether the image is inverted or not, and the normalised (or reduced) object-plane coordinates are $\tilde{\xi} = m\,\xi$ and $\tilde{\eta} = m\,\eta$. The diffraction pattern is centred on the image coordinates, $u = m\,\xi$ and $v = m\,\eta$, which are the transformed coordinates of the impulse response's position (ξ,η).

The impulse response obtained in Eq. 17 is Fraunhofer's diffraction pattern centred in $(u=\tilde{\xi}, v=\tilde{\eta})$ and up to a scaling factor of $1/\lambda z_o$. This result is the consequence of the choice of z_i, such that it verifies the lens law, allowing us to drop out quadratic phase terms in the integral. We have obtained a simple formulation for the impulse response, but the optical system that we studied is not used in practice to carry out galaxy imaging surveys. We have to extend the analysis to more general optical systems.

2.3 Analysis of a general optical imaging system

Let us now analyse a general optical imaging system composed of one or many lenses or mirrors of possibly different characteristics. We treat the optical system as a *black box* characterised by the transformations applied to an incident *object* scalar wave, $\mathcal{U}_{\mathrm{obj}}$, into an output *image* wave, $\mathcal{U}_{\mathrm{img}}$. Figure 4 illustrates the new interpretation of the optical system, where we have replaced the

previous single-lens system with a *black box*. In this general model, we assume that the effect of the optical system between the entrance and exit pupils is well described by geometrical optics, which is an affine transformation. We also assume that all the diffraction effects can be associated with one of the two pupils, input or output (for more discussion on both assumptions, see Goodman, 2005, §6.1). We choose the latter and consider the diffraction of the output wave between the output pupil and image plane. For the moment, our analysis continues to assume an ideal monochromatic illumination.

The *ideal image*, $\mathcal{U}_g$, is defined as the input image when applying the effect of geometrical optics inside the *black box* and is given as

$$\mathcal{U}_g\left(\tilde{\xi},\tilde{\eta}\right) = \frac{1}{|m|}\mathcal{U}_{obj}\left(\frac{\tilde{\xi}}{m},\frac{\tilde{\eta}}{m}\right), \quad \text{and} \quad \tilde{\xi} = m\xi, \quad \tilde{\eta} = m\eta, \tag{18}$$

where m is the magnification factor of the optical system, and we express the images in *reduced coordinates*.

Our analysis is based on the impulse response developed in the previous section. The approximations applied and the use of reduced object coordinates have made the system spatially invariant. This fact translates to having $\mathcal{H}(u,v;\tilde{\xi},\tilde{\eta}) = \mathcal{H}(u-\tilde{\xi},v-\tilde{\eta})$, as the approximated impulse response from Eq. 17 depends only on the difference of the image coordinates and the reduced object coordinates. The impulse response is given as

$$\mathcal{H}\left(u-\tilde{\xi},v-\tilde{\eta}\right) = \frac{a}{\lambda f}\iint_{-\infty}^{+\infty}\mathcal{P}(x,y) \times \exp\left[-\mathrm{j}\frac{2\pi}{\lambda f}\left(\left(u-\tilde{\xi}\right)x+\left(v-\tilde{\eta}\right)y\right)\right]\mathrm{d}x\mathrm{d}y, \tag{19}$$

where a is a constant amplitude that does not depend on the optical system under study. The superposition integral in Eq. 12 relates the waves at the object and image positions with the impulse response in a *spatially variant* system. However, if the system is *spatially invariant*, the equation can be reformulated as the convolution equation, which is given as

$$\mathcal{U}_{img}(u,v) = \iint_{-\infty}^{+\infty}\mathcal{H}\left(u-\tilde{\xi},v-\tilde{\eta}\right)\mathcal{U}_g\left(\tilde{\xi},\tilde{\eta}\right)\,\mathrm{d}\tilde{\xi}\,\mathrm{d}\tilde{\eta}. \tag{20}$$

The previous equation can be rewritten with the usual convolution notation as

$$\mathcal{U}_{img}(u,v) = \left(\mathcal{U}_g \star \mathcal{H}\right)(u,v). \tag{21}$$

In this general case of a system without aberrations and under the aforementioned approximations, we see that the output image is formed by a geometrical-optics transformation followed by a convolution with an impulse response from the Fresnel diffraction of the exit aperture.

2.3.1 Introducing optical aberrations

In the previous development, we considered an ideal optical system without any aberrations, known as *diffraction-limited*. An aberrated optical system produces the imperfect convergence of rays, which can be expressed equivalently in wavefront space by deviations from the ideal reference sphere. The aberrations produce leads and lags in the wavefront with respect to the ideal sphere (see Figure 5). A complementary interpretation, from Goodman (2005), is that we start with the previous diffraction-limited system producing converging spherical wavefronts. Then, we add a phase-shifting plate representing the system's aberrations. The plate is

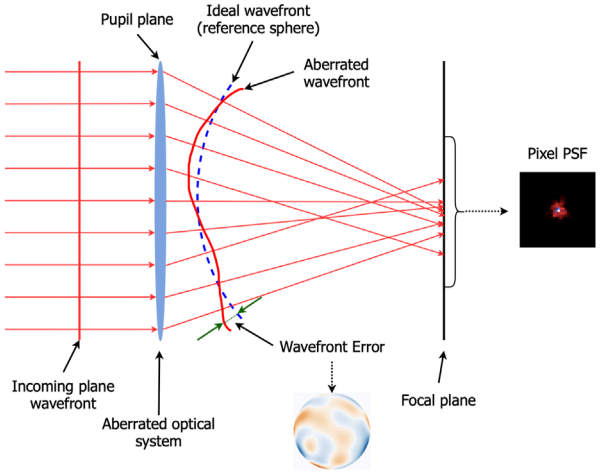


FIGURE 5
Illustration of the wavefront error in a one-dimensional projection of an ideal setting where the optical system is represented as a single lens. Figure reproduced from Liaudat et al. (2023).

located at the aperture after the exit pupil and affects the output wave's phase. To characterise the aberrations, we use the *generalised pupil function* that generalises the pupil function $\mathcal{P}$ from Eq. 19 and gives

$$\mathcal{P}_{gen}(x,y;u,v) = \mathcal{P}(x,y;u,v)\exp\left[\mathrm{j}\frac{2\pi}{\lambda}\mathcal{W}(x,y;u,v)\right], \tag{22}$$

where λ is the central wavelength of the incident wave, $\mathcal{P}$ is the pupil function that includes the telescope's obscurations, and $\mathcal{W}$ represents the optical path differences (OPDs) between a perfectly spherical and the aberrated wavefront. We also refer to the OPD as the wavefront error (WFE). Figure 5 illustrates the concept of WFE. It is common to represent the WFE using a Zernike polynomial decomposition (Noll, 1976) as these are orthogonal in the unit disk, and we generally use circular apertures in telescopes and optical systems. Figure 6B shows the first Zernike polynomials.

The aberrations, $\mathcal{W}$, and pupil function, $\mathcal{P}$, depend on the object's position in the focal plane as is seen in the (u,v) coordinate dependence in Eq. 22. Large telescopes with wide focal planes have spatially varying aberrations. The path travelled by the light rays changes considerably between distant points in the focal plane, also changing the aberrations, $\mathcal{W}$. The obscurations and aperture, represented by the pupil function $\mathcal{P}$, also change with the focal plane position. For example, Figure 6A illustrates the obscurations from the *Euclid* telescope. One can notice a circular aperture with several obscurations in it, a secondary mirror, and three arms supporting the mirror. What we observe in Figure 6A is a 2D projection of the 3D structure. This projection changes as a function of the focal plane position that we are analysing, making the function $\mathcal{P}$ dependent on the (u,v) coordinates.

In the impulse response of the optical system without aberrations from Eq. 19, we had a spatially invariant system. This invariance allowed us to use the convolution rather than the superposition integral, which is a computationally practical formulation. If we now consider aberrations, we must inject the generalised pupil function appearing in Eq. 22 into the impulse

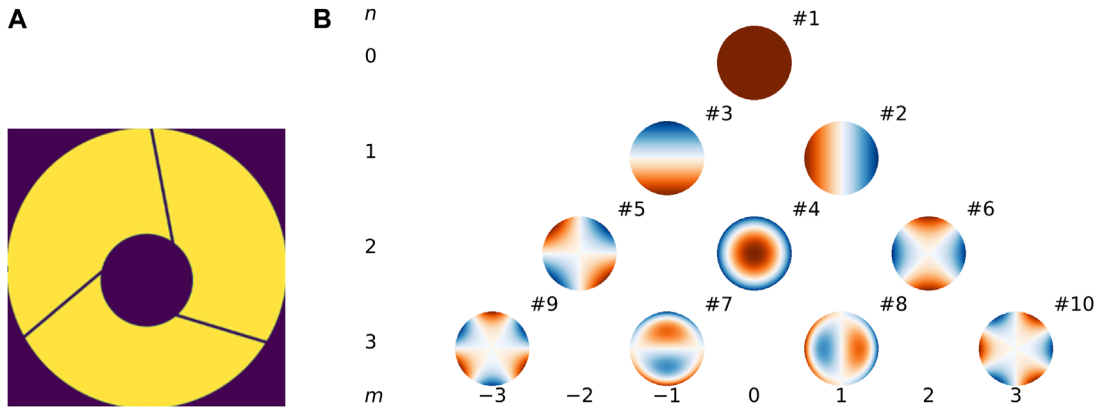


FIGURE 6
(A) Illustration of *Euclid*'s pupil function, which can be seen in Venancio et al. (2020), in the (x,y) plane for a given position in the (ξ,η). **(B)** Example of the first Zernike polynomial maps.

response formula from Eq. 19. The addition of the (u,v) dependency in $\mathcal{P}_{\text{gen}}$ makes the impulse response $\mathcal{H}$ spatially variant again.

The study of $\mathcal{H}$, the impulse response and the main topic of this review, is strongly spatially variant in systems with a large focal plane. Nevertheless, we can consider $\mathcal{H}$ spatially invariant in its *isoplanatic region*. This region consists of close-by points in the focal plane, where the light has travelled similar paths giving small deviations of $\mathcal{H}$. We are assuming a certain regularity in $\mathcal{H}$ due to the optical system under study that allows the deviations to be small. In other words, we consider $\mathcal{H}$ to be locally spatially invariant or spatially invariant in patches. Figure 7 illustrates the idea of an *isoplanatic region*. This local invariance assumption limits the size of the imaged objects under study, as they should have a certain size range with respect to the support of $\mathcal{H}$ so that all the objects being imaged lie within the aforementioned region. We consider the generalised pupil function evaluated at the centroid of the imaged object, (u_i,v_i), and note the locally spatially invariant generalised pupil function as follows:

$$\mathcal{P}^*_{\text{gen}}(x,y|u_i,v_i) = \mathcal{P}^*(x,y|u_i,v_i)\exp\left[\mathrm{j}\frac{2\pi}{\lambda}\mathcal{W}^*(x,y|u_i,v_i)\right]. \quad (23)$$

Injecting Eq. 23, instead of Eq. 22, to the impulse response in Eq. 19 gives

$$\begin{aligned}\mathcal{H}\left(u-\tilde{\xi},v-\tilde{\eta}|u_i,v_i\right) = \frac{a}{\lambda f}\iint_{-\infty}^{+\infty}&\mathcal{P}^*(x,y|u_i,v_i)\\ &\times\exp\left[\mathrm{j}\frac{2\pi}{\lambda}\mathcal{W}^*(x,y|u_i,v_i)\right]\exp\\ &\times\left[-\mathrm{j}\frac{2\pi}{\lambda f}\left(\left(u-\tilde{\xi}\right)x+\left(v-\tilde{\eta}\right)y\right)\right]\mathrm{d}x\mathrm{d}y, \quad (24)\end{aligned}$$

where we have made the system spatially invariant again, allowing us to exploit the convolution formula in Eq. 20.

We have considered aberrations that only depend on the object's position in the focal plane, also known as achromatic aberrations. However, depending on the optical system under study, there might be wavelength-dependent aberrations. For example, some refractive components, or some components implementing complex thin film coatings, may introduce spurious spectral dependences to the optical system's response. If this is the case, we can add a wavelength dependence to the WFE function $\mathcal{W}$ to account for these effects.

2.3.2 Polychromatic illumination: coherent and incoherent cases

We studied until now a system with ideal monochromatic light. It is time to shift to polychromatic light as telescopes have filters with finite bandwidths and hence allow multiple frequencies of light. For a more rigorous analysis of polychromatic illumination, we refer the readers to the theory of partial coherence in Beran and Parrent (1964), Goodman (1985), and Born and Wolf (1999, §10). Even if we were to study the system's behaviour to light with a particular wavelength, this has practically never been the case, as real illumination is never perfectly chromatic, even for lasers. Therefore, we consider a narrowband polychromatic illumination centred at a given wavelength λ. The *narrowband assumption* states that the bandwidth occupied is small with respect to the central wavelength. For polychromatic light, we follow Goodman (2005) and consider a time-varying phasor of the field, $\mathcal{U}_{\text{img}}(u,v;t)$, where its intensity is given by the time integration of its instantaneous intensity:

$$\mathcal{I}_{\text{img}}(u,v) = \left\langle\left|\mathcal{U}_{\text{img}}(u,v;t)\right|^2\right\rangle_t = \frac{1}{T}\int_{-T/2}^{T/2}\left|\mathcal{U}_{\text{img}}(u,v;t)\right|^2\mathrm{d}t, \quad (25)$$

where T is the detector integration time that is considered to be much greater than the optical wave period. We can generalise the field expression from Eq. 20 by considering that light is polychromatic and that the impulse response $\mathcal{H}$ is wavelength independent due to the narrowband assumption. The field then is given as

$$\mathcal{U}_{\text{img}}(u,v;t) = \iint_{-\infty}^{+\infty}\mathcal{H}\left(u-\tilde{\xi},v-\tilde{\eta}\right)\mathcal{U}_{\text{g}}\left(\tilde{\xi},\tilde{\eta};t-\tau\right)\,\mathrm{d}\tilde{\xi}\mathrm{d}\tilde{\eta}, \quad (26)$$

where τ represents the delay of the wave propagation from $(\tilde{\xi},\tilde{\eta})$ to (u,v). Continuing with the polychromatic analysis, we rewrite the intensity from Eq. 25 as

$$\begin{aligned}\mathcal{I}_{\text{img}}(u,v) = &\iint_{-\infty}^{+\infty}\mathrm{d}\tilde{\xi}_1\mathrm{d}\tilde{\eta}_1\iint_{-\infty}^{+\infty}\mathrm{d}\tilde{\xi}_2\mathrm{d}\tilde{\eta}_2\,\mathcal{H}\left(u-\tilde{\xi}_1,v-\tilde{\eta}_1\right)\\ &\times\mathcal{H}^*\left(u-\tilde{\xi}_2,v-\tilde{\eta}_2\right)\mathcal{J}_{\text{g}}\left(\tilde{\xi}_1,\tilde{\eta}_1;\tilde{\xi}_2,\tilde{\eta}_2\right), \quad (27)\end{aligned}$$

where $\mathcal{H}^*$ is the conjugate of $\mathcal{H}$, $\mathcal{J}_{\text{g}}$ is known as the *mutual intensity* which describes the spatial coherence of $\mathcal{U}_{\text{g}}$ at two points and is given

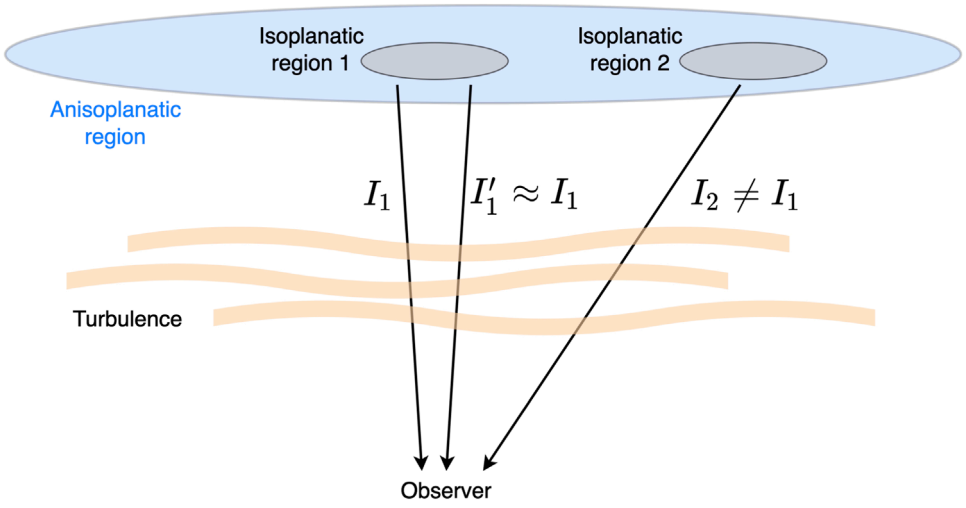


FIGURE 7
Illustration of isoplanatic regions. Two rays from the same isoplanatic region travel through almost the same turbulence and suffer almost the same distortions. Figure reproduced from Liaudat (2022).

as

$$\mathcal{J}_{\mathrm{g}}\left(\tilde{\xi}_1,\tilde{\eta}_1;\tilde{\xi}_2,\tilde{\eta}_2\right)=\left\langle \mathcal{U}_{\mathrm{g}}\left(\tilde{\xi}_1,\tilde{\eta}_1;t\right)\mathcal{U}_{\mathrm{g}}^{*}\left(\tilde{\xi}_2,\tilde{\eta}_2;t\right)\right\rangle. \tag{28}$$

We can distinguish two types of illuminations, *coherent* and *incoherent*. *Coherent* illumination refers to waves whose phases vary in a perfectly correlated way. This illumination is approximately the case of a laser. In *incoherent* illumination, the wave's phases vary in an uncorrelated fashion. Most natural light sources can be considered incoherent sources. The *mutual intensity* is helpful to represent both types of illumination. In the case of coherent light, we obtain

$$\mathcal{J}_{\mathrm{g}}^{\mathrm{co}}\left(\tilde{\xi}_1,\tilde{\eta}_1;\tilde{\xi}_2,\tilde{\eta}_2\right)=\mathcal{U}_{\mathrm{g}}\left(\tilde{\xi}_1,\tilde{\eta}_1\right)\mathcal{U}_{\mathrm{g}}^{*}\left(\tilde{\xi}_2,\tilde{\eta}_2\right), \tag{29}$$

where $\mathcal{U}_{\mathrm{g}}\left(\tilde{\xi}_1,\tilde{\eta}_1\right)$ and $\mathcal{U}_{\mathrm{g}}^{*}\left(\tilde{\xi}_2,\tilde{\eta}_2\right)$ are time-independent phasor amplitudes relative to their time-varying counterpart. As both time-varying phasors are synchronized, we have taken a reference phasor and normalised it against the amplitude with respect to a reference point that can be the origin (0, 0). For example,

$$\mathcal{U}_{\mathrm{g}}\left(\tilde{\xi}_1,\tilde{\eta}_1;t\right)=\mathcal{U}_{\mathrm{g}}\left(\tilde{\xi}_1,\tilde{\eta}_1\right)\frac{\mathcal{U}_{\mathrm{g}}(0,0;t)}{\left\langle\left|\mathcal{U}_{\mathrm{g}}(0,0;t)\right|^2\right\rangle^{\frac{1}{2}}}. \tag{30}$$

By substituting Eq. 29 in Eq. 27, we obtain

$$\begin{aligned}\mathcal{I}_{\mathrm{img}}^{\mathrm{co}}(u,v)&=\left|\mathcal{U}_{\mathrm{img}}^{\mathrm{co}}(u,v)\right|^2\\&=\left|\iint_{-\infty}^{+\infty}\mathcal{H}\left(u-\tilde{\xi},v-\tilde{\eta}\right)\mathcal{U}_{\mathrm{g}}\left(\tilde{\xi},\tilde{\eta}\right)\mathrm{d}\tilde{\xi}\mathrm{d}\tilde{\eta}\right|^2,\\\mathcal{I}_{\mathrm{img}}^{\mathrm{co}}(u,v)&=\left|\left(\mathcal{U}_{\mathrm{g}}\star\mathcal{H}\right)(u,v)\right|^2,\end{aligned} \tag{31}$$

where we observe that the *coherent illumination gives a linear system in the complex amplitude of the field* $\mathcal{U}_{\mathrm{g}}$. The previous result is related to the interference of coherent waves. If we now consider incoherent illumination, the mutual intensity is given by

$$\mathcal{J}_{\mathrm{g}}^{\mathrm{in}}\left(\tilde{\xi}_1,\tilde{\eta}_1;\tilde{\xi}_2,\tilde{\eta}_2\right)=\kappa\mathcal{I}_{g}\left(\tilde{\xi}_1,\tilde{\eta}_1\right)\delta\left(\tilde{\xi}_1-\tilde{\xi}_2,\tilde{\eta}_1-\tilde{\eta}_2\right), \tag{32}$$

where κ is a real constant, δ is the Dirac delta distribution, and $\mathcal{I}_{\mathrm{g}}$ is the intensity of the U_{g} field. The constant κ is the result of a simplification from statistical optics giving origin to Eq. 32. The constant depends on the degree of the extension of coherence when the evanescent wave phenomenon (Beran and Parrent, 1964) is taken fully into account. If the coherence extends over a wavelength, κ is equal to $\bar{\lambda}^2/\pi$, where $\bar{\lambda}$ is the mean wavelength (for a deeper discussion on incoherent illumination and the κ constant, see Goodman, 1985, §5.5.2). By replacing Eq. 32 in Eq. 27, the output (image) intensity obtained is

$$\begin{aligned}\mathcal{I}_{\mathrm{img}}^{\mathrm{in}}(u,v)&=\left|\mathcal{U}_{\mathrm{img}}^{\mathrm{in}}(u,v)\right|^2\\&=\kappa\iint_{-\infty}^{+\infty}\left|\mathcal{H}\left(u-\tilde{\xi},v-\tilde{\eta}\right)\right|^2\mathcal{I}_{\mathrm{g}}\left(\tilde{\xi},\tilde{\eta}\right)\mathrm{d}\tilde{\xi}\mathrm{d}\tilde{\eta},\\\mathcal{I}_{\mathrm{img}}^{\mathrm{in}}(u,v)&=\kappa\left(\mathcal{I}_{\mathrm{g}}\star\left|\mathcal{H}\right|^2\right)(u,v)=\kappa\left(\mathcal{I}_{\mathrm{g}}\star\mathcal{H}_{\mathrm{int}}\right)(u,v),\end{aligned} \tag{33}$$

where $\mathcal{H}_{\mathrm{int}}=|\mathcal{H}|^2$ is the *intensity impulse response*, also known as the PSF. In this case, *an optical system illuminated with incoherent light is linear in intensity*. Eq. 33 shows a commonly exploited fact; the output intensity is the convolution of the intensity PSF with ideal image intensity $\mathcal{I}_{\mathrm{g}}$.

2.4 Usual assumptions adopted in PSF modelling

PSF modelling articles generally implicitly assume specific hypotheses. We provide some of them in the following list:

- The scalar diffraction theory is valid.
- The lens law is verified, the paraxial approximation is valid, and the approximations discussed in Section 2.2 hold. These approximations allow us to discard quadratic phase terms from Fresnel's diffraction and exploit the simpler Fraunhofer diffraction formula.

- The incoming light from natural sources is assumed to be ideally incoherent. Then, the optical system is linear in intensity, as seen in Eq. 33.
- The PSF is considered to be spatially invariant in its isoplanatic region. In other words, the PSF is assumed not to change on the objects' typical length scales. This assumption allows us to use the convolution equation, i.e., Eq. 33, rather than the superposition integral, i.e., Eq. 12.

Although the previous assumptions are standard, certain precision levels require dropping simplifications. For example, the *Euclid* mission requirements on the PSF model accuracy as given in Laureijs et al. (2011) and Racca et al. (2016) is of 2×10^{-4} for the root mean square (RMS) error on each ellipticity component (δe_i^{PSF}) and 1×10^{-3} for the relative RMS error on the size ($\delta R^2_{\text{PSF}}/R^2_{\text{PSF}}$). The PSF model might have to include light polarisation to fulfil these extremely tight PSF requirements. Other assumptions might also be dropped for the precise imaging of widespread objects. This case might require discarding the spatially invariant assumption of the PSF or reducing the size of the isoplanatic region.

To conclude, the usual formulation of the PSF, i.e., the intensity of the impulse response, convolving an image seen in many articles, comes from the previous assumptions using the results from Eqs 23, 24, and 33. We rewrite this formula as follows:

$$\mathcal{I}_{\text{img}}(u,v) = \left(\mathcal{H}_{\text{int}} \star \mathcal{I}_g\right)(u,v), \tag{34}$$

where we remind the readers that (u,v) is the image plane, we have dropped the κ term from Eq. 33, and $\mathcal{H}_{\text{int}}$ is the intensity impulse response or PSF that is given as

$$\begin{aligned}\mathcal{H}_{\text{int}}(u,v|u_i,v_i) = \frac{a'^2}{\lambda^2 f^2}\Bigg| \iint_{-\infty}^{+\infty} \mathcal{P}^*(x,y|u_i,v_i) \\ \times \exp\left[\mathrm{j}\frac{2\pi}{\lambda}\mathcal{W}^*(x,y|u_i,v_i)\right] \\ \times \exp\left[-\mathrm{j}\frac{2\pi}{\lambda f}(ux+vy)\right]\mathrm{d}x\mathrm{d}y\Bigg|^2,\end{aligned} \tag{35}$$

where we are studying the PSF for a specific wavelength and focal plane position.

3 General observational model

We consider the PSF as the intensity impulse response, $\mathcal{H}_{\text{int}}$, of the imaging system under study to a point source. The concept of PSF (Born and Wolf, 1999) is used throughout many imaging applications, such as astronomical imaging (Schmitz, 2019; Liaudat et al., 2023), medical imaging (Dougherty and Kawaf, 2001; Joyce et al., 2018), or microscopy (Soulez et al., 2012; Debarnot et al., 2021a; Debarnot et al., 2021b). The central idea behind a PSF is that it represents transformations done to the imaged object by the imaging system. The PSF is, in a certain way, a characterisation of the imaging system. Considering incoherent illumination and what the hypotheses from the previous section hold, we can affirm that the optical system behaves linearly as in Eq. 34. Consequently, the PSF is considered the impulse response of the optical system and affects the ground truth image through a convolution operation. Focusing on astronomical imaging, the definition of the imaging system can vary between the different use cases and telescopes. For example, in a ground-based telescope, we will consider that the atmosphere belongs to the imaging system we are modelling. However, naturally, the atmosphere will not be considered in a space-based telescope. This article focuses on optical systems, which work with electromagnetic radiation with a wavelength close to the visible spectrum. For example, *Euclid* VIS instrument's theoretical wavelength range is from 550 to 900 nm.

The PSF describes the effects of the imaging system in the imaging process of the object of interest. The PSF is a convolutional kernel, as we have seen in Section 2.3. However, this convolutional kernel varies spatially, spectrally, and temporally. We give a non-exhaustive list that motivates each of these variations.

- *Spatial variations:* The optical system presents a certain *optical axis*, which is an imaginary line where the system has some degree of rotational symmetry. In simpler words, it can be considered as the direction of the light ray that produces a PSF in the centre of the focal plane for an unaberrated optical system. The angle of incidence is defined as the angle between an incoming light ray and the optical axis. The main objective of the optical systems that we study is to make the incoming light rays converge in the focal plane, where there will be some measurement instruments, e.g., a camera. Depending on the angle of incidence, the image will form in different positions in the focal plane. The path of the incoming light will be different for each angle of incidence, and therefore the system's response will be different too. In other words, the PSF will change depending on the angle of incidence or spatial position in the focal plane where the image is forming. Optical systems with wide focal planes, generally associated with wide field-of-views (FOVs), present significant PSF spatial variations.
- *Spectral variations:* Principally, due to the diffraction phenomena and its well-known wavelength dependence covered in Section 2, refractive[2] components of the optical system under study are also a source of spectral variations (Baron et al., 2022). Other sources of spectral variations are detector electronic components (Meyers and Burchat, 2015a) and atmospheric chromatic effects (Meyers and Burchat, 2015b).
- *Temporal variations:* The state of the telescope changes with respect to time, therefore the imaged object's transformation also changes. In space-based telescopes, high-temperature gradients cause mechanical dilations and contractions that affect the optical system. In ground-based telescopes, the atmosphere composition changes with time. Consequently, it temporally affects the response of the optical system, i.e., the PSF.

The PSF convolutional kernel varies with space, time, and wavelength. Once we have set up a specific wavelength and time to

2 Refraction refers to the change of direction in the propagation of a wave passing from one medium to another. Most of the wave energy is transmitted to the new medium. Reflection refers to the abrupt change of direction of the wave propagation due to a boundary between mediums. In this last case, most of the oncoming wave energy remains in the same medium.

analyse our system, we will have a different convolutional kernel for each position in the field of view. Let us refer to the PSF field $\mathcal{H}_{\text{int}}$ as all the PSF representing an optical system. Then, $\mathcal{H}_{\text{int}}(u, v; \lambda; t|u_i, v_i)$ is a specific PSF where (u_i, v_i) represents its centroid, i.e., the first-order moments. The same notation is maintained in Figure 4, where the (u, v) variables represent the image plane. We can define the PSF field as a varying convolutional kernel $\mathcal{H}_{\text{int}}: \mathbb{R}^2 \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^2 \to \mathbb{R}$. This definition would accurately describe how the PSF affects the images considering the assumptions from Section 2 are valid. We recall that we adopted the approximation that considers the PSF locally invariant in its *isoplanatic region* (Born and Wolf, 1999, §9.5.1); for an illustration, see Figure 7. This approximation means that in the vicinity of an observed object, we will consider that the PSF only varies with time and wavelength, thus facilitating the computation of the convolution. The close vicinity, or the isoplanatic region, will be defined as the postage stamp to image the object of interest. The typical galaxies observed for weak lensing have a comparable size with respect to the PSF size (for distribution of the relative galaxy to PSF size in the HSC survey, see Mandelbaum et al., 2018, Figure 7). Consequently, the approximation error is kept low as it is only done for small patches of the focal plane.

Let us define our object of interest with the subscript ground truth (GT), $\mathcal{I}_{\text{GT}}(u, v; \lambda; t|u_i, v_i)$, that is, the I_g object from Section 2, as a continuous light distribution $\mathcal{I}_{\text{GT}}: \mathbb{R}^2 \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^2 \to \mathbb{R}$. In this review, we are not considering transient objects, i.e., the time-dependence scale of the object is comparable with the exposure time used to image it. Therefore, we can ignore the temporal dependency of the GT object, $\mathcal{I}_{\text{GT}}(u, v; \lambda; t) \neq f(t)$. Let us write our general observational model that relates our GT object of interest, our PSF, and our observed image as follows:

$$I_{\text{img}}(\bar{u}, \bar{v}; t|u_i, v_i) = \mathcal{F}_p \left\{ \int_0^{+\infty} \mathcal{T}(\lambda) \; (\mathcal{I}_{\text{GT}} \star \mathcal{H}_{\text{int}})(u, v; \lambda; t|u_i, v_i) \; \mathrm{d}\lambda \right\} \circ N(\bar{u}, \bar{v}; t|u_i, v_i), \tag{36}$$

where $\mathcal{F}_p$ is a degradation operator discretising the image to $\mathbb{R}^{p \times p}$ that includes the image sampling from the instrument. The variables $(\bar{u}, \bar{v})$ denote the discrete (pixelised) version of the (u, v) variables. Then, $I_{\text{img}}(\bar{u}, \bar{v}; t|u_i, v_i) \in \mathbb{R}$ corresponds to the instrument's measurement at a single pixel $(\bar{u}, \bar{v})$, and $I_{\text{img},(\cdot;t|u_i,v_i)} \in \mathbb{R}^{p \times p}$ to the entire image. The variables (u_i, v_i) correspond to the centre location of the target object i. The instrument's transmission is represented by $\mathcal{T}: \mathbb{R}_+ \to \mathbb{R}_+$, a function with finite support, and $N_{(\cdot;t|u_i,v_i)} \in \mathbb{R}^{p \times p}$ corresponds to the noise affecting our observation and possibly a modelling error, where $\circ$ is some composition operator. We have carried out the spectral integration (Hopkins, 1957; Eriksen and Hoekstra, 2018) on the instrument's passband defined in τ. Although Eq. 36 provides a general observational model, it can be unpractical. The continuous functions $\mathcal{H}_{\text{int}}$, $\mathcal{T}$, and $\mathcal{I}_{\text{GT}}$ are practically inaccessible. We make several assumptions to simplify the problem.

(a) The continuous functions $\mathcal{H}_{\text{int}}$ and $\mathcal{I}_{\text{GT}}$ are well approximated by piece-wise constant functions over a regular grid in $\mathbb{R}^2$. We assume $\mathcal{H}_{\text{int}} \approx H$ and $\mathcal{I}_{\text{GT}} \approx I_{\text{GT}}$, where $H, I_{\text{GT}} \in \mathbb{R}^{P \times P}$ with $P \geq p$. The resolution of these two variables has to be greater or equal to the observation resolution.

(b) The noise is additive, i.e., $\circ \equiv +$, although the formulation could be adapted to consider other types of noise, e.g., Poisson.

(c) The degradation operator is approximated by its discrete counterpart, $\mathcal{F}_p \approx F_p$, where $F_p: \mathbb{R}^{P \times P} \to \mathbb{R}^{p \times p}$, which has been discretized in a regular grid. We assume that the degradation operator is *linear* and that includes pixellation, possibly downsampling, intra-pixel shifts, and linear detector effects.

(d) We keep the approximation that the PSF is locally constant within the postage stamp of $P \times P$ values of the target image.

(e) The integral can be well approximated by a discretised version using n_λ bins.

Taking into account the aforementioned assumptions, we can define our practical observational model as follows:

$$I_{\text{img}}(\bar{u}, \bar{v}; t|u_i, v_i) = F_p \left\{ \sum_{k=1}^{n_\lambda} T(\lambda_k) \; (I_{\text{GT}} \star H)(\bar{u}, \bar{v}; \lambda_k; t|u_i, v_i) \; \Delta\lambda_k \right\} + N(\bar{u}, \bar{v}; t|u_i, v_i), \tag{37}$$

where $I_{\text{img},(\cdot;t|u_i,v_i)} \in \mathbb{R}^{p \times p}$, T is a discretized version of $\mathcal{T}$, and $b^k = [b_0^k, b_1^k]$ is the kth wavelength bin centred in λ_k, with a width of $\Delta\lambda_k = b_1^k - b_0^k$.

3.1 Particular case: star observation

The case of star observations is of particular interest, as some stars in the FOV can be considered as a spatial impulse, i.e., $\mathcal{I}_{\text{star}}(u, v; \lambda|u_i, v_i) = \delta(u_i, v_i; \lambda) = f_{(u_i,v_i)}(\lambda)$. Therefore, if we plug the impulse in Eq. 37, we obtain a degraded observation of the PSF field. These observations will be crucial to constrain the PSF models. Unluckily, we do not always have access to the star's spectral variation, $f_{(u_i,v_i)}(\lambda)$. However, we dispose of complementary photometric observations that can be useful to characterise the spectral variations. These observations provide us with the star's spectral energy distribution (SED), which can be defined as the calibrated flux density as a function of wavelength, usually at low spectral resolution. The photometric observations are done in several spectral bands. Figure 8 shows the bands from the MegaCam instrument at the Canada–France–Hawaii Telescope (CFHT)[3]. For more information about SEDs and stellar photometry, we refer the readers to Hogg (2022). The SED is a normalised low-resolution sampling of the star's spectral variations. We can write the SED definition we will use as

$$\text{SED}_{b^k}(\lambda_k) = \frac{1}{z_{n_\lambda(b)}} \int_{b_0^k}^{b_1^k} f_{(u_i,v_i)}(\lambda) \; \mathrm{d}\lambda, \tag{38}$$

where we continued to use the b^k bin definition from Eq. 37, and $z_{n_\lambda(b)}$ is a constant used so that the SED is normalised to unity. We have that $\sum_{k=1}^{n_\lambda} \text{SED}_{b^k}(\lambda_k) = 1$, we continue by considering that the GT image in Eq. 37 is a star, and we use the spectral bins from the SED definition to discretise the spectral integration. Finally, we write the practical star observation model as

3 The filter curves can be downloaded from the Spanish Virtual Observatory (SVO) webpage, http://svo2.cab.inta-csic.es/svo/theory/fps/index.php.

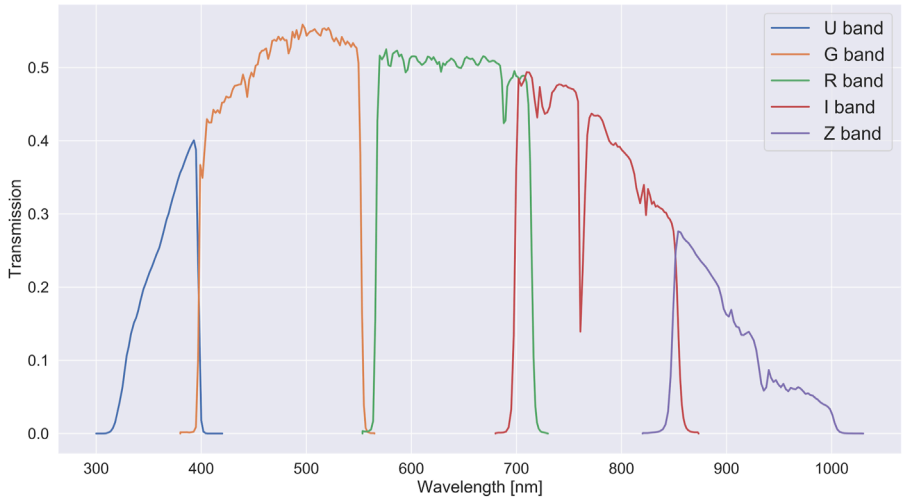


FIGURE 8
Third generation set of filters of the MegaCam instrument at the Canada–France–Hawaii Telescope that is currently being used for the Canada–France Imaging Survey. The transmission filter response includes the full telescope and 1.25 air masses of atmospheric attenuation. The full telescope includes mirrors, optics, and CCDs.

$$I_{\text{star}}(\bar{u},\bar{v};t|u_i,v_i) = F_p\left\{\sum_{k=1}^{n_\lambda} T(\lambda_k)\,\text{SED}_{b^k}(\lambda_k)\;H(\bar{u},\bar{v};\lambda_k;t|u_i,v_i)\;\Delta\lambda_k\right\} + N(\bar{u},\bar{v};t|u_i,v_i), \tag{39}$$

where we consider the star observation $I_{\text{star},(\cdot;t|u_i,v_i)} \in \mathbb{R}^{p\times p}$ as a degraded version of the PSF field $\bar{H}_{(\cdot;t|u_i,v_i)} \in \mathbb{R}^{p\times p}$.

4 PSF field contributors and related degradations

So far, we have described how the PSF interacts with the images we observe and how we can model an observation. However, we have not given much information about the different PSF field contributors and the different degradations represented by F_p in Eq. 37 that can occur when modelling observations. We provide a non-exhaustive list of contributors to the PSF field, sources of known degradations, and the atmosphere's effect on our PSF modelling problem.

4.1 Image coaddition

A fundamental contributor to the PSF is the choice of image coaddition scheme. A *coadded image* is a composite image created by combining multiple individual exposures of the same region of the sky in some way. This process can help increase the signal-to-noise ratio of the observation. Motivated by the analysis of the LSST data, Mandelbaum et al. (2022) have explored different coaddition schemes and studied how they affect the PSF of the resulting coadded image. In particular, Mandelbaum et al. (2022) have defined the schemes under which the coadded image accepts a well-defined PSF, i.e., the observation can be described by the convolution of an extended object and a uniquely defined coadded PSF. Bosch et al. (2017) have described the strategy for image and PSF coaddition in the HSC survey.

4.2 Dithering and super-resolution

Dithering consists of taking a series of camera exposures shifted by a fractional or a few pixel amount. There are several advantages of using a dithering strategy, which include the removal of cosmic rays and malfunctioning pixels, improving photometric accuracy, filling the gap between the detectors, and improving the sampling of the observed scene. Dithering allows estimating a sampling density of the images that is denser than the original pixel grid; in other words, to super-resolve the image. Regarding PSFs, it allows recovering Nyquist sampled PSF from undersampled observations. Bernstein (2002) studied the effect of dithering and the choice of pixel sizes in imaging strategies. Naturally, as we will see later, the dithering strategy is helpful for space-based telescopes thanks to their stability. In ground-based telescopes, the atmosphere constantly changes the PSF, making the dithering strategy less effective. However, a dithering strategy can be helpful if the telescope is equipped with adaptive optics technology, which will be described in Section 4.6. An example is the Spectro-Polarimetic High-contrast imager for Exoplanet REsearch (SPHERE) instrument (Beuzit et al., 2019) built for the European Southern Observatory's (ESO's) Very Large Telescope (VLT) in Chile.

Lauer (1999b) discussed the limiting accuracy effect of undersampled PSFs in stellar photometry and proposed ways to correct it with dithered data (Lauer, 1999a). Fruchter and Hook (2002) presented the widely used Drizzle algorithm that consists of shifting and adding the dithered images onto a finer grid. Rowe et al. (2011) proposed a linear coaddition method coined IMCOM to obtain a super-resolved image from several undersampled images. Hirata et al. (2023) later studied the use of IMCOM on simulations (Troxel et al., 2023) from the *Roman* Space Telescope,

while a companion paper by Yamamoto et al. (2023) explored its implications for weak-lensing analyses. Ngolè et al. (2015) proposed a super-resolution method coined SPRITE targeting the *Euclid* mission based on a sparse regularisation technique. More recent PSF models handle the undersampling of observations directly in their algorithms for estimating a well-sampled PSF field, as we will see later.

4.3 Optic-level contributors

These contributors affect the PSF by modifying the wave propagation in the optical system. In other words, they affect the wavefront's amplitude and phase.

- *Diffraction phenomena and aperture size:* As we have seen in Section 2, the diffraction phenomena occurring in the optical system play an essential role in the formation of the PSF. The size of the optical system's aperture and the wavelength of light being studied are of particular interest. Eq. 35 shows us that under some approximations, the PSF is the Fourier transform of the aperture. Therefore, the size of the aperture and the PSF are closely related. For example, if we consider an ideal circular aperture, its diffraction pattern is the well-known *Airy disk*. The relationship between the width of the PSF and diameter of the aperture is given by

$$\theta_{\text{FWHM}} = 1.025 \frac{\lambda}{d}, \tag{40}$$

where θ_{FWHM} is the full width at half maximum (FWHM) expressed in radians, λ is the wavelength of the light being studied, and d is the diameter of the aperture. The width of the PSF is a fundamental property of an optical system as it defines the resolution of the system. In other words, the PSF size defines the optical system's ability to distinguish small details in the image.

- *Optical aberrations:* These aberrations are due to imperfections in the optical elements, e.g., a not ideally spherical mirror or not perfectly aligning optical components. The optical aberrations play a significant role in the morphology of the PSF and can be modelled using the WFE introduced in the generalised pupil function from Eq. 22. Some aberrations have a distinctive name, e.g., coma, astigmatism, and defocus, and they represent a specific Zernike polynomial (Noll, 1976).
- *Surface errors or polishing effects:* One would ideally like perfectly smooth surfaces in mirrors and lenses. However, imperfections arise in the optical surfaces due to imperfect surface polishing. Krist et al. (2011) showed the measurement of surface errors (SFE) in the Hubble Space Telescope (HST). Gross et al. (2006, §3.5.2) gave a more in-depth analysis of surface errors focusing on the tolerancing of SFE. Figure 9A shows the surface errors measured for the Hubble Space Telescope (HST). Krist and Burrows (1995) studied HST's SFE before and after its iconic repair in 1993 with parametric and non-parametric (Gerchberg and Saxton, 1972) phase-retrieval algorithms.
- *Obscurations:* Complex optical systems have telescope designs where some elements can obscure some parts of the pupil. Obscurations are an essential contributor to PSF morphology and result from projecting a 3D structure onto the 2D focal plane. The resulting projection depends on the considered position of the focal plane. Accurate modelling of telescopes with wide-field imagers, e.g., *Euclid*, requires the computation of the obscuration's position dependence arising from the 3D projection. The *Euclid*'s obscurations are presented in Figure 6A. Fienup et al. (1993) and Fienup (1994) studied HST's obscuration from phase-retrieval algorithms and noticed a misalignment that caused a pupil shift.
- *Stray and scattered light:* Optical elements and instruments give rise to light reaching the detectors. Krist (1993) studied this problem for the HST. Storkey et al. (2004) developed methods to clean observations with scattered light from the SuperCOSMOS Sky Surveys (Hambly et al., 2001). Sandin (2014) studied the effect of scattered light on the outer parts of the PSF.
- *Material outgassing and ice contamination:* Material outgassing leads to molecular contamination that alters different properties of the imaging system. Water is the most common contaminant in cryogenic spacecraft, which then turns into thin ice films. A notable example is the *Gaia* mission which suffered from ice contamination (see Gaia Collaboration et al., 2016, §4.2.1) and required several decontamination procedures to slowly remove the ice from the optical system. Euclid Collaboration et al. (2023a) studied the ice formation and contamination for *Euclid*. The article also reviews the lessons learnt from other spacecraft on the topic of material outgassing. A companion paper by Euclid Collaboration et al. (2023a) is expected to be published soon that addresses the quantification of iced optics impact on *Euclid*'s data.
- *Chromatic optical components:* These components have a particular wavelength dependence, excluding the natural chromaticity due to diffraction. They are usually spectral filters and depend on the optical system design. A particular example is a dichroic filter which serves as an ideal band-pass filter. The *Euclid* optical system includes a dichroic filter which allows using both instruments, VIS and NISP, simultaneously as their passbands are disjoint. A dichroic filter is made of a stack of thin coatings of specific materials and thicknesses. Even if these components have a high-quality manufacturing process, they can induce significant chromatic variations in reflection that affect the PSF morphology. Baron et al. (2022) proposed a test bench to characterise *Euclid*'s dichroic filter and a numerical model of its chromatic dependence.
- *Light polarisation:* In Section 2, we studied the scalar diffraction theory, thus neglecting light polarisation. Firstly, the optical system can induce polarisation even when the incoming light is not polarised. Breckinridge et al. (2015) studied the effect of polarisation aberrations on the PSF of astronomical telescopes. The study of polarisation was carried out using Jones matrices (Jones, 1941). These matrices describe a ray's polarisation change when going through an optical system. For more information on polarisation aberrations, see McGuire and Chipman (1990, 1991) and Yun et al. (2011). Secondly, there are some regions in space where the incoming light has been polarised by different sources, e.g., galactic foreground dust. Lin et al. (2020) studied the impact of light polarisation

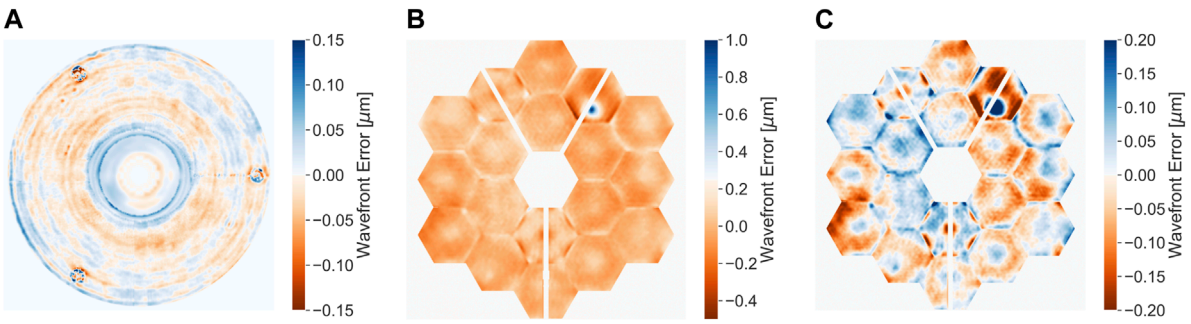


FIGURE 9
(A) Reduced range of variations to show the surface errors measured for the Hubble Space Telescope, where the scale has been reduced from ±22 nm to better show details. **(B,C)** Wavefront error measurements from the JWST Cycle 1 science operation on 30 July 2022 with the total and reduced range of variations, respectively.

on weak-lensing systematics for the Roman Space Telescope (Spergel et al., 2015). The study found that the systematics introduced by light polarisation is comparable to the Roman Space Telescope's requirements.

- *Thermal variations:* The thermal variations in a telescope introduce mechanical variations in its structure that affect the performance of the optical system. The origin of thermal variations is the strong temperature gradients due to the Sun's illumination. It is sometimes referred to as the *telescope's breathing* (Bély et al., 1993) for the periodical pattern consequence of its orbit. Thermal variations can introduce a small defocusing of the system that will change the PSF morphology. This phenomenon was first identified in the HST (Hasan et al., 1993). Nino et al. (2007) studied HST focus variations with temperature, and Lallo et al. (2006) studied HST temporal optical behaviour, where temperature variations play a principal role. Later works (Suchkov and Casertano, 1997; Makidon et al., 2006; Sahu et al., 2007) studied the impact of thermal variations, and consequently PSF variations, on different science applications. A Structural–Thermal–Optical Performance (STOP) test helps predict the thermal variations' impact on the optical system. This effect is naturally more significant in space-based telescopes as the temperature gradients in space are considerably more prominent than the ones found on the ground. Space-based telescopes located at the stable L_2 Lagrange point, e.g., *Euclid* and James Webb Space Telescope (JWST), are less prone to thermal variations than telescopes orbiting the Earth, e.g., HST.

As an example, Figures 9B,C show the measured optical contribution for the James Webb Space Telescope (JWST) PSF. Rigby et al. (2022) have presented a detailed analysis of JWST's state, since its commissioning, which includes its PSF.

4.4 Detector-level degradations

Detector-level degradations are related to the detectors being used and, therefore, to the intensity of the PSF. They affect the observed images through the degradation operator F_p from Eq. 37, and as we will use star images, or eventually other observations, to constrain PSF models, it is necessary to consider their effects. Some of these degradations are non-convolutional and will not be well modelled by a convolutional kernel. Nevertheless, we expect that image preprocessing steps will mainly correct these effects. However, the correction will not be perfect, and some modelling errors can propagate to the observations.

- *Undersampling and pixellation:* The EM wave that arrives at the detectors is a continuous function. The discrete pixels in the detectors integrate the functions and measure the intensity of the wave in their respective areas. We name this process *pixellation*, also known as *sampling*. Some authors, e.g., Anderson and King (2000), Bernstein (2002), and Kannawadi et al. (2016), have defined an *effective* PSF as the convolution of the optical PSF, i.e., the flux distribution at the focal plane from a point source, with the pixel response of the instrument, e.g., a 2D top-hat function. High et al. (2007) performed an early study on the effects of pixellation in WL and the choice of pixel scale for a WL space-based mission. Krist et al. (2011, §3) gave some insights on the pixellation effects for the HST. Two aspects of pixellation play a crucial role in PSF modelling. Firstly, the sampling is done with the same grid, but it is indispensable to consider that the continuous function is not necessarily centred on the grid. This difference means that intra-pixel shifts between the different pixellations will be found. Figure 10 shows how two pixel representations of the same light profile change due to two different pixellations. When optimising a PSF model to reproduce some observed stars, the centroids of both images must be the same. Suppose the image centroids are the same, and the underlying model represents the observations satisfactorily. In such a case, the residual image between the two pixelated images will be close to zero. If the centroids are not the same, the residual can be far from zero even though the model is a good representation of the observation, as illustrated in the residual image in Figure 10. The second aspect is related to the Nyquist–Shannon sampling theorem. The theorem states the required number of samples that we have to use to determine perfectly a signal of a given bandwidth. In the telescopes we study, the bandwidth and number of samples are related to the aperture's diameter and pixel size. Depending on the telescope's design,

the sampling may not verify the Nyquist–Shannon theorem. If the images are undersampled, i.e., the theorem is not verified, a super-resolution step is required in the PSF modelling, which is the case in *Euclid*. Using an observation strategy with dithering, as described in Section 4.2, can significantly mitigate the undesired effects of undersampling and pixellation. Kannawadi et al. (2021) studied ways of mitigating the effects of undersampling in WL shear estimations using metacalibration (Huff and Mandelbaum, 2017; Sheldon and Huff, 2017; Sheldon et al., 2020), which is a method for measuring WL shear from well-sampled galaxy images. Finner et al. (2023) studied near-IR weak-lensing (NIRWL) measurements in the CANDELS fields from HST images. The authors find that the most significant contributing systematic effect to WL measurements is caused by undersampling.

- *Optical throughput and CCD quantum efficiency (QE):* The optical throughput of the system is the combined effect of the different elements that compose the optical system, such as mirrors and optical elements like coatings (Venancio et al., 2016). The filter being used in the telescope forms a part of the optical throughput, as can be seen in Figure 8 for the MegaCam set of filters. Figure 8 also includes the CCD QE, which describes the sensibility of the CCD to detect photos of different wavelengths. Commonly, CCDs do not have a uniform response to the different wavelengths. Therefore, we must multiply the CCD QE with the telescope's optical throughput to compute the total transmission.
- *CCD misalignments:* Ideally, we expect that all the CCDs in the detector lie in a single plane that happens to be the focal plane of the optical system. However, this is not the case in practice, as there might be small misalignments between the CCDs that introduce small defocuses that change from CCD to CCD. For a study of this effect for the Vera C. Rubin Observatory, see Jee and Tyson (2011, Figure 8).
- *Guiding errors:* Even if space telescopes are expected to be very stable during observations thanks to the Attitude and Orbit Control System (AOCS), there will exist a small residual motion that is called pointing jitter. The effect on the observation is the introduction of a small blur that can be modelled by a specific convolutional kernel that depends on the pointing time series. Fenech Conti (2017, §4.8.3) proposed to model the effect for *Euclid* with a Gaussian kernel.
- *Charge transfer inefficiency:* CCD detectors are in charge of converting incoming photons to electrons and collecting them in a potential well in the pixel during an exposure. The charge on each pixel is read when the exposure is complete. The collected electrons are transferred through a chain of pixels to the edge of the CCD, amplified, and then read. High-energy radiation above the Earth's atmosphere gradually damages the CCD detector (Prod'homme et al., 2014b; Prod'homme et al., 2014a). The silicon damage in the detectors creates traps for the electrons that are delayed during the reading procedure. This effect is known as charge transfer inefficiency (CTI), producing a trailing of bright objects and blurring the image. This effect is noticeably significant for space telescopes, given the harsh environment. CTI effects are expected to be corrected in the VIS image preprocessing. Rhodes et al. (2010) carried out a study on the impact of CTI on WL studies. Massey et al. (2009) developed a model to correct CTI for the HST and later improved it in Massey et al. (2014).
- *Brighter-fatter effect:* The assumption that each pixel photon count is independent of its neighbours does not hold in practice. There is a photoelectron redistribution in the pixels as a function of the number of photoelectrons in each pixel. The brighter-fatter effect (BFE) is due to the accumulation of charge in the pixels' potential wells and the build-up of a transverse electric field. The effect is stronger for bright sources. Antilogus et al. (2014) studied the effect and observed that the images from the CCDs do not scale linearly with flux, so bright star sizes appeared larger than fainter stars. Guyonnet et al. (2015) and Coulton et al. (2018) proposed methods to model and correct this effect. The preprocessing of VIS images is supposed to correct for the BFE, but there might be some residuals.
- *Wavelength-dependent sub-pixel response:* There exists a charge diffusion between neighbouring pixels in the CCD. Niemi et al. (2015) studied this effect for *Euclid*'s VIS CCD and modelled the response of the CCD. They proposed to model the effect as a Gaussian convolutional kernel where the standard deviations of the 2D kernel are wavelength dependent: $\sigma_x(\lambda)$ and $\sigma_y(\lambda)$. They measured the proposed model with a reference VIS CCD. Krist (2003) studied the charge diffusion in the HST and proposed spatially varying blur kernels to model the effect.
- *Noise:* There are several noise sources in the measurements. *Thermal noise* (Nyquist, 1928) refers to the signal measured in the detector due to the random thermal motion of electrons which is usually modelled as Gaussian. *Readout noise* (Basden et al., 2004) refers to the uncertainty in the photoelectron count due to imperfect electronics in the CCD. *Dark-current shot noise* (Baer, 2006) refers to the random generation of electrons in the CCD, and even though it is related to the temperature, it is not Gaussian. There are also unresolved and undetected background sources that contribute to the observation noise. These are the statistics of the predominant noise that depends on the imaging setting of the instrument and its properties.
- *Tree rings and edge distortions:* There exist electric fields in the detector that are transverse to the surface of the CCD. The origin of these fields includes doping gradients or physical stresses on the silicon lattice. This electric field displaces charge, modifying the effective pixel area. Consequently, it changes the expected astrometric and photometric measurements. This electric field also generates concentric rings, *tree rings*, and bright stripes near the boundaries of the CCD, *edge distortions*. Given the close relationship between this effect and the detector, its importance depends strongly on the instrument being used. This effect is unnoticeable in the MegaCam used in the Canada–France Imaging Survey (CFIS) as it depends on the CCD design. However, it is a major concern in the Dark Energy Camera that is used in the Dark Energy Survey (DES), as was shown by Plazas et al. (2014). Jarvis et al. (2020, Figure 9) illustrated the consequence of tree rings in PSF modelling.
- *Other effects:* These include detector nonlinearity (Stubbs, 2014; Plazas et al., 2016), *sensor interpixel correlation* (Lindstrand, 2019), *interpixel capacitance* (McCullough,

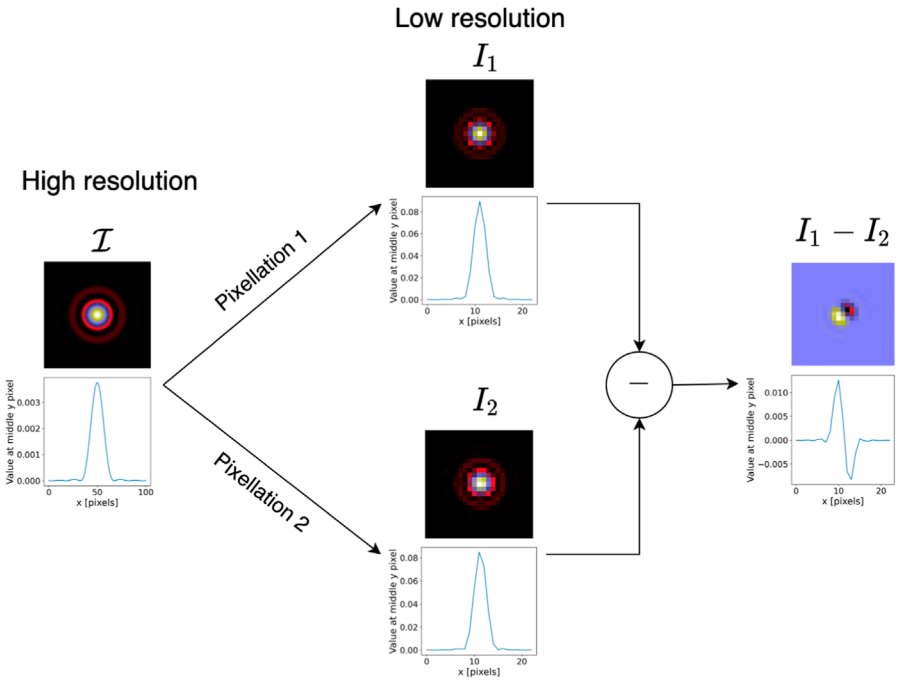


FIGURE 10
Example of two different pixellations on the same high-resolution image representing an Airy PSF. The difference between the two pixellations is an intra-pixel shift of $(\Delta x, \Delta y) = (0.35, 0.15)$ between them. Figure reproduced from Liaudat (2022).

2008; Kannawadi et al., 2016; Donlon et al., 2018), *charge-induced pixel shifts* (Gruen et al., 2015), *persistence* (Smith et al., 2008a; Smith et al., 2008b), *reciprocity failure (flux-dependent nonlinearity)* (Bohlin et al., 2005; Biesiadzinski et al., 2011), and *detector analogue-to-digital nonlinearity.*

4.5 Atmosphere

The atmosphere plays a central role in ground-based telescopes' PSFs. For an in-depth study of the subject, see Roddier (1981). How the atmosphere affects our images will strongly depend on the exposure time used to image an object. The PSF induced by the atmosphere for a very short exposure will look like a speckle, while a long exposure will produce a PSF that resembles a 2D Gaussian, or more precisely, a Moffat profile (Moffat, 1969). Figure 12 shows examples of atmospheric PSFs with different exposure times. The atmosphere's effect on the PSF for a long exposure can be approximated by the effect of a spatially varying low-pass filter, thereby broadening the PSF and limiting the telescope's resolution. Astronomers usually use the term *seeing* to refer to the atmospheric conditions of the telescope, and it is measured as the FWHM of the PSF. The loss of resolution due to the atmosphere is one of the main motivations for building space telescopes like *Euclid* and *Roman*, where the PSF is close to the diffraction limit and very stable.

The atmosphere is a heterogeneous medium whose composition changes with the three spatial dimensions and time. The inhomogeneity of the atmosphere affects the propagation of light waves that arrive at the telescope. Instead of supposing that the incoming light waves are plane, as emitted by the faraway source under study, these waves already have some phase lags or leads with respect to an ideal plane wave. The atmosphere introduces a WFE contribution to the optical system. These effects can be resumed as an effective phase-shifting plate, $\Phi_{\text{eff}}(x, y, t)$. However, calculating this effective plate is cumbersome as it involves having a model of the atmosphere and integrating the altitude, z, so that we have the spatial distribution, (x, y), of the effective WFEs. The model of the atmosphere is represented by the continuous $C_n^2(z)$ (Roddier, 1981) profile, which represents the variations of the refractive index due to atmospheric turbulence as a function of height. However, the $C_n^2(z)$ is challenging to model and measure, and even if it is possible, it is computationally expensive to exploit.

We can discretise the integral over the altitude into M thin phase screens of variable strengths at different altitudes to simulate the effect of the atmosphere. Each phase screen will have specific properties and move at different speeds in different directions. These assumptions are known as the frozen flow hypothesis. Each phase screen will be characterised by its power spectrum that can be modelled by a von Kármán model of turbulence (Kármán, 1930). The power spectrum of the atmosphere's WFE contribution is given by

$$\Psi(\nu) = 0.023\, r_0^{-5/3} \left(\nu^2 + \frac{1}{L_0^2} \right)^{-11/6}, \tag{41}$$

where ν is the spatial frequency, r_0 is the Fried parameter, and L_0 is the outer scale. Both parameters, r_0 and L_0, are generally expressed in metres. The Fried parameter relates to the turbulence amplitude, and the outer scale relates to the correlation length. For an example of atmospherical phase screens, see Figure 11. For lengths longer than L_0, the power of the turbulence asymptotically flattens. If we

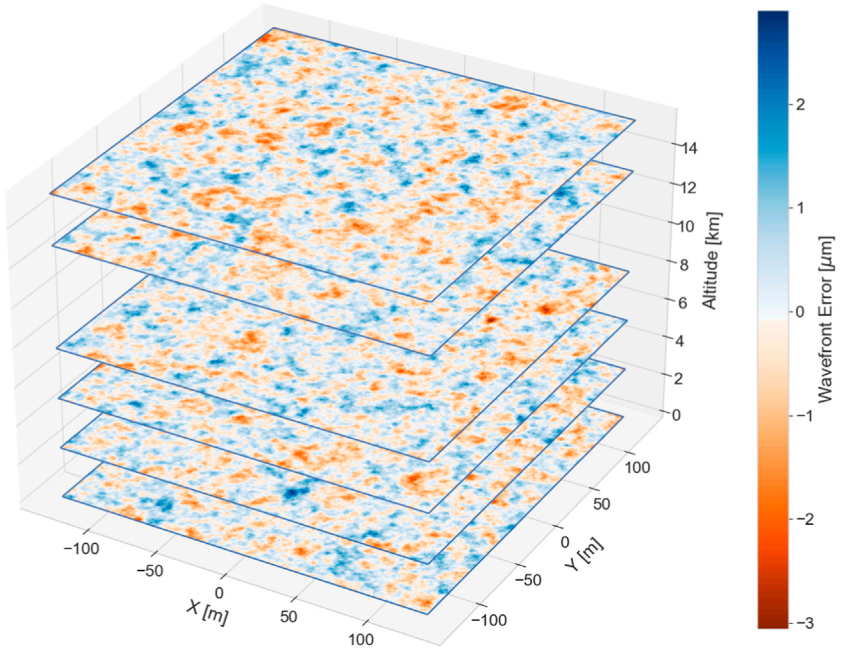


FIGURE 11
Illustration of six von Kármán phase screen layers at different altitudes simulated for LSST. The simulations were produced with the GalSim package (Rowe et al., 2015) using the parameters from Jee and Tyson (2011).

take the limit of L_0 to infinite, we converge to the Kolmogorov model of turbulence (Kolmogorov, 1991). For more information on electromagnetic wave propagation in turbulence, see Sasiela (1994).

Once the phase screens, $\Phi_m(x,y|u_i,v_i)$, have been simulated following Eq. 41, the temporal variation of the screen has to be taken into account. The phase screens contribute to the WFEs of the PSF, which is why it depends on the pupil plane variables (x,y). The temporal variation is usually modelled with the wind's properties at the phase screen's reference altitude. We describe the wind with two components, v_u and v_v, where we have assumed that $v_z = 0$. We then obtain the effective phase screen by a weighted average of the phase screens at the different altitudes as

$$\Phi_{\text{eff}}(x,y;t|u_i,v_i) = \sum_{m=1}^{M} c_m \Phi_m(x,y;t|u_i,v_i), \tag{42}$$

where $\{c_m\}$ are relative weights corresponding to the different phase screen. The difficulty of modelling the atmosphere is that the time scales are comparable with the exposure time. Therefore, the PSF that we estimate for a given time snapshot will change with respect to another PSF at another snapshot within the same camera exposure. This change means that we have to integrate the instantaneous PSF over time to model the PSF physically, which corresponds to

$$I_{\text{img}}(\bar{u},\bar{v}|u_i,v_i) = \int_{t_0}^{t_0+T_{\text{exp}}} I_{\text{img}}(\bar{u},\bar{v};t|u_i,v_i)\ dt, \tag{43}$$

where $I_{\text{img}}(\bar{u},\bar{v},t|u_i,v_i)$ is the instantaneous image of the object affected by the PSF $\mathcal{H}(u,v;\lambda;t|u_i,v_i)$, t_0 is a random initial time, and T_{exp} is the exposure time.

Finally, we have to choose the time step size to discretise the integral from Eq. 43. Each instantaneous PSF will look like a speckle. Once we add them up in the integral, the PSF starts to become rounder and smoother. Figure 12 shows examples of atmospheric PSFs using different exposure times that were simulated using six-phase screens using the parameters from Jee and Tyson (2011) that correspond to an LSST-like scenario. It is interesting to see how the short-exposure PSF looks like a speckle, and then the profile becomes more and more smooth as the exposure time increases. As a reference, the exposure time used for the r-band observations in CFIS is 200 s[4]. de Vries et al. (2007) studied the PSF ellipticity change due to atmospheric turbulences as a function of the exposure time. They observed that the ellipticity of the PSF decreases its amplitude as the exposure time increases.

Another effect that should be considered is *atmospheric differential chromatic refraction*. This effect represents the refraction due to the change of medium from vacuum to the Earth's atmosphere. The effect varies as a function of the zenith angle and wavelength. Meyers and Burchat (2015b) performed a study on the impact of the atmospheric chromatic effect on weak lensing for surveys like LSST and DES.

Heymans et al. (2012) performed a study on the impact of atmospheric distortions on weak-lensing measurement with real data from CFHT. They characterised the ellipticity contribution of the atmosphere to the PSF for different exposure times. To achieve this, they computed the two-point correlation function of the residual PSF ellipticity between the observations and a PSFEx-like PSF model (described in detail in Section 5). Salmon et al. (2009) studied the image quality and the observing environment at the CFHT (Xin et al., 2018) and carried out a study of the PSF and the variation of its width with time and wavelength for

4 https://www.cfht.hawaii.edu/Science/CFIS/

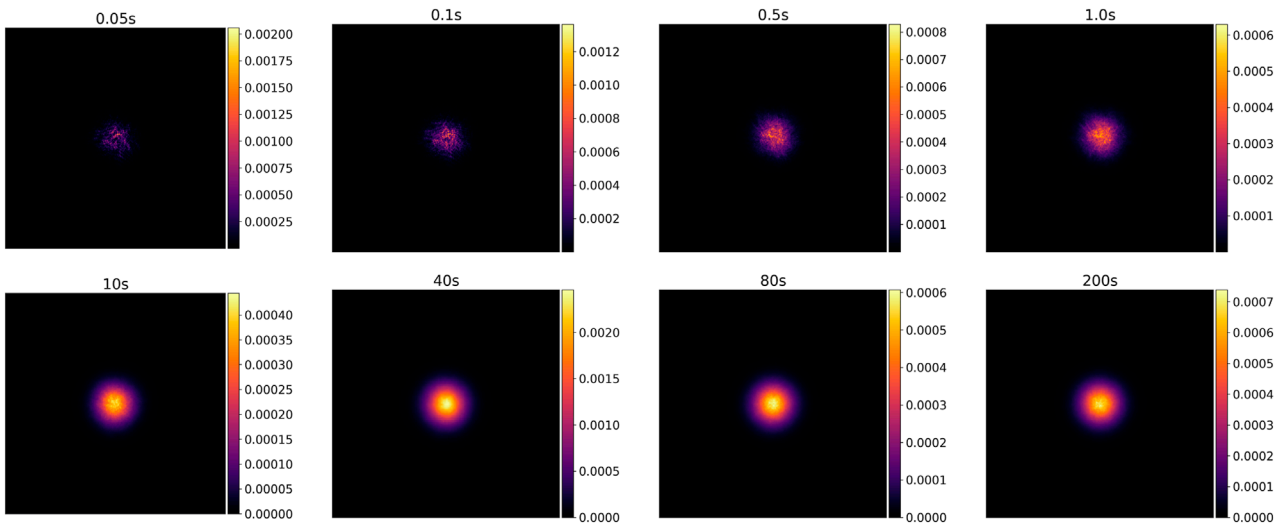


FIGURE 12
Example of atmospheric PSFs with different exposure times. The simulation was done using the atmospherical parameters from Jee and Tyson (2011) for an LSST-like scenario.

the Sloan Digital Sky Survey (SDSS) (Gunn et al., 1998; York et al., 2000). Jee and Tyson (2011) carried out a simulation study to evaluate the impact of atmospheric turbulence on weak-lensing measurement in LSST. Jee and Tyson (2011) used the atmospherical parameters from (Ellerbroek, 2002) that were measured in the LSST site in Cerro Pachón, Chile. There is an ongoing project at LSST DESC[5] to leverage atmospheric and weather information at or near the observation site to produce realistically correlated wind and turbulence parameters for atmospheric PSF simulations.

Another way to simulate the atmosphere and the PSF is to use a photon Monte Carlo approach. This line of work was carried out in Peterson et al. (2015, 2019, 2020) with a simulator available coined[6] that is capable of simulating several telescopes including the Vera C. Rubin observatory and JWST. The method consists of sampling photons from astronomical sources and simulating their interactions with their models of the atmosphere, the optics and the detectors. PhoSim is characterised by being a remarkably fast simulator regarding the level of simulation complexity handled. This simulator has proved helpful in several studies (Walter, 2015; Xin et al., 2015; Beamer et al., 2015; Carlsten et al., 2018; Xin et al., 2018; Burke et al., 2019; Sánchez et al., 2020; Nie et al., 2021b; Euclid Collaboration, 2023b; Euclid Collaboration, 2023c; Merz et al., 2023). PhoSim is a powerful simulation tool that can help to study the PSF of systems by following a forward model approach where simulations are compared with observations (Chang et al., 2012; Chang and Peterson, 2017). The PhoSim parameters can then be fitted or modified to match the simulation output with the observed images. Another known simulation software[7], Rowe et al. (2015), incorporates options to simulate atmospheric PSFs from phase-screen exploiting the photon Monte Carlo approach from Peterson et al. (2015).

To conclude, we have seen that it is possible to develop a physical model of the atmosphere based on the optical understanding that we have from Section 2 and the studies of atmospheric turbulence of Kármán and Kolmogorov. However, this approach has two inconveniences. Firstly, the approach requires physical measurements of the atmosphere at the telescope's site, which is not always available. Secondly, it is computationally expensive, as there is an altitude and a temporal integration to handle varying atmospherical properties and reach the exposure time, respectively. In practice, it is required to use long exposure times to obtain deeper observations, i.e., observing fainter objects that are important for weak-lensing studies. This fact simplifies the PSF modelling task as the long temporal integration smooths the PSF profile and PSF spatial variations over the FOV. Therefore, a data-driven approach to modelling the PSF can offer a feasible and effective solution in this scenario.

4.6 Adaptive optics

An alternative approach to work with ground-based observations affected by the atmosphere is to use adaptive optics (AO) systems (Beckers, 1993). This technology significantly improves the observation resolution in ground-based telescopes that is severely limited due to the atmosphere, as we have seen in Section 4.5. An AO system tries to counteract the effect of the atmosphere on the incoming wavefront by changing the shape of a deformable mirror. The key components of the AO system are wavefront sensors (WFS), wavefront reconstruction techniques, and deformable mirrors, which operate together inside a control loop. The WFS provides information about the incoming wavefront and usually incorporates a phase-sensitive device. The wavefront reconstruction has to compute a correction vector for

5 github.com/LSSTDESC/psf-weather-station
6 bitbucket.org/phosim/phosim_release/wiki/Home
7 github.com/GalSim-developers/GalSim

the deformable mirrors by estimating the incoming wavefront from the WFS information. The control loop works in real-time sensing and modifies the deformable mirrors so that the wavefront received by the underlying instrument is free from optical path differences introduced by the atmosphere. Davies and Kasper (2012) have provided a detailed review of the AO systems for astronomy. The LSST that carries out weak-lensing studies contains an AO system (Angeli et al., 2014; Neill et al., 2014; Thomas et al., 2016). Exoplanet imaging studies have greatly benefited from AO systems and impose strict requirements for these systems. Some examples are the SPHERE instrument in the VLT (Beuzi et al., 2019) and the Gemini Planet Imager (Graham et al., 2007; Macintosh et al., 2014; Macintosh et al., 2018) in the Gemini South Telescope.

5 Current PSF models

Let us now discuss some of the most known PSF models, which can be divided into two main families, *parametric* and *non-parametric*, also known as *data-driven*.

5.1 Parametric PSF models

This family of PSF models is characterised by trying to build a physical optical system model that aims to be as close as possible to the telescope. Once the physical model is defined, a few parameters are estimated using star observations. Such an estimation, also called calibration, is required as some events, like launch vibrations, ice contaminations, and thermal variations, introduce significant variations in the model. These events prevent a complete on-ground characterisation from being a successful model. Parametric models are capable of handling chromatic variations of the PSF as well as complex detector effects. Nevertheless, parametric models have only been developed for space missions and are custom-made for a specific telescope. The parametric model is compelling if the proposed PSF model matches the underlying PSF field. However, if there are mismatches between both models, significant errors can arise due to the rigidity of the parametric models. The difficulty of building a physical model for the atmosphere, already discussed in Section 4.5, makes them impractical for ground-based telescopes.

The parametric model Tiny Tim[8] (Krist, 1993; Krist, 1995; Krist et al., 2011) has been used to model the PSF of the different instruments on board the HST. The Advanced Camera for Surveys (ACS) in the HST was used to image the Cosmic Evolution Survey (COSMOS), which covers a $2-\text{deg}^2$ field that was used to create a widely used space-based weak-lensing catalogue. The first WL shape catalogue used the Tiny Tim model (Leauthaud et al., 2007). Rhodes et al. (2007) studied the stability of the HST's PSF, noticing a temporal change of focus in the images. In addition to the parametric Tiny Tim model, Anderson and King (2000) developed the concept of *effective PSF*, which is the continuous PSF arriving at the detectors, i.e., Equation 35, convolved with the pixel-response function of the detector. They proposed an algorithm to model the effective PSF iteratively from observed stars and continued the work on the effective PSF, adding some improvements and detailed a model for the HST instruments ACS and Wide Field Camera (WFC). Hoffmann and Anderson (2017) then carried out a comparison between Tiny Tim and the effective PSF approach for ACS/WFC. The study showed that the effective PSF approach consistently outperforms the Tiny Tim PSFs, exposing the limitations of parametric modelling. Anderson (2016) describes the adoption of the effective PSF approach applied to Wide Field Camera 3 (WFC3/IR) observations, which are undersampled. The software Photutils[9] (Bradley et al., 2022) provides an implementation of the effective PSF approach from Anderson and King (2000) with enhancements from Anderson (2016). Schrabback, T. et al. (2010) used a data-driven PSF model based on PCA when studying the COSMOS field.

The early severe aberrations of HST's optical system were an important driver of phase-retrieval algorithms. Several efforts to characterise HST were published in an *Applied Optics* special issue (Breckinridge and Wood, 1993). Solving the phase-retrieval problem provides a reliable approach to characterise the optical system accurately. Fienup (1993) studied new phase-retrieval algorithms and Fienup et al. (1993) presented several results characterising HST's PSF.

Regarding recently launched space telescopes, *Euclid*'s VIS parametric model constitutes the primary approach for *Euclid*'s PSF modelling. The model will soon be published and used to work with observations from *Euclid*. JWST has a Python-based simulating toolkit, WebbPSF[10] (Perrin et al., 2012, 2014). Recent works have developed and compared data-driven PSF models for JWST's NIRCam (Nardiello et al., 2022; Zhuang and Shen, 2023).

5.2 Data-driven (or non-parametric) PSF models

The *data-driven* PSF models, also known as *non-parametric*, only rely on the observed stars to build the model in pixel space. They are blind to most of the physics of the inverse problem. These models assume regularity in the spatial variation of the PSF field across the FOV and usually differ in how they exploit this regularity. Data-driven models can easily adapt to the current state of the optical system. However, they have difficulties modelling complex PSF shapes occurring in diffraction-limited settings. One limitation shared by all the data-driven models is their sensitivity to the available number of stars to constrain their estimation. A low star number implies that there might be not enough stars to sample the spatial variation of the PSF. When the number of stars in an FOV falls below some threshold, the model built is usually considered unusable for WL purposes. This family of models has been widely used for modelling ground-based telescope PSFs. Nevertheless, they are not yet capable of successfully modelling the chromatic variations in addition to the spatial variations and super-resolution.

We proceed by describing several PSF models in chronological order. The first models, described in more detail, were used

8 github.com/spacetelescope/tinytim

9 github.com/astropy/photutils

10 github.com/spacetelescope/webbpsf

to process real data from different surveys, except for Resolved Component Analysis (Ngolè et al., 2016). The latter models are worth mentioning but have yet to be used to produce a WL shape catalogue with all the validation and testing it implies.

5.2.1 Shape interpolation

The first approach for PSF modelling consisted of estimating some parameters of the PSF at the positions of interest. It was done for early studies in WL and is closely related to the widely employed galaxy shape measurement method KSB (Kaiser et al., 1995). This precursor method only required the PSF's ellipticity and size at the positions of the galaxies. Therefore, a full-pixel image of the PSF was unnecessary. Then, the KSB method can correct the galaxy shape measurement for the effects of the PSF. The method to interpolate the shape parameters to other positions is usually a polynomial interpolation. For example, this was the case for the early WL study of Canada-France-Hawaii Telescope Legacy Survey (CFHTLS) (Fu, L. et al., 2008). Gentile et al. (2013) reviewed the different interpolation methods and studied their performance for WL studies. Viola et al. (2011) performed a study showcasing different biases present in the KSB (Kaiser et al., 1995) method. These biases are a consequence of problematic KSB assumptions: 1) KSB gives a shear estimate per individual image and then takes an average, while WL shear should be estimated from averaged galaxy images; 2) KSB works under the assumption that galaxy ellipticities are small, but in the context of weak lensing, what is considered 'small' pertains to the alteration in ellipticity caused by the shear; and 3) KSB gives an approximate PSF correction that only holds in the limit of circular sources and does not invert the convolution with the PSF. Recent WL studies no longer use this approach. The WL studies have evolved to more sophisticated galaxy shape measurement techniques that require a full pixel image of the PSF at the position of galaxies.

5.2.2 Principal component analysis

The principal component analysis is a widely known method for multivariate data analysis and dimensionality reduction. Let us start with a set of star observations in $\mathbb{R}^{p^2}$ that we concatenate in a matrix $\bar{\mathbf{I}} = [\bar{I}_1, \ldots, \bar{I}_n]$. We have flattened the 2D images into an array to simplify expressions. One would like to represent the observations with r components $\{X_i\}_{i=1}^{r}$ in $\mathbb{R}^{p^2}$, where $r \geq n$, assuming that $p^2 > n$. The PCA method gives r orthonormal components in $\mathbb{R}^{p^2}$ which define directions in the $\mathbb{R}^{p^2}$ space where the variance of the data set $\bar{\mathbf{I}}$ is maximized.

If n components are used to represent the observations, then the learned components in the PCA procedure represent a basis of the subspace spanned by the observations or the columns of $\bar{\mathbf{I}}$. The method can be interpreted as a linear transformation to a new representation with orthogonal components. As it is usual to observe regularity in the spatial variations of the PSF, most of the data set variability can be described with a few components. Then, one can only use the first r principal components and achieve a dimensionality reduction of the observations. The dimensionality reduction technique allows denoising the model as the observational noise cannot be represented with r components and only the PSF trends are well described.

The PCA method was used to model the PSF for the SDSS (Lupton et al., 2001), although it was referenced as the Karhunen–Loève transform. Jarvis and Jain (2004) then proposed its use in a WL context. Jee et al. (2007) used the PCA to model the spatial and temporal variations of the HST PSF. Jee and Tyson (2011) also used the PCA to model the PSF in LSST simulations. HST's COSMOS catalogue (Schrabback, T. et al., 2010) used the PCA to model the PSF. The PCA showcased the utility and robustness of data-driven methods and the importance of using a pixel representation of the PSF and is the precursor of several of the following models.

5.2.3 PSFEx

PSFEx[11] (Bertin, 2011) has been widely used in astronomy for weak-lensing surveys, e.g., DES year 1 (Zuntz et al., 2018), HSC (Mandelbaum et al., 2017), and CFIS (Guinot et al., 2022). It was designed to work together with the SExtractor (Bertin and Arnouts, 1996) software which builds catalogues from astronomical images and measures several properties of the observed stars. The PSFEx models the variability of the PSF in the FOV as a function of these measured properties. It builds independent models for each CCD in the focal plane and works with polychromatic observations. It cannot model the chromatic variations of the PSF field. The model is based on a matrix factorisation scheme, where one matrix represents the PSF features and the other matrix represents the feature weights. Each observed PSF is represented as a linear combination of PSF features. The feature weights are defined as a polynomial law of the selected measured properties. This choice allows having an easy interpolation framework for target positions. In practice, the properties that are generally used are both components of the PSF's FOV position. The PSF features are shared by all the observed PSFs and are learnt in an optimisation problem. The PSF reconstruction at a FOV position (u_i, v_i) can be written as

$$\bar{I}^{\text{PSFEx}}_{\text{star}}(\bar{u}, \bar{v} | u_i, v_i) = F^{\text{PSFEx}} \underbrace{\left\{ \sum_{\substack{p,q \geq 0 \\ p+q \leq d}} u_i^p \, v_i^q \, S_{p,q}(\bar{u}, \bar{v}) + S_0(\bar{u}, \bar{v}) \right\}}_{\bar{H}^{\text{PSFEx}}(\bar{u}, \bar{v} | u_i, v_i)}, \quad (44)$$

where $\bar{I}^{\text{PSFEx}}_{\text{star},(\cdot|u_i,v_i)} \in \mathbb{R}^{p \times p}$ is the PSFEx reconstruction of the observed star $\bar{I}_{(\cdot|u_i,v_i)}$, $S_{p,q} \in \mathbb{R}^{P \times P}$ represents the learned PSF features or *eigenPSFs*, $S_0 \in \mathbb{R}^{P \times P}$ represents the first guess of the PSF, the polynomial law is defined to be of degree d, and F^{PSFEx} represents the degradations required to match the model with the observations. The model's PSF reconstruction is represented by $\bar{H}^{\text{PSFEx}}$. The first guess can be computed as a function of the median of all the observations. The dimensions p and P are the same if no downsampling operation is included in F^{PSFEx}.

The PSF features are learnt in an optimisation problem that aims to minimise the reconstruction error between the PSFEx model and observations, which are given by

$$\min_{\substack{S_{p,q} \\ \forall p,q \geq 0,\, p+q \leq d}} \left\{ \sum_{i=1}^{n_{\text{obs}}} \left\| \frac{\bar{I}_{(\cdot|u_i,v_i)} - \bar{I}^{\text{PSFEx}}_{\text{star},(\cdot|u_i,v_i)}}{\widehat{\sigma}_i} \right\|_F^2 + \sum_{\substack{p,q \geq 0 \\ p+q \leq d}} \left\| T_{p,q} \, S_{p,q} \right\|_F^2 \right\}, \quad (45)$$

11 github.com/astromatic/psfex

where $\widehat{\sigma}_i^2$ represents the estimated per-pixel variances, $\bar{I}$ represents the noisy observations, and $\|\cdot\|_F$ the Frobenius norm of a matrix. The second term in Eq. 45 corresponds to a Tikhonov regularisation, where $T_{p,q}$ represents some regularisation weights to favour smoother PSF models. The PSF recovery at target positions is straightforward. One has to introduce new positions in Eq. 44 after learning the PSF features $S_{p,q}$. The recovery at a new FOV position (u_j, v_j) is simply given by

$$\bar{H}^{\text{PSFEx}}\left(\bar{u},\bar{v}|u_j,v_j\right) = \sum_{\substack{p,q\geq 0\\ p+q\leq d}} u_j^p\, v_j^q\, S_{p,q}\left(\bar{u},\bar{v}\right) + S_0\left(\bar{u},\bar{v}\right), \tag{46}$$

where $\bar{H}^{\text{PSFEx}}$ is the model's PSF reconstruction, and S_0 and $S_{p,q}$ are learnt during the training procedure.

5.2.4 Resolved component analysis

The resolved component analysis (RCA)[12] (Ngolè et al., 2016) is a state-of-the-art data-driven method designed for the space-based *Euclid* mission (Schmitz et al., 2020). The model builds an independent model for each CCD, can handle super-resolution, and, like the PSFEx, is based on a matrix factorisation scheme. However, there are three fundamental changes with respect to the PSFEx. The first difference is that, in RCA, the feature weights are defined as a further matrix factorisation and are also learnt from the data. The feature weights are constrained to be part of a dictionary[13] built with different spatial variations based on the harmonics of a fully connected undirected weighted graph. The graph is built using the star positions as the nodes and a function of the inverse distance between the stars to define the edge weights. The rationale of the graph-harmonics dictionary is to capture localised spatial variations of the PSF that occur in space-based missions exploiting the irregular structure of the star positions. The RCA reconstruction of an observed star is then

$$\begin{aligned}
\bar{I}^{\text{RCA}}_{\text{star}}\left(\bar{u},\bar{v}|u_i,v_i\right) &= F^{\text{RCA}}\left\{\bar{H}^{\text{RCA}}\left(\bar{u},\bar{v}|u_i,v_i\right)\right\}, \text{ where}\\
\bar{H}^{\text{RCA}}\left(\bar{u},\bar{v}|u_i,v_i\right) &= \sum_{r=1}^{N_{\text{comp}}} S_r\left(\bar{u},\bar{v}\right) A\left[r,i\right]\\
&= \sum_{r=1}^{N_{\text{comp}}} S_r\left(\bar{u},\bar{v}\right) \left(\alpha V^{\top}\right)\left[r,i\right],
\end{aligned} \tag{47}$$

where $\bar{I}^{\text{RCA}}_{\text{star},(\cdot|u_i,v_i)} \in \mathbb{R}^{p\times p}$, F^{RCA} corresponds to the degradation model of the RCA, which includes downsampling, intra-pixel shifts, among others, $S_r \in \mathbb{R}^{Dp\times Dp}$ corresponds to the data-driven feature, i.e., eigenPSF, r from a total of N_{comp} features, D is the upsampling factor in case a super-resolution step is required, and $A\left[r,i\right]$ is the rth feature weight of the ith star. The feature weight matrix A is decomposed into a sparse matrix α and a dictionary matrix $V^{\top}$ of graph-based spatial variations; for more information, see Ngolè et al. (2016).

To regularise the inverse problem, the RCA enforces a low-rank solution by fixing N_{comp}, a positivity constraint on the modelled PSF, a denoising strategy based on a sparsity constraint in the starlet (Starck et al., 2015) domain, which is a wavelet representation basis, and a constraint to learn the useful spatial variations from the graph-harmonics-based dictionary. The optimisation problem that the RCA method targets is

$$\begin{aligned}
\min_{S_k,\alpha_k} \Bigg\{ &\frac{1}{2}\sum_{i=1}^{n_{\text{obs}}} \left\| \bar{I}_{(\cdot|u_i,v_i)} - F^{\text{RCA}}\left\{\bar{H}^{\text{RCA}}_{(\cdot|u_i,v_i)}\right\}\right\|_F^2\\
&+ \sum_{r=1}^{N_{\text{comp}}} \|w_r \odot \Phi S_r\|_1 + \iota_+\left(\bar{H}^{\text{RCA}}_{(\cdot|u_i,v_i)}\right) + \iota_\Omega\left(\alpha\right)\Bigg\} \quad \text{s.t.}\\
\bar{H}^{\text{RCA}}_i &= \sum_{r=1}^{N_{\text{comp}}} S_r\left(\alpha V^{\top}\right)\left[r,i\right],
\end{aligned} \tag{48}$$

where w_r is the weight, Φ represents a transformation allowing the eigenPSFs to have a sparse representation, e.g., a wavelet transformation, $\odot$ denotes the Hadamard product, ι_+ is the indicator function of the positive orthant, and ι_Ω is the indicator function over a set Ω, which is defined as a set of sparse vectors and is used to enforce sparsity on α.

The second difference with respect to PSFEx corresponds to the regularisations used in the objective function from Eq. 48, which ends up being non-convex due to matrix factorisation and non-smooth due to the $\|\cdot\|_1$ constraint. The optimisation is solved through a block coordinate descent, as it is a multi-convex problem, and proximal optimisation algorithms that tackle the non-smooth subproblems (Beck and Teboulle, 2009; Condat, 2013).

The third difference is handling the PSF recovery at a new position (u_j, v_j). The recovery is carried out by a radial basis function (RBF) interpolation of the learned columns of the A matrix, issuing a vector, $\hat{\mathbf{a}}_j \in \mathbb{R}^{N_{\text{comp}}}$; for more details, see Schmitz et al. (2020). This way, the spatial constraints encoded in the A matrix are preserved when estimating the PSF at galaxy positions. The interpolated feature weights $\hat{\mathbf{a}}_j$ can be introduced in the Eq. 47 formulation to generate the PSF at the new j position.

The RCA model has yet to be used to generate a WL shape catalogue from real observations. Liaudat et al. (2021a) showed that the RCA is not robust enough to handle real ground-based observations from CFIS as some CCDs exhibited significant errors in the PSF shape.

5.2.5 Multi-CCD PSF model

The multi-CCD (MCCD)[14] (Liaudat et al., 2021a) model is a state-of-the-art data-driven method originally designed for the ground-based CFIS from the CFHT. The MCCD can model the full focal plane at once by incorporating the CCD mosaic geometry into the PSF model. The MCCD model rationale is explained by the limitations of other PSF models that build independent PSF models for every CCD, e.g., RCA and PSFEx: 1) fundamentally limited in the possible model complexity due to the lack of constraining power of a reduced number of star observations, and 2) the difficulty of modelling PSF spatial variations spanning the entire focal plane, i.e., several CCDs, from independently modelled CCDs. The MCCD overcomes these issues by building a PSF model containing two types of variations: global and CCD-specific. Both variations are modelled by a matrix factorisation approach, building over the success of PSFEx and RCA. The global features are shared between all CCDs, and the local CCD-specific features aim to provide

12 github.com/CosmoStat/rca

13 In the signal processing community, a dictionary is a collection of templates, or basic elements, used to decompose a signal.

14 github.com/CosmoStat/mccd

corrections for the global features. The feature weights are defined as a combination of the polynomial variations from PSFEx and the graph-based variations from RCA. The model's optimisation is more challenging than the previous models and is based on a novel optimisation procedure based on iterative schemes involving proximal algorithms (Parikh and Boyd, 2014).

The MCCD model has proven robust enough to handle real observations from CFIS (Liaudat et al., 2021a), giving state-of-the-art results. It has been incorporated into the recent ShapePipe shape measurement pipeline (Farrens, S. et al., 2022) originally designed to process the CFIS survey and generate a WL shape catalogue. The first version of the shape catalogue (Guinot et al., 2022), spanning 1,700 deg^2, from ShapePipe used PSFEx. However, the next version, spanning ~3500 deg^2, uses the MCCD PSF model and will be released soon.

5.2.6 *lens*fit

The *lens*fit (Miller et al., 2007; Kitching et al., 2008; Miller et al., 2013) refers to a Bayesian galaxy shape measurement method for WL surveys. It also includes a data-driven PSF model that will also be referenced as *lens*fit and is sparsely described throughout the different publications involving the shape measurement *lens*fit (Miller et al., 2013; Kuijken et al., 2015; Giblin et al., 2021). This method has been used with real data to produce the WL shape catalogues of CFHTLenS (Erben et al., 2013; Miller et al., 2013), KiDS+VIKING-450 (Wright et al., 2019), KiDS-450 (Hildebrandt et al., 2016; Fenech Conti et al., 2017), KiDS-1000 (Giblin et al., 2021), and VOICE (Fu et al., 2018). However, the code is not publicly available.

This PSF model differs from the previous ones. The PSFEx and RCA learn some features or *eigenPSFs* that all the PSFs share. The *lens*fit model is fitted on a pixel-by-pixel basis. Each pixel is represented as a polynomial model of degree d of the FOV positions. The *lens*fit model can use all the observations in one exposure, i.e., it uses several CCDs at once. The model uses the low-order polynomials, up to degree $n_c < d$, to be fitted independently for each CCD, and the rest of the monomials are fitted using the observations from all the CCDs. This multi-CCD modelling is a significant change with respect to previous methods that built independent models for each CCD. The total number of coefficients *per pixel* is given by

$$N_{\text{coeff}} = \frac{1}{2}\left((d+1)(d+2) + (N_{\text{CCD}}-1)(n_c+1)(n_c+2)\right), \quad (49)$$

where N_{CCD} is the total number of CCDs in the camera, d represents the degree of the polynomial varying in the full FOV, and n_c is the polynomial that is CCD-dependent. We can write the description of the pixel $(\bar{u}, \bar{v})$ of the PSF model for a FOV position (u_j, v_j) in CCD k as follows:

$$\bar{H}^{\text{lensfit}}\left(\bar{u}, \bar{v} | u_j, v_j\right) = \sum_{\substack{p,q\geq 0 \\ p+q\leq n_c}} u_j^p \, v_j^q \, a^k_{(p,q),(\bar{u},\bar{v})} + \sum_{\substack{p,q>n_c \\ p+q\leq d}} u_j^p \, v_j^q \, b_{(p,q),(\bar{u},\bar{v})}, \quad (50)$$

where $a^k_{(p,q),(\bar{u},\bar{v})}$ is the coefficient specific of CCD k, pixel $(\bar{u}, \bar{v})$, and polynomial (p, q) to be fitted to the observations. The coefficient $b_{(p,q),(\bar{u},\bar{v})}$ is shared by all the CCDs.

One thing to notice in this approach is that as the modelling of the PSF is done pixel-by-pixel, then every observation should share the same pixel grid of the PSF. There is no guarantee that an observation will have its centroid aligned with the chosen pixel grid. Therefore, the PSF model has to be aligned with the observations. Other methods, like PSFEx and RCA, interpolate the model to the observation's centroids. However, *lens*fit interpolates all the observations to the model's pixel grid with a sinc function interpolation which implies interpolating noisy images. This procedure is described in Kuijken et al. (2015).

For the KiDS DR2 (Kuijken et al., 2015), the hyperparameters used by *lens*fit are $n_c = 1$, $d = 3$, and $N_{\text{CCD}} = 32$ (from the CFHT's OmegaCAM instrument), where the images used belong to a 32×32 pixel grid. When fitting the model's parameters, each star is given a weight that is a function of its SNR with the following empirical formula

$$w_i = \frac{s_i^2}{s_i^2 + 50^2}, \quad (51)$$

where s_i is the measured SNR of the star i.

5.2.7 PSFs in full field-of-view

The PSFs in the full field-of-view (PIFF)[15] (Jarvis et al., 2020) is a recently developed PSF model that was used for the DES year 3 WL shape catalogue (Gatti et al., 2021) replacing the PSFEx that was used for the DES year 1 release. The PIFF model targets the LSST survey. Some improvements of the PIFF with respect to the PSFEx are the ability to use the full focal plane to build the model and modelling the PSF in sky coordinates rather than pixel coordinates. The PIFF offers a modular and user-friendly design that will enable further improvements. The change of modelling coordinates was motivated by the strong tree-ring detector effect observed in the DES instrument, Dark Energy Camera, which introduces astrometric distortions that are easier to correct in sky coordinates. Pixel coordinates refer to the coordinates defined on the pixel grid of the instrument. By contrast, sky coordinates refer to the angles in the celestial sphere, known as right ascension (RA) and declination (DEC). The geometric transformations that allow going back and forth between the pixel and sky coordinates are known as the World Coordinate System (WCS).

Being a modular PSF modelling code, the PIFF allows choosing between different options for the PSF model and the interpolation method. For example, the model can be an analytical profile like a Gaussian, Moffat, or Kolmogorov profile, or a more general non-parametric profile called PixelGrid. The interpolation method can be a simple polynomial interpolation, a K-nearest neighbours method, a Gaussian process (also known as Kriging), or a *basis*-function polynomial interpolation. Let us clarify the difference between the first and last mentioned interpolations. The simple polynomial interpolation first fits the PSF model's parameters $\mathbf{p}$ for each observed star. Then, it fits the coefficients of a polynomial of the 2D star positions that will later be used to interpolate. In the *basis-function polynomial interpolation*, the position polynomial's interpolation coefficients are fitted simultaneously with the model's parameters using all the available stars (from a single CCD or the entire exposure). If this last option is used with the PixelGrid model, it comes closer to the approaches of PSFEx and RCA without the

15 github.com/rmjarvis/Piff

specific characteristics of each model. We have only mentioned position polynomials, but, as in the PSFEx, the interpolation polynomial can be built on any parameter of the PSF, e.g., a colour parameter.

The current PIFF PSF model includes an outlier check after the algorithm has converged. The outlier check is based on the chi-squared, χ^2, pixel residual error between the model and observations. The model implements an iterative refining approach which means that after the model has converged, one (or more) outlier star(s) is (are) removed from the observations. A new iteration then starts with the model being recomputed. Although this approach effectively removes outlier stars not representative of the PSF (because they are binary stars or have some contamination), it can be very computationally demanding. The computing time increases linearly with the number of iterations used, which might be problematic depending on the total area to analyse or the available computing resources. For more details, we refer the readers to Jarvis et al. (2020).

The DES year 3 shape catalogue (Gatti et al., 2021) uses the PIFF model. The model is a PixelGrid with the *basis*-function polynomial interpolation using a third-order polynomial. Even if the PIFF has the *potential* to build a model across several CCD chips, in practice, each model is built independently for each CCD.

5.2.8 WaveDiff and differentiable optics approaches

The WaveDiff[16] (Liaudat et al., 2023, 2021b) PSF model was recently developed targeting space telescopes, in particular the *Euclid* mission (Laureijs et al., 2011). The WaveDiff proposes a paradigm shift for *data-driven* PSF modelling. Instead of building a data-driven model of the PSF in the pixel space as the previous models, the WaveDiff builds its model in the wavefront error (WFE) space. This change relies on a differentiable optical forward model that allows propagating the wavefront from the pupil plane to the focal plane and then computing the pixel PSF. The model is based on two components: a parametric WFE and a data-driven WFE. The parametric WFE can be based on optical simulations, characterisations of the optical system, or complementary measurements such as phase diversity observations. The parametric part should aim to model very complex dependencies that cannot be inferred from the degraded star observations. Liaudat et al. (2023) propose a parametric model built using fixed features, namely, the Zernike polynomials (Noll, 1976). For several reasons, the obtained parametric WFE model often cannot accurately represent the observed PSF, e.g., the telescope changes over time, there are errors in the built parametric model, and relevant effects are not considered or neglected. Consequently, the data-driven WFE should be capable of correcting the aforementioned mismatches. This data-driven part is based on a matrix factorisation with spatial variations inspired by PSFEx and RCA. It is crucial to model smooth variations that have a reliable generalisation of the PSF to different positions in the FOV. An overview of the model is presented in Figure 13.

Estimating the model's parameters in the WFE space is a challenging, ill-posed inverse problem known as phase retrieval (Fienup, 1993; Fienup et al., 1993; Shechtman et al., 2015), i.e., estimating a complex signal from intensity-only observations. The optimisation problem is non-convex and non-smooth, the star observations are not very informative, and there is no guarantee that the WFE model's structure can represent the underlying ground truth WFE. Liaudat et al. (2023) show that under the aforementioned conditions, targeting the estimation of the ground truth WFE is not the best option. The PSF model's objective is to have a good pixel representation of the PSF. It is, therefore, possible to estimate a WFE manifold far away from the underlying WFE but very close in the pixel space. The data-driven features, the basis of the WFE manifold, are estimated with a stochastic gradient descent method widely used for estimating neural network parameters.

The WaveDiff model can handle spatial variations, super-resolve the PSF, and model chromatic variations thanks to the WFE formulation and the optical forward model, which also considers more general degradations as in Eq. 39. To the best of our knowledge and at the time of writing, this is the *only* data-driven PSF model that can cope with the spectral variations of the PSF. The WaveDiff shows a breakthrough in performance for data-driven PSF models in a simplified *Euclid*-like setting (Liaudat et al., 2023). It is flexible enough to be adapted to different space telescopes. The framework proposed shows an exciting research direction for future data-driven PSF modelling. The WaveDiff model has yet to be tested with real space-telescope observations and has to incorporate detector-level effects, which are more naturally modelled in pixel space. For more details, we refer the readers to Liaudat et al. (2023) and Liaudat (2022).

More general approaches based on differentiable optics have been recently emerging, e.g., ∂Lux[17] (Desdoigts et al., 2023). Approaches based on automatic differentiation can be useful not only for modeling the PSF but also for designing apodizing phase plate coronagraphs (Wong et al., 2021) and detector calibrations (Desdoigts et al., 2023).

5.2.9 Other PSF models

- *Shapelets*: Refregier (2003) proposed a framework to analyse images based on a series of localised basis functions of different shapes named *shapelets* and images could then be decomposed using these basis functions. Refregier and Bacon (2003) continued the work to propose the *shapelet* framework to be used for building shear estimates and modelling the PSF. The PSF modelling consists of decomposing the star images in the *shapelet* basis and then performing an interpolation of the coefficients to positions of interest. Essentially, it is an extension of the approach seen in shape interpolation. Expressing the image in *shapelet* coefficients allows denoising the star images and providing an easier framework for the galaxy-PSF deconvolution. However, capturing all the PSF structures in a finite expansion over analytical functions is not always possible, leading to the loss of information. Massey and Refregier (2005) extended the framework from Cartesian to polar *shapelets*.
- *Moffatlets and Gaussianlets*: Li et al. (2016) proposed two other series of basis functions to decompose the PSF named *Moffatlets*

16 github.com/CosmoStat/wf-psf

17 github.com/LouisDesdoigts/dLux

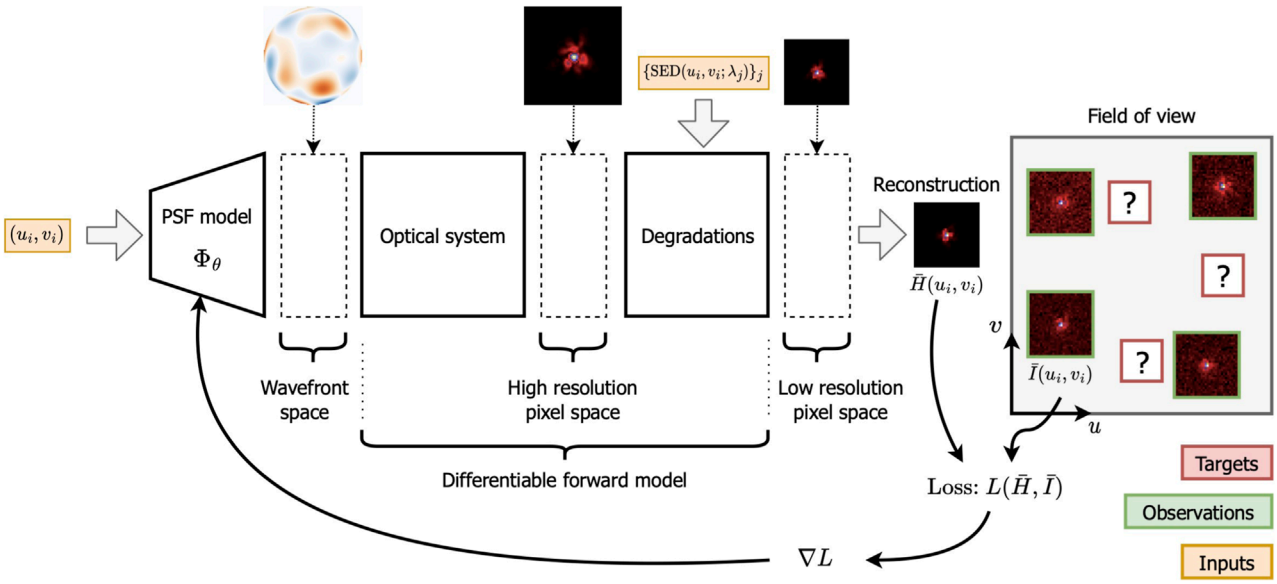


FIGURE 13
Overview of the WaveDiff approach to model the PSF with a differentiable optical forward model. Figure adapted from Liaudat et al. (2023).

and *Gaussianlets*. They compared the PSF reconstruction using the aforementioned basis with a PCA-based method on LSST-like images. Using analytical basis functions leads to more denoised models, as expected. Nie et al. (2021a) continued the approach and forced the principal components being learnt in the PCA-like algorithm to be built using the *Moffatlets* basis. This choice avoids the principal components of learning noise as the *Moffatlets* basis avoids it. Both analyses lack a performance benchmark with a reference PSF model like PSFEx. In addition, the models' performance is computed at the same position as the observed stars, so the models' generalisation to other positions, a fundamental task of the PSF model, still needs to be studied.

- *Fourier-based methods*: Zhang (2007) proposed a Fourier-based method for directly measuring the cosmic shear, taking into account the PSF, which was further developed in several publications (Zhang, 2011; Zhang et al., 2015; Lu et al., 2017; Zhang et al., 2019). The method is based on the quadrupole moments of the galaxy images (described in detail in Section 7.2) but is measured in Fourier space. The handling of the PSF is also done in Fourier space. Lu et al. (2017) explore different interpolation approaches for the PSF in the aforementioned Fourier framework. The 2D power spectrum of the observed PSFs is interpolated to target positions. The interpolation is done pixel-by-pixel, and the best-performing method is a well-parametrised polynomial interpolation. An advantage of the Fourier interpolation is that the 2D power spectrum is automatically centred in Fourier space, simplifying the handling of images with intra-pixel shifts or, what is the same, different centroids. Another Fourier-based shear measurement method is Bayesian Fourier Domain (BFD) (Bernstein and Armstrong, 2014; Bernstein et al., 2016) built on the Bayesian formalism. However, it does not include a specific PSF model.
- *Optimal transport (OT)-based methods*: There exist two approaches involving OT (Peyré and Cuturi, 2019) to tackle the PSF modelling problem. Ngolè and Starck (2017) proposed to use Wasserstein barycenters as a nonlinear geometry-aware interpolation procedure of a low-dimensional embedding representation of the PSFs. Although elegant, the performance of the approach does not seem to justify its computational burden. In the comparison method, an RBF interpolation of the principal components obtained through the PCA achieves a similar performance. The performance of the PCA method is better in terms of ellipticity but slightly worse in terms of pixel error and PSF FWHM. Schmitz (2019) worked on a data-driven PSF model based on the RCA that could model the chromatic variations of the PSF through the use of Wasserstein barycenters that were previously developed by Schmitz et al. (2018). The OT-based PSF model named λRCA was compared to the RCA. The comparison showed a lower pixel and size error for λRCA, although the ellipticity error was similar or better for RCA. This method assumes that the PSF's chromatic variation is smooth over all the passband. This assumption holds if the only chromatic effect of the PSF is due to the diffraction phenomena, which exhibits a smooth variation with the $1/\lambda$ dependence in the WFE that has already been presented in Eq. 35. However, if this is not the case and another non-smooth chromatic variation is present, currently occurring in *Euclid* (Venancio et al., 2016; Baron et al., 2022), there is no straightforward way to adapt the λRCA model to account for it.
- *Wavefront approach*: Soulez et al. (2016) proposed to model the propagation of light through the mirrors of the optical system. The PSF modelling problem is recast into the phase-retrieval problem. The article is a proof-of-concept as there are only qualitative results, and many PSF modelling difficulties remain unaddressed.

- *Exploit out-of-focus images*: Some instruments, like the Dark Energy Camera (DECam) (Flaugher et al., 2015), are equipped with wavefront sensors that are helpful for focus, alignment, and the adaptive optics system (AOS) (Roddier and Roddier, 1993; Roodman, 2010; Roodman, 2012). The LSST camera (LSST Science Collaboration et al., 2009) is also equipped with wavefront sensors (Manuel et al., 2010; Claver et al., 2012; Xin et al., 2015; Xin et al., 2016). Roodman et al. (2014) proposed to use the data from the wavefront sensors to constrain the optical contribution of the PSF. The work was continued by Davis et al. (2016) who proposed a wavefront-based PSF model for the DECam instrument using out-of-focus observations. The PSF model was based on Zernike polynomials fitted to out-of-focus stars, also called doughnuts, that contain considerably more wavefront information than in-focus stars. Then, a new fit is done for each exposure based on the measured quadrupole moments of the in-focus star images. It is not easy to understand at which point the quadrupole moments constrain the proposed PSF model and at which point it is the base physical wavefront measured from the out-of-focus images that are the only parts driving the performance of the model. Snyder et al. (2016) proposed using the AOS measurements to characterise atmospheric turbulence in terms of Zernike decomposition.
- *Deep learning approaches*: A model coined PSF-NET was proposed by Jia et al. (2020a) and is based on two convolutional neural networks (CNNs) trained jointly. One network transforms high-resolution images into low-resolution images, while the other does the opposite. The CNNs are trained in a supervised way expecting that the first network will learn a PSF manifold. However, it is unclear how the approach handles the spatial variation of the PSF, and it has not been tested for WL purposes. Jia et al. (2020b) proposed another approach for PSF modelling based on denoising auto-encoders, but the spatial variation of the PSF remains untackled. Another approach is followed by Herbel et al. (2018), where the PSF profile is modelled by a parametric function consisting of a base profile of two Moffat profiles and several parametrised distortions to increase the expressivity. A CNN is trained in a supervised manner to predict the parameters of the parametric profile from a noisy star observation. The neural network provides a good estimation of the parameters, but the spatial variation of the PSF is, again, not addressed. Having a PSF model that can model the observations is important. However, in PSF modelling for WL analysis, a crucial part is to capture the spatial variations of the PSF and that the model outputs the PSF at different positions in the FOV.

6 Desirable properties of PSF models

In the previous section, we reviewed some of the most relevant PSF models developed so far. After studying many models, we can conclude on desirable PSF model properties. The PSF model should

(a) *Have an accurate modelling of the PSF light profile.* This modelling is essential for any target task, as the light profile is the convolutional kernel for a given position. The smoothness and structure in the PSF profile are a consequence of the PSF being the Fourier transform, in Fraunhofer's approximation, of a particular finite-length aperture that limits the frequency content of the PSF. This frequency limitation prevents us from having a Dirac distribution as a PSF, as it would require an infinite frequency content to build it. One difficulty is in accurately modelling the PSF's wings or outer region, which is often below the noise level. In ground-based telescopes, the effect of the atmosphere can be interpreted as a low-pass filter for the PSF, smoothing the PSF light profile.

(b) *Produce noiseless estimations of the PSF.* The presence of noise in the PSF estimations is an issue for purely data-driven models, which sometimes tend to overfit noisy observations. Some regularisations have to be introduced in the PSF model parameter optimisation to avoid producing noisy PSFs. A seemingly straightforward solution to this problem is to rely on PSF models based on fixed basis functions like *shapelets* or *Moffatlets*. These models will always output denoise PSFs as their basis functions cannot reproduce the noise. However, they introduce modelling errors if they cannot accurately model the observed PSF light profiles.

(c) *Capture the PSF field's variations.* It is often the case that a good quality PSF is required at the position of a certain object where no direct information about the PSF is available. The PSF model first has to capture most of the relevant information from the observations at other positions and wavelengths. Then, the model exploits this information and predicts the PSF at the required position and wavelength. The PSF model relies upon its generalisation power as it has to exploit the PSF field information from other positions.

(d) *Be capable of exploiting the structure of the PSF field variations.* This desired property is related to the previous point (c). An exciting approach to obtain good generalisation power is to learn the structure of the PSF field variations. This structure is a consequence of the physical properties of the telescope's optical system. The following subsection provides a physical understanding of the PSF field structure, which imposes a certain smoothness to the variations. The spatial variations are also structured due to the atmosphere if we study the PSF field of a ground-based telescope. A data-driven PSF model should use a low-complexity representation of the PSF field, which can learn its structure and spatial variations.

(e) *Handle discontinuities of the PSF field.* The CCD misalignments are a source of discontinuity of the PSF's spatial variations. The PSF field is piece-wise continuous, and the borders delimiting the discontinuity are well known as the geometry of the focal plane is accurately known. A straightforward way to handle the discontinuities is to model the PSF for each CCD independently, e.g., PSFEx. Although this is simple to implement, it limits the number of stars available to constrain the PSF model. Another more difficult but potentially more powerful approach is to build a PSF model for the entire focal plane, accounting for the focal plane discontinuities, e.g., MCCD. Another source of discontinuity is segmented mirrors, e.g., JWST's hexagonal mirrors seen in Figure 9B.

(f) *Be robust to outliers and contaminations of the star sample.* Contaminations can come from the selection of stars, and the fact that the objects classified as stars are good representations

of the PSF and are not small galaxies or binary stars (Kuntzer and Courbin, 2017). Outliers can come from imperfect image preprocessing where detector effects, like CTI, have not been adequately removed. In addition, the model should be robust to different observation conditions such as spatial distributions of the observed stars, SNRs, and the number of observed stars.

(g) *Work appropriately with the target task.* The PSF model can be exploited differently according to the target task's objective, e.g., estimating a deconvolved object or estimating some summary statistic of the deconvolved image. The model should be developed with the task in mind, as each task might be more or less susceptible to different kinds of errors.

(h) *Be fast.* Upcoming surveys will be processing a vast amount of observations. Consequently, they put significant pressure on the computing time of PSF models as they have to cope with the data intake. The requirements in terms of computing time can drive many design choices in a PSF model, preventing the use of costly physical simulations.

Once the PSF model has been developed with all the aforementioned properties in mind, it is essential to validate the model's performance. The validation should ensure that the expected performance of the model is achieved and helps to identify sources of problems and provide directions for the improvement of the model. In the next section, we give an overview of validation methods for PSF models.

7 Validation of PSF models

The validation of PSF models is a challenging problem. To derive a validation method, it would be necessary to quantify the impact of PSF modelling errors on the final objective of our analysis. We consider, as an example, a weak-lensing-based cosmological analysis, where the objective is to derive constraints on the parameter of the cosmological model under analysis. This exercise is challenging, given the analysis' complexity and the large data volume. Nevertheless, with some simplifying assumptions, the PSF modelling requirements were set for the *Euclid* mission as shown in Section 7.3. In this analysis, some assumptions on the PSF shape used do not always hold for the high complexity of the PSF in a space-based mission like *Euclid*. Even though it is essential to derive requirements for the PSF model, these do not give much information on the nature of errors and possible problems that the PSF model has. Therefore, it is necessary to derive different diagnoses or null tests. Jarvis et al. (2016) proposed a set of null tests for the DES WL shear catalogues Science Verification, which includes the PSF model validation.

The most basic rule for any validation of the PSF models is to separate the observations in the FOV into two data sets for estimation and testing, i.e., validation, which could be 80{%} and 20{%}, respectively. The first one should be used to estimate the PSF model. The second one should help validate its performance and not be used in the PSF model estimation. This rule tests the PSF model's generalisation power to unseen positions in the FOV. Next, we will describe the most used PSF diagnosis that will help us validate the performance of the PSF models.

7.1 Pixel-based metrics

The most straightforward diagnostic that we can think of is to compute the pixel residual of our PSF model. Once trained, the model is used to recover the PSF at test positions. We can then compute the RMSE of the pixel reconstruction residuals. The PSF model can ideally predict the observed test stars without error, and the reconstruction residual would only contain the observational noise. If we work with simulations, we can produce noiseless stars for our testing set, and the RMSE will directly indicate the pixel reconstruction error. Even though pixel-based metrics can give insight into the PSF model performance, they are not easily interpretable regarding scientific impact. Errors in the PSF core or the PSF wings can impact the observed galaxy's estimated shape differently. With the existing methodology, it is difficult to translate a pixel-based metric into a scientifically meaningful quantity in terms of error propagation.

When working with real data, the PSF model's validation with pixel-based metrics becomes more complicated. The different noise levels in the data can hide the pixel reconstruction error, making it difficult to compare different PSF models or even assess the performance of a single one. Liaudat et al. (2021a) proposed pixel-based reconstruction metrics for real observations. Let us denote with $I_{\text{star}}(\bar{u},\bar{v}|u_i,v_i), I_{\text{model}}(\bar{u},\bar{v}|u_i,v_i) \in \mathbb{R}^{p\times p}$ a test star and the predicted PSFs, respectively, at the FOV position (u_i,v_i), where p^2 is the total number of pixels in the image. To simplify the notation, we write $I(\bar{u},\bar{v}|u_i,v_i) = I_i$. In star images, most of the PSF flux is concentrated in the centre of the square image, and the noise level can be considered constant in the image. Therefore, we can mask the image to only consider the central pixels within a given radius of R pixels and compute the pixel RMSE of the masked images. We note $\tilde{I} = I \odot M_R$ the masked image, where $M_R \in \{0,1\}^{p\times p}$ is a binary mask and $\odot$ is the Hadamard or element-wise product. Let us define the σ value as

$$\sigma(\tilde{I}_i) = \sigma(M_R \odot I_i) = \left(\frac{1}{\tilde{p}^2}\sum_{\bar{u},\bar{v}=1}^{p}\left(\tilde{I}_i(\bar{u},\bar{v})\right)^2\right)^{1/2}, \tag{52}$$

where $\tilde{p}^2$ is the number of unmasked pixels, and the sum is done on the unmasked pixel values. The first pixel metric is Q_{p_1} and is defined as

$$Q_{p_1} = \left(\text{Err}^2 - \sigma^2_{\text{noise}}\right)^{1/2}, \tag{53}$$

where

$$\text{Err}^2 = \frac{1}{N_s}\sum_{i=1}^{N_s}\sigma\left(\tilde{I}_{\text{star},i} - \tilde{I}_{\text{model},i}\right)^2, \text{ and } \sigma^2_{\text{noise}} = \frac{1}{N_s}\sum_{i=1}^{N_s}\sigma\left(\tilde{I}^{*}_{\text{star},i}\right)^2, \tag{54}$$

where the general noise standard deviation, σ_{noise}, is computed from the pixels on the outer region of the test stars, i.e., $\tilde{I}^{*}$, which we define as $\tilde{I}^{*} = I_{\text{star}} \odot M^{*}_R$, where M^{*}_R is such that $M^{*}_R + M_R = 1_{p\times p}$. Subtracting our estimated model, I_{model}, from an observed star, I_{star}, should lead to a residual map containing only noise if the model is perfect. The probability of having our model correlated with the noise is minimal. Therefore, the method with the smallest Q_{p_1} can be considered the best from the Q_{p_1} point of view.

The next two metrics, Q_{p_2} and Q_{p_3}, help quantify the model noise. Let us define $\sigma^2_{\text{model},i} = [\sigma(\tilde{I}_{\text{star},i} - \tilde{I}_{\text{model},i})^2 - \sigma(\tilde{I}^{*}_{\text{star},i})^2]_+$, where

the operator $[\cdot]_+$ sets to zero negative values. Then, both metrics are defined as follows:

$$Q_{p_2} = \left(\frac{1}{N_s} \sum_{i=1}^{N_s} \sigma^2_{\text{model},i} \right)^{1/2}, \quad Q_{p_3} = \left(\frac{1}{N_s} \frac{1}{N_s} \sum_{i=1}^{N_s} \left(\sigma^2_{\text{model},i} - Q^2_{p_2} \right)^2 \right)^{1/4}. \tag{55}$$

The Q_{p_2} metric represents the modelling error expectation for a given star, and the Q_{p_3} metric indicates the fluctuation of the modelling error. A perfect PSF model would give values close to zero for the three metrics. We have assumed that there is no background contamination in the observed test stars or that it has been removed.

7.1.1 Chromatic PSF models

Some applications or analyses require a chromatic PSF model, and it is essential to validate the chromaticity of the PSF model. This monochromatic validation means validating the PSF at every single wavelength or validating the monochromatic PSF before it is integrated into the instrument's passband. A PSF model with a good performance in reproducing the polychromatic stars does not necessarily have a good monochromatic performance. Supposing that is the case and even if the spectra of the different objects are known in advance, the PSF errors will be more significant when used with objects with considerably different spectra, e.g., galaxies. Chromatic PSF models will generally be required if the observing instrument has a wide passband, e.g., the *Euclid*'s VIS instrument passband goes from 550 nm to 900 nm. The pixel RMSE can be computed for monochromatic PSFs in the passband as done in Liaudat et al. (2023). However, it is cumbersome to validate with real data as we usually do not have access to the monochromatic PSFs of the instrument under study. Consequently, the monochromatic validation might only be possible with simulations.

7.2 Moment-based metrics

Weak gravitational lensing analyses are interested in measuring the shape of galaxies as the measured ellipticity is an estimator of the shear. Cosmologists have developed formulations to relate the PSF errors, expressed in terms of shape and size metrics (Massey et al., 2012), to the cosmological parameters of interest (Cropper et al., 2013). Therefore, it seems natural to have diagnosis metrics based on the ellipticity and size of the PSF. These metrics are determined using the moments of the polychromatic observation $\bar{I}[u, v]$. Following Hirata and Seljak (2003), we redefine the image moments that we will use in practice, which include a weight function as follows:

$$< \mu > = \frac{\int \mu \, \bar{I}(\bar{u}, \bar{v}) \; w(\bar{u}, \bar{v}) \; \mathrm{d}\bar{u} \, \mathrm{d}\bar{v}}{\int \bar{I}(\bar{u}, \bar{v}) \; w(\bar{u}, \bar{v}) \; \mathrm{d}\bar{u} \, \mathrm{d}\bar{v}}, \tag{56}$$

$$M_{\mu\nu} = \frac{\int \bar{I}(\bar{u}, \bar{v}) \; (\mu - < \mu >) \; (\nu - < \nu >) \; w(\bar{u}, \bar{v}) \; \mathrm{d}\bar{u} \, \mathrm{d}\bar{v}}{\int \bar{I}(\bar{u}, \bar{v}) \, w(\bar{u}, \bar{v}) \; \mathrm{d}\bar{u} \, \mathrm{d}\bar{v}}, \tag{57}$$

where $\mu, \nu \in \{\bar{u}, \bar{v}\}$, $<\mu>$ denotes the mean of μ, and $w(\bar{u}, \bar{v})$ is a weight function that helps in noisy settings. The weight function is also useful to compute the moments from diffraction-limited PSFs, e.g., an Airy profile, as they prevent the integral from diverging due to the wings of the PSF. Eq. 56 defines the first-order moments, while Eq. 57 defines the second-order moments. The ellipticities, or *shape* metrics, are defined as

$$e = e_1 + \mathrm{i} e_2 = \frac{(M_{\bar{u}\bar{u}} - M_{\bar{v}\bar{v}}) + \mathrm{i} \, 2 M_{\bar{u}\bar{v}}}{M_{\bar{u}\bar{u}} + M_{\bar{v}\bar{v}}}, \tag{58}$$

where i is the imaginary unit, and the size metric is defined as

$$R^2 = T = M_{\bar{u}\bar{u}} + M_{\bar{v}\bar{v}}. \tag{59}$$

One widely used method to estimate these metrics is the adaptive moment algorithm from GalSim's HSM (Hirata and Seljak, 2003; Mandelbaum et al., 2005). The adaptive moment algorithm measurement provides σ as size, which relates to the aforementioned size metric as $R^2 = 2\sigma^2$. The measurements are carried out on well-resolved polychromatic images. If the observations are undersampled, as is the case for *Euclid*, a super-resolution step is required for the model. Gillis et al. (2020) proposed alternative metrics, based on the image moments, that target the validation of space-based PSFs with emphasis on the HST PSF.

The measurements of the shape parameters based on the image moments are susceptible to image noise. If we are working with real data, we do not have access to ground truth images and are obliged to work with noisy observations. Therefore, we have to average over many objects in order to conclude from the different diagnostics. We continue by presenting different moment-based metrics.

7.2.1 Shape RMSE

We start with a set of test stars and their corresponding PSF estimations. Then, the most direct moment-based metric is to compute the RMSE of the ellipticities and size residuals between the observations and model prediction. However, this metric could be more insightful as it does not provide any information about the composition of the residuals and the estimation biases involved.

7.2.2 meanShapes

A useful diagnostic is to compute the spatial distribution of the ellipticities and size residuals, which we coin *meanShapes*. For an example, see Figure 14. This diagnostic allows us to inspect if there are spatial correlations in the shape and size residuals, indicating that the PSF model is not capturing certain spatial variations from the PSF field. In order to have a finely sampled distribution, we have to average over many exposures, as the available stars from a single exposure are not enough to observe patterns. The shape measurements are also noisy, therefore averaging over many exposures allows us to smooth out the residuals and observe systematic modelling errors. In practice, we divide the focal plane into several spatial bins, consider several exposures, and then the value of each bin is built by averaging the residuals of all the stars within that bin. A ground-based survey allows us to average the ellipticity contribution of the atmosphere (Heymans et al., 2012), as it can be considered a random field with a zero mean. Then, the observed ellipticity distribution over the focal plane is due to the telescope's optical system that is consistent in every exposure. It is also possible to plot the same spatial distribution but observe the positions of the stars. Such a plot will help observe if there are regions of stellar under-density that could eventually affect the PSF model.

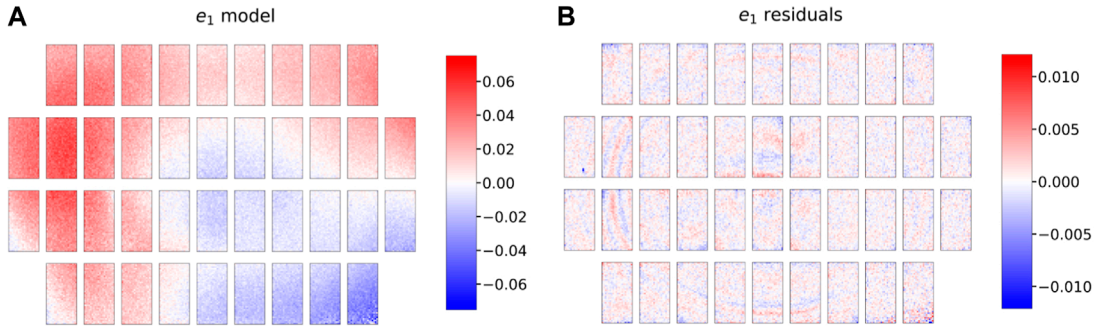


FIGURE 14
meanShapes plots showing the first component of the ellipticity of the PSF model **(A)** and of its residuals with the observed stars **(B)**. The figure shows the 40-CCD mosaic from the MegaCam instrument's focal plane at the Canada–France–Hawaii Telescope. Figure reproduced from Guinot et al. (2022).

We have assumed that the ellipticities and sizes are good descriptors, or summary statistics, of the PSF shape. If this is not the case, the diagnostic could be extended with more accurate descriptors.

7.2.3 ρ-statistics

Rowe (2010) proposed to compute the auto- and cross-correlations of the ellipticities and their residuals as a diagnostic. The diagnostic was then expanded by Jarvis et al. (2016) to a combination of ellipticities, sizes, and residuals. The ρ-statistics are helpful in identifying PSF modelling biases and detecting the scales at which it most affects the weak-lensing analysis. Following Jarvis et al. (2016), we define the ρ-statistics as follows:

$$\rho_1(\theta) = \left\langle \delta e^*_{\rm PSF}(\boldsymbol{\theta}')\ \delta e_{\rm PSF}(\boldsymbol{\theta}'+\boldsymbol{\theta}) \right\rangle, \tag{60}$$

$$\rho_2(\theta) = \left\langle e^*_{\rm PSF}(\boldsymbol{\theta}')\ \delta e_{\rm PSF}(\boldsymbol{\theta}'+\boldsymbol{\theta}) \right\rangle, \tag{61}$$

$$\rho_3(\theta) = \left\langle \left(e^*_{\rm PSF}\frac{\delta R^2_{\rm PSF}}{R^2_{\rm PSF}}\right)(\boldsymbol{\theta}') \left(e_{\rm PSF}\frac{\delta R^2_{\rm PSF}}{R^2_{\rm PSF}}\right)(\boldsymbol{\theta}'+\boldsymbol{\theta}) \right\rangle, \tag{62}$$

$$\rho_4(\theta) = \left\langle \delta e^*_{\rm PSF}(\boldsymbol{\theta}') \left(e_{\rm PSF}\frac{\delta R^2_{\rm PSF}}{R^2_{\rm PSF}}\right)(\boldsymbol{\theta}'+\boldsymbol{\theta}) \right\rangle, \tag{63}$$

$$\rho_5(\theta) = \left\langle e^*_{\rm PSF}(\boldsymbol{\theta}') \left(e_{\rm PSF}\frac{\delta R^2_{\rm PSF}}{R^2_{\rm PSF}}\right)(\boldsymbol{\theta}'+\boldsymbol{\theta}) \right\rangle, \tag{64}$$

where * denotes complex conjugation, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ denote sky positions, θ denotes the modulus of $\boldsymbol{\theta}$, and δ denotes the residual error that can be computed as $\delta e_{\rm PSF} = e_{\rm PSF} - e_{\rm star}$ in case of PSF ellipticity. Suppose that the ellipticities are random fields that are isotropic and statistically homogenous. In such a case, we can compute the correlation $\rho(\boldsymbol{\theta},\boldsymbol{\theta}')$ as $\rho(|\boldsymbol{\theta}-\boldsymbol{\theta}'|) = \rho(\theta)$, using the modulus θ. This choice means that we are assuming translational and rotational symmetry, a consequence of the cosmological principle. We define several θ bins in a logarithmic scale, corresponding to $\ln\theta - \Delta\ln\theta/2 < \ln\theta_{ij} < \ln\theta + \Delta\ln\theta/2$, where $\theta_{ij} = |\boldsymbol{\theta}_i - \boldsymbol{\theta}_j|$ is the distance between two objects at $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$. Consequently, the correlation function at θ can be computed using the following unbiased estimator of ρ:

$$\widehat{\rho}(\theta) = \frac{\sum_{i,j} w_i w_j e^{\rm A\,*}_i e^{\rm B}_j}{\sum_{i,j} w_i w_j}, \tag{65}$$

where we are computing, as an example, the correlation of two ellipticities $e^{\rm A}$ and $e^{\rm B}$, and the weights depend on the SNR of the ellipticity measurements. We carry out the weighted sum over the pairs of objects within each bin.

The ρ-statistics are interesting as they can be propagated to the shear two-point correlation function (2PCF) (Kilbinger, 2015), which allows for studying the properties of the weak-lensing convergence field. Following Jarvis et al. (2020), we include the PSF errors into the shear 2PCF, making the ρ-statistics appear, and then express the systematic error in the shear 2PCF.

7.2.4 Other shape metrics

Another shape metric that gives insights into the performance of the PSF model in a weak-lensing analysis is the PSF leakage α from Jarvis et al. (2016). It is related to the linear modelling of the shear bias and where it has been decomposed into a multiplicative bias and an additive bias further decomposed into PSF-dependent (leakage) and PSF-independent terms. The PSF leakage helps quantify how the PSF affects the shear estimation through shape measurement. It measures the leakage of the PSF shape to the galaxy shapes.

7.3 Weak lensing: PSF error propagation and PSF requirements definition

The pioneers in the PSF error propagation for WL were Paulin-Henriksson et al. (2008), followed by Massey et al. (2012). The proposed framework is based on the second-order moments of the images, i.e., complex ellipticity e and size R^2. It expresses how the PSF, or some other effect, affects the observed ellipticity and size. Let us consider the effect of the PSF on the unweighted moments from Eqs 58, 59 where unweighted represents computing Eqs 56, 57 without the w weight function. Then, we obtain

$$e_{\rm obs} = e_{\rm gal} + \frac{R^2_{\rm PSF}}{R^2_{\rm PSF}+R^2_{\rm gal}}\left(e_{\rm PSF} - e_{\rm gal}\right) \quad \text{and} \quad R^2_{\rm obs} = R^2_{\rm gal} + R^2_{\rm PSF}, \tag{66}$$

where the subscript $_{\rm obs}$ refers to the quantity measured to the observed galaxy, the subscript $_{\rm gal}$ refers to the intrinsic quantity of

the galaxy, and $_{\rm PSF}$ refers to the quantity measured to the PSF. There are intrinsic assumptions in Eq. 66, that the observational model is $I_{\rm obs} = I_{\rm gal} \star H_{\rm PSF}$, and that all the moments are well-defined. Then, Eq. 66 can be rewritten to express the quantity of interest in weak lensing, the intrinsic galaxy ellipticity, as follows:

$$e_{\rm gal} = \frac{e_{\rm obs} R^2_{\rm obs} - e_{\rm PSF} R^2_{\rm PSF}}{R^2_{\rm obs} - R^2_{\rm PSF}}. \tag{67}$$

The error propagation consists of expanding the previous equation in a first-order Taylor series with respect to the quantities of interest. In this case, it will be the shape and size of the PSF, and the propagation is given by

$$\hat{e}_{\rm gal} \approx e_{\rm gal} + \frac{\partial e_{\rm gal}}{\partial (R^2_{\rm PSF})} \delta (R^2_{\rm PSF}) + \frac{\partial e_{\rm gal}}{\partial e_{\rm PSF}} \delta e_{\rm PSF}, \tag{68}$$

where δ refers to errors in the model with respect to the ground truth. It is straightforward to compute the partial derivatives in Eq. 68 from Eq. 67. We then obtain the following expression:

$$\hat{e}_{\rm gal} \approx e_{\rm gal} \left(1 + \frac{\delta (R^2_{\rm PSF})}{R^2_{\rm gal}} \right) - \left(\frac{R^2_{\rm PSF}}{R^2_{\rm gal}} \delta e_{\rm PSF} + \frac{\delta (R^2_{\rm PSF})}{R^2_{\rm gal}} e_{\rm PSF} \right). \tag{69}$$

The previous ellipticity estimator can be used to obtain a shear estimator assuming that the intrinsic galaxy distribution has a zero mean. The estimator can then be used in the linear shear bias parametrization from Jarvis et al. (2016). At this point, we can express the additive and multiplicative biases as a function of the elements from Eq. 69. This analysis shows us that the multiplicative bias is related to the size of the PSF with its estimation error and the size of the galaxy. The result is as expected if we pay attention to the first term of Eq. 69.

This framework allows us to consider different types of errors. Massey et al. (2012) use it to include errors due to non-convolutional detector effects, imperfect shape measurement, and the fact that the shape measurement method used weighted, i.e., Eq. 57, instead of unweighted moments. The procedure consists of adding the desired effect to the galaxy ellipticity expression, Eq. 67, and then adding their corresponding partial derivatives to the Taylor expansion seen in Eq. 68. Cropper et al. (2013) use this formalism to derive requirements for a WL mission in space. The aforementioned framework was used to derive the current PSF model requirements for the *Euclid* space mission (Laureijs et al., 2011).

The previous formalism is based on *unweighted* moments. However, in practice, the moments are always computed from noisy images using a compact weight function to ensure that the measurement yields significant results. Melchior et al. (2011) and Melchior and Viola (2012) showed that using a weight function mixes the image's moments. Consequently, the second-order moments are affected by higher-order moments even in the absence of noise, thus exposing the Paulin-Henriksson et al. (2008) framework's fundamental limitations. Schmitz et al. (2020) then noted and verified empirically in an *Euclid*-like scenario that the propagation is based on second-order moments of the PSF, which do not accurately describe the shape of a space-based PSF. A perfect second-order moment estimation of the PSF would have a zero shear bias contribution in the formalism described. However, the PSF's higher moments error (HME) will impact the shear biases and are not considered in the framework proposed by Paulin-Henriksson et al. (2008). The higher the contribution of the HME to the PSF, the more significant the deviations are. A space mission like *Euclid* will have a PSF close to the diffraction limit, i.e., its shape will be complex and not well described by a Gaussian (or by its second-order moments). As a space PSF is not well described by second-order moments, the previous requirements should be used *with caution*. The LSST collaboration, concerned with the previous issue, studied the contribution to systematic biases of the HME of the PSF model on the shear measurement (Zhang et al., 2021; Zhang et al., 2022). Zhang et al. (2021) showed that the HME of the PSF model might be a significant source of systematics in upcoming WL analyses. Zhang et al. (2022) studied the impact of moments from the third to sixth order on the cosmological parameter inference, concluding that the HME of PSF models like PSFEx and PIFF should be reduced for future surveys like LSST if the WL analysis is to remain unchanged.

The use and adoption of automatic differentiable (Baydin et al., 2017) models could make a significant contribution to error propagation. The derivatives of the target estimators with respect to the model's parameters, or intermediate products with physical meaning, would be available. This fact allows us to consider more complex scenarios than the one seen in Eq. 66, as we would not require explicitly writing the equations nor their derivatives. A differentiable forward model should be enough to describe how the PSF interacts with the target task.

8 Conclusion

This review gives an overview of point spread function (PSF) modelling for astronomical telescopes, emphasising cosmological analyses based on weak gravitational lensing. This application sets the tightest constraints on the PSF models and has driven much of the last progress in PSF models. The development of new instruments and telescopes seeking higher precision and accuracy requires more powerful PSF models to keep up with the reduction of other sources of errors. We differentiate two scenarios that fundamentally change how the PSF is modelled: the ground- and space-based telescopes. The main difference is the atmosphere, how it affects the observations, and how challenging it is to build an atmospherical physical model that can be exploited in a reasonable amount of time. The difficulty of handling the temporal integration of a representative atmospherical model fosters the use of purely data-driven PSF models built in the pixel space for ground-based telescopes. The stability of space-based telescopes allows for exploiting models more physically based on the wavefront.

The optics fundamentals to properly define the PSF and understand its effect on the underlying imaged object are not often introduced in PSF modelling articles. One of our goals was to solve this issue by providing a concise yet comprehensive introduction to optical principles. The provided optical background should cover most of the available PSF models and motivate the general observational model that we have proposed. This observational model can be further simplified and adapted to different use cases, which include ground or space telescopes. We described several assumptions that might not always hold. After describing how the PSF affects our observations, we presented an extended list

of the leading optical- and detector-level contributors to the PSF field. These contributors are the sources of PSF's spatial, spectral, and temporal variations in addition to its morphology. A detailed description of the atmospheric contribution based on phase screens was presented. We then gave a brief description of the most relevant PSF models.

The PSFEx model has been successful in modelling the PSF for several surveys with its robust and fast implementation. However, the next-generation telescopes set higher PSF requirements, demanding novel models to achieve such performances. Recent models target upcoming telescopes, e.g., the Vera C. Rubin observatory and the *Euclid* telescope and *Roman* Space Telescope. These models are continuously being developed, and they push forward the capabilities of PSF modelling. A common bottleneck for these is the computing time required to estimate the model from observations. It is unclear if the solution can be achieved through better software implementations that exploit parallel computing architectures or better-performing programming languages. A refactoring of the methods allowing for simplifications that accelerate calculations might be required, or even both approaches. One big challenge of PSF modelling is to build *fast* yet *powerful* models.

Another challenge is to include complex effects and contributions that cannot be directly constrained from the observations into the PSF model. These contributions can be modelled with simulations and obtained from complementary observations or instrument characterisations. However, this complementary information will not precisely match the state of the telescope during the imaging procedure due to several reasons, e.g., changes in the telescope, measurement errors, and imperfect modelling. The better way to correct this information and adapt it to real observations has to be further studied.

The validation of PSF models from real observations is a challenging subject that requires further development, as access to the ground truth PSF field is unavailable. Although some validation methods exist, they are generally not very informative or are based on second-order moments that are not well-suited to describe diffraction-limited PSFs. We have presented the error propagation of a galaxy-shape measurement, where several limitations were exposed. Current error propagation methods have simplifying assumptions, e.g., the PSF is well described by its quadrupole moments that do not hold anymore with recent and upcoming telescopes. Further development of these methods is required to define realistic PSF model requirements and study how PSF modelling errors affect the target task.

Author contributions

TL coordinated the entire effort, produced the figures, and wrote the text. J-LS and MK provided valuable comments and feedback on the different sections of the article. All authors contributed to the article and approved the submitted version.

Funding

This work was partially funded by the TITAN ERA Chair project (Contract No. 101086741) within the Horizon Europe Framework Programme of the European Commission. This work is also partially supported by EPSRC (grant number EP/W007673/1).

Acknowledgments

Part of this review has been adapted from the PhD thesis (Liaudat, 2022). The authors acknowledge and resume in the following list the open-source simulating and modelling software mentioned in this review:

- GalSim (Rowe et al., 2015): github.com/GalSim-developers/GalSim
- WebbPSF (Perrin et al., 2014): github.com/spacetelescope/webbpsf
- PhoSim (Peterson et al., 2015): bitbucket.org/phosim/phosim_release/wiki/Home
- WaveDiff (Liaudat et al., 2023): github.com/CosmoStat/wf-psf
- ShapePipe (Farrens, S. et al., 2022): github.com/CosmoStat/shapepipe
- PIFF (Jarvis et al., 2020): github.com/rmjarvis/Piff
- MCCD (Liaudat et al., 2021a): github.com/CosmoStat/mccd
- RCA (Ngolè et al., 2016; Schmitz et al., 2020): github.com/CosmoStat/rca
- PSFEx (Bertin, 2011): github.com/astromatic/psfex
- Tiny Tim (Krist et al., 2011): github.com/spacetelescope/tinytim
- Photutils (Bradley et al., 2022): github.com/astropy/photutils
- ∂Lux (Desdoigts et al., 2023): github.com/LouisDesdoigts/dLux
- psf-weather-station: github.com/LSSTDESC/psf-weather-station.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akeson, R., Armus, L., Bachelet, E., Bailey, V., Bartusek, L., Bellini, A., et al. (2019). *The wide field infrared survey telescope: 100 hubbles for the 2020s*. arXiv e-prints , arXiv:1902.05569. doi:10.48550/arXiv.1902.05569

Anderson, J. (2016). Empirical models for the WFC3/IR PSF. *Tech. Rep*.

Anderson, J., and King, I. R. (2000). Toward high-precision astrometry with wfpc2. i. deriving an accurate point-spread function. *Publ. Astronomical Soc. Pac.* 112, 1360. doi:10.1086/316632

Anderson, J., and King, I. R. (2006). PSFs, photometry, and astronomy for the ACS/WFC. *Tech. Rep*.

Angeli, G. Z., Xin, B., Claver, C., MacMartin, D., Neill, D., Britton, M., et al. (2014). "Real time wavefront control system for the large synoptic survey telescope (LSST)," in *Proceeding volume 9150 modeling, systems engineering, and project management for astronomy VI*. Editors G. Z. Angeli, and P. Dierickx (Society of Photo-Optical Instrumentation Engineers SPIE), 91500H. doi:10.1117/12.2055390

Antilogus, P., Astier, P., Doherty, P., Guyonnet, A., and Regnault, N. (2014). The brighter-fatter effect and pixel correlations in CCD sensors. *J. Instrum.* 9, C03048. doi:10.1088/1748-0221/9/03/c03048

Baer, R. L. (2006). "A model for dark current characterization and simulation," in *Proceedings volume 6068, sensors, cameras, and systems for scientific/industrial applications VII; 606805*. Editor M. M. Blouke (International Society for Optics and Photonics SPIE), 37–48. doi:10.1117/12.639844

Baron, M., Sassolas, B., Frugier, P.-A., Venancio, L. M. G., Amiaux, J., Castelnau, M., et al. (2022). "Measurement and modelling of the chromatic dependence of a reflected wavefront on the Euclid space telescope dichroic mirror," in *Space telescopes and instrumentation 2022: optical, infrared, and millimeter wave*. Editors L. E. Coyle, S. Matsuura, and M. D. Perrin (International Society for Optics and Photonics SPIE), 12180, 121804V. doi:10.1117/12.2630072

Basden, A., Tubbs, B., and Mackay, C. (2004). "L3ccd's: low readout noise ccds in astronomy," in *Scientific detectors for astronomy*. Editors P. Amico, J. W. Beletic, and J. E. Beletic (Dordrecht: Springer Netherlands), 599–602.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2017). Automatic differentiation in machine learning: a survey. *J. Mach. Learn. Res.* 18, 5595–5637.

Beamer, B., Nomerotski, A., and Tsybychev, D. (2015). A study of astrometric distortions due to "tree rings" in CCD sensors using LSST Photon Simulator. *J. Instrum.* 10 (05), C05027. doi:10.1088/1748-0221/10/05/C05027

Beck, A., and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.* 2, 183–202. doi:10.1137/080716542

Beckers, J. M. (1993). Adaptive optics for astronomy: principles, performance, and applications. *Annu. Rev. Astronomy Astrophysics* 31, 13–62. doi:10.1146/annurev.aa.31.090193.000305

Bély, P., Hasan, H., and Miebach, M. (1993). Orbital focus variations in the hubble space telescope. *Tech. Rep*.

Beran, M. J., and Parrent, G. B. (1964). "Theory of partial coherence," in *Prentice-Hall International series in physics* (Englewood Cliffs, N.J: Prentice-Hall).

Bernstein, G. (2002). Advanced exposure-time calculations: undersampling, dithering, cosmic rays, astrometry, and ellipticities. *Publ. Astronomical Soc. Pac.* 114, 98–111. doi:10.1086/337997

Bernstein, G. M., and Armstrong, R. (2014). Bayesian lensing shear measurement. *Mon. Notices R. Astronomical Soc.* 438, 1880–1893. doi:10.1093/mnras/stt2326

Bernstein, G. M., Armstrong, R., Krawiec, C., and March, M. C. (2016). An accurate and practical method for inference of weak gravitational lensing from galaxy images. *Mon. Notices R. Astronomical Soc.* 459, 4467–4484. doi:10.1093/mnras/stw879

Bertin, E., and Arnouts, S. (1996). Sextractor: software for source extraction. *Astron. Astrophys. Suppl. Ser.* 117, 393–404. doi:10.1051/aas:1996164

Bertin, E. (2011). "Automated morphometry with SExtractor and PSFEx," in *Astronomical data analysis software and systems XX*. Editors I. N. Evans, A. Accomazzi, D. J. Mink, and A. H. Rots (Astronomical Society of the Pacific Conference Series), 442, 435.

Beuzit, J.-L., Vigan, A., Mouillet, D., Dohlen, K., Gratton, R., Boccaletti, A., et al. (2019). Sphere: the exoplanet imager for the very large telescope. *A&A* 631, A155. doi:10.1051/0004-6361/201935251

Biesiadzinski, T., Lorenzon, W., Newman, R., Schubnell, M., Tarlé, G., and Weaverdyck, C. (2011). Measurement of reciprocity failure in near-infrared detectors. *Publ. Astronomical Soc. Pac.* 123, 179. doi:10.1086/658282

Bohlin, R. C., Lindler, D. J., and Riess, A. (2005). Grism sensitivities and apparent non-linearity. *Tech. Rep*.

Born, M., and Wolf, E. (1999). *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. 7 edn. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139644181

Bosch, J., Armstrong, R., Bickerton, S., Furusawa, H., Ikeda, H., Koike, M., et al. (2017). The Hyper Suprime-Cam software pipeline. *Publ. Astronomical Soc. Jpn.* 70. doi:10.1093/pasj/psx080.S5

Bradley, L., Sipőcz, B., Robitaille, T., Tollerud, E., Vinícius, Z., Deil, C., et al. (2022). *astropy/photutils: 1.5.0*. Zenodo. doi:10.5281/zenodo.6825092

Breckinridge, J. B., Lam, W. S. T., and Chipman, R. A. (2015). Polarization aberrations in astronomical telescopes: the point spread function. *Publ. Astronomical Soc. Pac.* 127, 445. doi:10.1086/681280

Breckinridge, J. B., and Wood, H. J. (1993). Space optics: an introduction by the editors. *Appl. Opt.* 32, 1677–1680. doi:10.1364/AO.32.001677

Burke, C. J., Peterson, J. R., Egami, E., Leisenring, J. M., Sembroski, G. H., and Rieke, M. J. (2019). PhoSim-NIRCam: photon-by-photon image simulations of the James Webb Space Telescope's near-infrared camera. *J. Astron. Telesc. Instrum. Syst.* 5 (3), 1–14. doi:10.1117/1.JATIS.5.3.038002

Carlsten, S. G., Strauss, M. A., Lupton, R. H., Meyers, J. E., and Miyazaki, S. (2018). Wavelength-dependent PSFs and their impact on weak lensing measurements. *Monthly Notices of the Royal Astronomical Society*.

Chang, C., Marshall, P. J., Jernigan, J. G., Peterson, J. R., Kahn, S. M., Gull, S. F., et al. (2012). Atmospheric point spread function interpolation for weak lensing in short exposure imaging data. *Monthly Notices of the Royal Astronomical Society*), 427 (3). 2572–2587. doi:10.1111/j.1365-2966.2012.22134.x

Cheng, J., and Peterson, J. P. (2017). *Strong gravitational lens modeling of the cosmic horseshoe and photon simulation of DECam images*. Indiana: Purdue University (10266501), 159.

Claver, C. F., Chandrasekharan, S., Liang, M., Xin, B., Alagoz, E., Arndt, K., et al. (2012). "Prototype pipeline for LSST wavefront sensing and reconstruction," in *Ground-based and airborne telescopes IV*. Editors L. M. Stepp, R. Gilmozzi, and H. J. Hall (International Society for Optics and Photonics SPIE), 8444. 84444P. doi:10.1117/12.926472

Condat, L. (2013). A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.* 158, 460–479.

Coulton, W. R., Armstrong, R., Smith, K. M., Lupton, R. H., and Spergel, D. N. (2018). Exploring the brighter-fatter effect with the hyper suprime-cam. *ApJ* 155, 258.

Crew, H., Huygens, C., Young, T., Fresnel, A. J., and Arago, F. (1900). *The wave theory of light: memoirs of huygens, young and fresnel*. New York: American Book Co.

Cropper, M., Hoekstra, H., Kitching, T., Massey, R., Amiaux, J., Miller, L., et al. (2013). Defining a weak lensing experiment in space. *MNRAS* 431, 3103–3126.

Davies, R., and Kasper, M. (2012). Adaptive optics for astronomy. *Annu. Rev. Astronomy Astrophysics* 50, 305–351. doi:10.1146/annurev-astro-081811-125447

Davis, C. P., Rodriguez, J., and Roodman, A. (2016). "Wavefront-based PSF estimation," in *Ground-based and airborne telescopes VI*. Editors H. J. Hall, R. Gilmozzi, and H. K. Marshall (International Society for Optics and Photonics SPIE), 9906, 2156–2168. doi:10.1117/12.2233366

de Vries, W. H., Olivier, S. S., Asztalos, S. J., Rosenberg, L. J., and Baker, K. L. (2007). Image ellipticity from atmospheric aberrations. *Astrophysical J.* 662, 744–749. doi:10.1086/517873

Debarnot, V., Escande, P., Mangeat, T., and Weiss, P. (2021a). Learning low-dimensional models of microscopes. *IEEE Trans. Comput. Imaging* 7, 178–190. doi:10.1109/TCI.2020.3048295

Debarnot, V., Escande, P., Mangeat, T., and Weiss, P. (2021b). "Modelling a microscope as low dimensional subspace of operators," in *2020 28th European signal processing conference* (EUSIPCO), 765–769. doi:10.23919/Eusipco47968.2020.9287603

Desdoigts, L., Pope, B. J. S., Dennis, J., and Tuthill, P. G. (2023). Differentiable optics with ∂Lux: I—deep calibration of flat field and phase retrieval with automatic differentiation. *J. Astronomical Telesc. Instrum. Syst.* 9, 028007. doi:10.1117/1.JATIS.9.2.028007

Donlon, K., Ninkov, Z., and Baum, S. (2018). Point-spread function ramifications and deconvolution of a signal dependent blur kernel due to interpixel capacitive coupling. *Publ. Astronomical Soc. Pac.* 130, 074503. doi:10.1088/1538-3873/aac261

Dougherty, G., and Kawaf, Z. (2001). The point spread function revisited: image restoration using 2-d deconvolution. *Radiography* 7, 255–262. doi:10.1053/radi.2001.0341

Ellerbroek, B. L. (2002). Efficient computation of minimum-variance wave-front reconstructors with sparse matrix techniques. *J. Opt. Soc. Am. A* 19, 1803–1816. doi:10.1364/JOSAA.19.001803

Erben, T., Hildebrandt, H., Miller, L., van Waerbeke, L., Heymans, C., Hoekstra, H., et al. (2013). CFHTLenS: the Canada–France–Hawaii telescope lensing survey – imaging data and catalogue products. *Mon. Notices R. Astronomical Soc.* 433, 2545–2563. doi:10.1093/mnras/stt928

Eriksen, M., and Hoekstra, H. (2018). Implications of a wavelength-dependent PSF for weak lensing measurements. *Mon. Notices R. Astronomical Soc.* 477, 3433–3448. doi:10.1093/mnras/sty830

Euclid Collaboration, Schirmer, M., Thürmer, K., Bras, B., Cropper, M., Martin-Fleitas, J., et al. (2023a). *Euclid preparation. XXIX. Water ice in spacecraft part I: the physics of ice formation and contamination*. arXiv e-prints, arXiv:2305.10107. doi:10.48550/arXiv.2305.10107

Euclid Collaboration, Merlin, E., Castellano, M., Bretonnière, M., Huertas-Company, M., Kuchner, U., Tuccillo, D., et al. (2023b). Euclid preparation - XXV. The Euclid Morphology Challenge: towards model-fitting photometry for billions of galaxies. *A&A* 671, A101. doi:10.1051/0004-6361/202245041

Euclid Collaboration, Bretonnière, H., Kuchner, U., Huertas-Company, M., Merlin, E., Castellano, M., Tuccillo, D., et al. (2023c). Euclid preparation - XXVI. The Euclid Morphology Challenge: Towards structural parameters for billions of galaxies. *A&A* 671, A102. doi:10.1051/0004-6361/202245042

Farrens, S., Guinot, A., Kilbinger, M., Liaudat, T., Baumont, L., Jimenez, X., et al. (2022). Shapepipe: a modular weak-lensing processing and analysis pipeline. *A&A* 664, A141. doi:10.1051/0004-6361/202243970

Fenech Conti, I., Herbonnet, R., Hoekstra, H., Merten, J., Miller, L., and Viola, M. (2017). Calibration of weak-lensing shear in the kilo-degree survey. *Mon. Notices R. Astronomical Soc.* 467, 1627–1651. doi:10.1093/mnras/stx200

Fenech Conti, I. (2017). *Point spread function modelling and shear calibration for weak lensing surveys*. Malta: Ph.D. thesis, University of Malta.

Fienup, J. (1994). "Diagnosing the aberrations of the hubble space telescope," in *Current trends in optics*. Editor J. Dainty (Academic Press), 279–287. chap. 20.

Fienup, J. R., Marron, J. C., Schulz, T. J., and Seldin, J. H. (1993). Hubble space telescope characterized by using phase-retrieval algorithms. *Appl. Opt.* 32, 1747–1767. doi:10.1364/AO.32.001747

Fienup, J. R. (1993). Phase-retrieval algorithms for a complicated optical system. *Appl. Opt.* 32, 1737–1746. doi:10.1364/AO.32.001737

Finner, K., Lee, B., Chary, R.-R., Jee, M. J., Hirata, C., Congedo, G., et al. (2023). *Near-IR weak-lensing (NIRWL) measurements in the CANDELS fields I: point-spread function modeling and systematics*. arXiv e-prints , arXiv:2301.07725. doi:10.48550/arXiv.2301.07725

Flaugher, B., Diehl, H. T., Honscheid, K., Abbott, T. M. C., Alvarez, O., Angstadt, R., et al. (2015). The dark energy camera. *Astronomical J.* 150, 150. doi:10.1088/0004-6256/150/5/150

Fresnel, A. (1819). *Mémoire sur la diffraction de la lumière (De l'Imprimerie de Feugueray, rue du Cloitre Saint-Benoit, no. 4)*.

Fruchter, A. S., and Hook, R. N. (2002). Drizzle: a method for the linear reconstruction of undersampled images. *Publ. Astronomical Soc. Pac.* 114, 144–152.

Fu, L., Liu, D., Radovich, M., Liu, X., Pan, C., Fan, Z., et al. (2018). Weak-lensing study in VOICE survey – I. Shear measurement. *Mon. Notices R. Astronomical Soc.* 479, 3858–3872. doi:10.1093/mnras/sty1579

Fu, L., Semboloni, E., Hoekstra, H., Kilbinger, M., van Waerbeke, L., Tereno, I., et al. (2008). Very weak lensing in the cfhtls wide: cosmology from cosmic shear in the linear regime. *A&A* 479, 9–25. doi:10.1051/0004-6361:20078522

Gaia Collaboration, , Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., Vallenari, A., Babusiaux, C., et al. (2016). The gaia mission. *A&A* 595, A1. doi:10.1051/0004-6361/201629272

Gaskill, J. D. (1978). "Linear systems, Fourier transforms, and optics," in *Wiley series in pure and applied optics* (New York: Wiley).

Gatti, M., Sheldon, E., Amon, A., Becker, M., Troxel, M., Choi, A., et al. (2021). Dark energy survey year 3 results: weak lensing shape catalogue. *Mon. Notices R. Astronomical Soc.* 504, 4312–4336. doi:10.1093/mnras/stab918

Gentile, M., Courbin, F., and Meylan, G. (2013). Interpolating point spread function anisotropy. *A&A* 549, A1.

Gerchberg, R. W., and Saxton, W. O. (1972). A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* 35, 237–246.

Giblin, B., Heymans, C., Asgari, M., Hildebrandt, H., Hoekstra, H., Joachimi, B., et al. (2021). Kids-1000 catalogue: weak gravitational lensing shear measurements. *A&A* 645, A105. doi:10.1051/0004-6361/202038850

Gillis, B. R., Schrabback, T., Marggraf, O., Mandelbaum, R., Massey, R., Rhodes, J., et al. (2020). Validation of PSF models for HST and other space-based observations. *Mon. Notices R. Astronomical Soc.* 496, 5017–5038. doi:10.1093/mnras/staa1818

Goodman, J. W. (2005).Introduction to fourier optics. Editor J. W. Goodman 3rd ed. (Englewood, CO: Roberts & Co. Publishers), 1.

Goodman, J. W. (1985). *Statistical optics*. New York: Wiley-Interscience, 567.

Graham, J. R., Macintosh, B., Doyon, R., Gavel, D., Larkin, J., Levine, M., et al. (2007). *Ground-based direct detection of exoplanets with the gemini planet imager (GPI)*. arXiv e-prints, arXiv:0704.1454. doi:10.48550/arXiv.0704.1454

Gross, H. (2005). *Handbook of optical systems: Volume 1: fundamentals of technical optics*. John Wiley & Sons, Ltd.

Gross, H. H., Zügge, H., Peschka, M., and Blechinger, F. (2006). *Handbook of optical systems: Volume 3: aberration theory and correction of optical systems*. John Wiley & Sons, Ltd.

Gruen, D., Bernstein, G., Jarvis, M., Rowe, B., Vikram, V., Plazas, A., et al. (2015). Characterization and correction of charge-induced pixel shifts in decam. *J. Instrum.* 10, C05032. doi:10.1088/1748-0221/10/05/C05032

Guinot, A., Kilbinger, M., Farrens, S., Peel, A., Pujol, A., Schmitz, M., et al. (2022). *Shapepipe: a new shape measurement pipeline and weak-lensing application to unions/cfis data*. arXiv:2204.04798. doi:10.48550/ARXIV.2204.04798

Gunn, J. E., Carr, M., Rockosi, C., Sekiguchi, M., Berry, K., Elms, B., et al. (1998). The sloan digital sky survey photometric camera. *Astronomical J.* 116, 3040. doi:10.1086/300645

Guyonnet, A., Astier, P., Antilogus, P., Regnault, N., and Doherty, P. (2015). Evidence for self-interaction of charge distribution in charge-coupled devices. *Astronomy Astrophysics* 575, A41. doi:10.1051/0004-6361/201424897

Hambly, N., MacGillivray, H., Read, M., Tritton, S., Thomson, E., Kelly, B., et al. (2001). The SuperCOSMOS sky survey - i. introduction and description. *Mon. Notices R. Astronomical Soc.* 326, 1279–1294. doi:10.1111/j.1365-2966.2001.04660.x

Hasan, H., Burrows, C. J., and Schroeder, D. J. (1993). Focus history of the hubble space telescope—Launch to may 1993. *Publ. Astronomical Soc. Pac.* 105, 1184–1191.

Hecht, E. (2017). *Optics*. fifth edition edn. Harlow, England: Pearson Education.

Herbel, J., Kacprzak, T., Amara, A., Refregier, A., and Lucchi, A. (2018). Fast point spread function modeling with deep learning. *J. Cosmol. Astropart. Phys.* 054. doi:10.1088/1475-7516/2018/07/054

Heymans, C., Rowe, B., Hoekstra, H., Miller, L., Erben, T., Kitching, T., et al. (2012). The impact of high spatial frequency atmospheric distortions on weak-lensing measurements. *Mon. Notices R. Astronomical Soc.* 421, 381–389. doi:10.1111/j.1365-2966.2011.20312.x

High, F. W., Rhodes, J., Massey, R., and Ellis, R. (2007). Pixelation effects in weak lensing. *Publ. Astronomical Soc. Pac.* 119, 1295–1307.

Hildebrandt, H., Viola, M., Heymans, C., Joudaki, S., Kuijken, K., Blake, C., et al. (2016). KiDS-450: cosmological parameter constraints from tomographic weak gravitational lensing. *Mon. Notices R. Astronomical Soc.* 465, 1454–1498. doi:10.1093/mnras/stw2805

Hirata, C., and Seljak, U. (2003). Shear calibration biases in weak-lensing surveys. *MNRAS* 343, 459–480.

Hirata, C. M., Yamamoto, M., Laliotis, K., Macbeth, E., Troxel, M. A., Zhang, T., et al. (2023). *Simulating image coaddition with the nancy grace roman space telescope: I. Simulation methodology and general results*. arXiv e-prints , arXiv:2303.08749. doi:10.48550/arXiv.2303.08749

Hoffmann, S. L., and Anderson, J. (2017). *A study of PSF models for ACS/WFC, 8*.

Hogg, D. W. (2022). *Magnitudes, distance moduli, bolometric corrections, and so much more*. doi:10.48550/ARXIV.2206.00989

Hopkins, H. H. (1957). Applications of coherence theory in microscopy and interferometry. *J. Opt. Soc. Am.* 47, 508–526. doi:10.1364/JOSA.47.000508

Huff, E., and Mandelbaum, R. (2017). *Metacalibration: direct self-calibration of biases in shear measurement*. arXiv preprint arXiv:1702.02600.

Huygens, C. (1690). *Traité de la lumière: où sont expliquées les causes de ce qui lui arrive dans la reflexion; dans la refraction et particulièrement dans l'étrange réfraction du cristal d'Islande (Chez Pierre vander Aa, marchand libraire)*. doi:10.5479/sil.294285.39088000545160

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., et al. (2019). Lsst: from science drivers to reference design and anticipated data products. *Astrophysical J.* 873, 111. doi:10.3847/1538-4357/ab042c

Jarvis, M., and Jain, B. (2004). *Principal component analysis of psf variation in weak lensing surveys*. doi:10.48550/ARXIV.ASTRO-PH/0412234

Jarvis, M., Sheldon, E., Zuntz, J., Kacprzak, T., Bridle, S., Amara, A., et al. (2016). The des science verification weak lensing shear catalogues. *MNRAS* 460, 2245–2281.

Jarvis, M., Bernstein, G. M., Amon, A., Davis, C., Léget, P. F., Bechtol, K., et al. (2020). Dark energy survey year 3 results: point spread function modelling. *Mon. Notices R. Astronomical Soc.* 501, 1282–1299. doi:10.1093/mnras/staa3679

Jee, M. J., Blakeslee, J. P., Sirianni, M., Martel, A. R., White, R. L., and Ford, H. C. (2007). Principal component analysis of the time- and position-dependent point-spread function of the advanced camera for surveys. *Publ. Astronomical Soc. Pac.* 119, 1403–1419. doi:10.1086/524849

Jee, M. J., and Tyson, J. A. (2011). Toward precision lsst weak-lensing measurement. i. impacts of atmospheric turbulence and optical aberration. *Publ. Astronomical Soc. Pac.* 123, 596–614. doi:10.1086/660137

Jia, P., Li, X., Li, Z., Wang, W., and Cai, D. (2020a). Point spread function modelling for wide-field small-aperture telescopes with a denoising autoencoder. *Mon. Notices R. Astronomical Soc.* 493, 651–660. doi:10.1093/mnras/staa319

Jia, P., Wu, X., Yi, H., Cai, B., and Cai, D. (2020b). PSF-NET: a nonparametric point-spread function model for ground-based optical telescopes. *AJ* 159, 183. doi:10.3847/1538-3881/ab7b79

Jones, R. C. (1941). A new calculus for the treatment of optical systemsi. description and discussion of the calculus. *J. Opt. Soc. Am.* 31, 488–493. doi:10.1364/JOSA.31.000488

Joyce, K. T., Bardsley, J. M., and Luttman, A. (2018). Point spread function estimation in x-ray imaging with partially collapsed gibbs sampling. *SIAM J. Sci. Comput.* 40, B766–B787. doi:10.1137/17M1149250

Kaiser, N., Squires, G., and Broadhurst, T. (1995). A method for weak lensing observations. *ApJ* 449, 460.

Kannawadi, A., Shapiro, C. A., Mandelbaum, R., Hirata, C. M., Kruk, J. W., and Rhodes, J. D. (2016). The impact of interpixel capacitance in CMOS detectors on PSF shapes and implications for WFIRST. *Publ. Astronomical Soc. Pac.* 128, 095001. doi:10.1088/1538-3873/128/967/095001

Kannawadi, A., Rosenberg, E., and Hoekstra, H. (2021). Mitigating the effects of undersampling in weak lensing shear estimation with metacalibration. *Mon. Notices R. Astronomical Soc.* 502, 4048–4063. doi:10.1093/mnras/stab211

Kármán, T. V. (1930). "Mechanische aenlichkeit und turbulenz," in *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 58–76.

Kilbinger, M. (2015). Cosmology with cosmic shear observations: a review. *Rep. Prog. Phys.* 78, 086901.

Kitching, T. D., Miller, L., Heymans, C. E., van Waerbeke, L., and Heavens, A. F. (2008). Bayesian galaxy shape measurement for weak lensing surveys - II. application to simulations. *Mon. Notices R. Astronomical Soc.* 390, 149–167. doi:10.1111/j.1365-2966.2008.13628.x

Kolmogorov, A. N. (1991). The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *Proc. R. Soc. Lond. Ser. A* 434, 9–13. doi:10.1098/rspa.1991.0075

Krist, J. (1993). "Tiny tim: an HST PSF simulator," in *Astronomical data analysis software and systems II*. Editors R. J. Hanisch, R. J. V. Brissenden, and J. Barnes (Astronomical Society of the Pacific Conference Series) 52, 536.

Krist, J. E. (1995). "Wfpc2 ghosts, scatter and psf field dependence," in *Calibrating hubble space telescope. Post servicing mission. Proceedings of a workshop held at the space telescope science institute*. Editors A. Koratkar, and C. Leitherer Baltimore, Maryland.

Krist, J. (2003). ACS WFC & HRC fielddependent PSF variations due to optical and charge diffusion effects. *Tech. Rep*. Space Telescope Science Institute.

Krist, J. E., and Burrows, C. J. (1995). Phase-retrieval analysis of pre- and post-repair hubble space telescope images. *Appl. Opt.* 34, 4951–4964. doi:10.1364/AO.34.004951

Krist, J. E., Hook, R. N., and Stoehr, F. (2011). "20 years of hubble space telescope optical modeling using tiny tim," in *Optical modeling and performance predictions V*. Editor M. A. Kahan (International Society for Optics and Photonics SPIE), 8127, 166–181. doi:10.1117/12.892762

Kuijken, K., Heymans, C., Hildebrandt, H., Nakajima, R., Erben, T., de Jong, J. T., et al. (2015). Gravitational lensing analysis of the kilo-degree survey. *MNRAS* 454, 3500–3532.

Kuntzer, T., and Courbin, F. (2017). Detecting unresolved binary stars in euclid vis images. *A&A* 606, A119.

Lallo, M. D., Makidon, R. B., Casertano, S., and Krist, J. E. (2006). "Temporal optical behavior of HST: focus, coma, and astigmatism history," in *Observatory operations: strategies, processes, and systems*. Editors D. R. Silva, and R. E. Doxsey (Society of Photo-Optical Instrumentation Engineers SPIE), 6270, 62701N. doi:10.1117/12.672040

Lauer, T. R. (1999a). Combining undersampled dithered images. *Publ. Astronomical Soc. Pac.* 111, 227. doi:10.1086/316319

Lauer, T. R. (1999b). The photometry of undersampled point-spread functions. *Publ. Astronomical Soc. Pac.* 111, 1434. doi:10.1086/316460

Laureijs, R., Amiaux, J., Arduini, S., Augueres, J.-L., Brinchmann, J., Cole, R., et al. (2011). *Euclid definition study report*. ArXiv e-prints.

Leauthaud, A., Massey, R., Kneib, J.-P., Rhodes, J., Johnston, D. E., Capak, P., et al. (2007). Weak gravitational lensing with COSMOS: galaxy selection and shape measurements. *Astrophysical J. Suppl. Ser.* 172, 219–238. doi:10.1086/516598

Li, B.-S., Li, G.-L., Cheng, J., Peterson, J., and Cui, W. (2016). The point spread function reconstruction by using moffatlets — I. *Res. Astronomy Astrophysics* 16, 007. doi:10.1088/1674-4527/16/9/139

Liaudat, T., Bonnin, J., Starck, J.-L., Schmitz, M. A., Guinot, A., Kilbinger, M., et al. (2021a). Multi-CCD modelling of the point spread function. *A&A* 646, A27. doi:10.1051/0004-6361/202039584

Liaudat, T., Starck, J.-L., Kilbinger, M., and Frugier, P.-A. (2021b). "Rethinking the modeling of the instrumental response of telescopes with a differentiable optical model," in *NeurIPS 2021 machine learning for physical sciences workshop*.

Liaudat, T. I. (2022). "Data-driven modelling of ground-based and space-based telescope's point spread functions," in *Ph.D. Thesis, université paris-saclay, saclay, France. 2022UPASP118*.

Liaudat, T., Starck, J.-L., Kilbinger, M., and Frugier, P.-A. (2023). Rethinking data-driven point spread function modeling with a differentiable optical model. *Inverse Probl.* 39, 035008. doi:10.1088/1361-6420/acb664

Lin, C.-H., Tan, B., Mandelbaum, R., and Hirata, C. M. (2020). The impact of light polarization effects on weak lensing systematics. *Mon. Notices R. Astronomical Soc.* 496, 532–539. doi:10.1093/mnras/staa1298

Lindstrand, M. (2019). Sensor interpixel correlation analysis and reduction: a review. *J. Imaging Sci. Technol.* 63, 010502. doi:10.2352/J.ImagingSci.Technol.2019.63.1.010502

LSST Science Collaboration, Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., et al. (2009). *LSST science book*. Version 2.0. arXiv e-prints.

Lu, T., Zhang, J., Dong, F., Li, Y., Liu, D., Fu, L., et al. (2017). Testing PSF interpolation in weak lensing with real data. *Astronomical J.* 153, 197. doi:10.3847/1538-3881/aa661e

Lupton, R., Gunn, J. E., Ivezic, Z., Knapp, G. R., Kent, S., and Yasuda, N. (2001). *The sdss imaging pipelines*. doi:10.48550/ARXIV.ASTRO-PH/0101420

Macintosh, B., Graham, J. R., Ingraham, P., Konopacky, Q., Marois, C., Perrin, M., et al. (2014). First light of the gemini planet imager. *Proc. Natl. Acad. Sci.* 111, 12661–12666. doi:10.1073/pnas.1304215111

Macintosh, B., Chilcote, J. K., Bailey, V. P., de Rosa, R., Nielsen, E., Norton, A., et al. (2018). "The gemini planet imager: looking back over five years and forward to the future," in *Adaptive optics systems VI*. Editors L. M. Close, L. Schreiber, and D. Schmidt (Society of Photo-Optical Instrumentation Engineers SPIE), 10703, 107030. doi:10.1117/12.2314253

Makidon, R. B., Lallo, M. D., Casertano, S., Gilliland, R. L., Sirianni, M., and Krist, J. E. (2006). "The temporal optical behavior of the hubble space telescope: the impact on science observations," in *Observatory operations: strategies, processes, and systems*. Editors D. R. Silva, and R. E. Doxsey (Society of Photo-Optical Instrumentation Engineers SPIE), 6270, 62701L. doi:10.1117/12.671604

Mandelbaum, R., Hirata, C. M., Seljak, U., Guzik, J., Padmanabhan, N., Blake, C., et al. (2005). Systematic errors in weak lensing: application to sdss galaxy-galaxy weak lensing. *MNRAS* 361, 1287–1322.

Mandelbaum, R., Miyatake, H., Hamana, T., Oguri, M., Simet, M., Armstrong, R., et al. (2017). The first-year shear catalog of the subaru hyper suprime-cam subaru strategic program survey. *Publ. Astronomical Soc. Jpn.* 70. doi:10.1093/pasj/psx130.S25

Mandelbaum, R. (2018). Weak lensing for precision cosmology. *Annu. Rev. Astronomy Astrophysics* 56, 393–433. doi:10.1146/annurev-astro-081817-051928

Mandelbaum, R., Lanusse, F., Leauthaud, A., Armstrong, R., Simet, M., Miyatake, H., et al. (2018). Weak lensing shear calibration with simulations of the HSC survey. *Mon. Notices R. Astronomical Soc.* 481, 3170–3195. doi:10.1093/mnras/sty2420

Mandelbaum, R., Jarvis, M., Lupton, R. H., Bosch, J., Kannawadi, A., Murphy, M. D., et al. (2022). *PSFs of coadded images*. arXiv e-prints, arXiv:2209.09253.

Manuel, A. M., Phillion, D. W., Olivier, S. S., Baker, K. L., and Cannon, B. (2010). Curvature wavefront sensing performance evaluation for active correction of the large synoptic survey telescope (lsst). *Opt. Express* 18, 1528–1552. doi:10.1364/OE.18.001528

Massey, R., Hoekstra, H., Kitching, T., Rhodes, J., Cropper, M., Amiaux, J., et al. (2012). Origins of weak lensing systematics, and requirements on future instrumentation (or knowledge of instrumentation). *MNRAS* 429, 661–678.

Massey, R., and Refregier, A. (2005). Polar shapelets. *Mon. Notices R. Astronomical Soc.* 363, 197–210. doi:10.1111/j.1365-2966.2005.09453.x

Massey, R., Schrabback, T., Cordes, O., Marggraf, O., Israel, H., Miller, L., et al. (2014). An improved model of charge transfer inefficiency and correction algorithm for the Hubble Space Telescope. *Mon. Notices R. Astronomical Soc.* 439, 887–907. doi:10.1093/mnras/stu012

Massey, R., Schrabback, T., Cordes, O., Marggraf, O., Israel, H., Miller, L., et al. (2014). An improved model of charge transfer inefficiency and correction algorithm for the Hubble Space Telescope. *Mon. Notices R. Astronomical Soc.* 439, 887–907. doi:10.1093/mnras/stu012

McCullough, P. (2008). Inter-pixel capacitance: prospects for deconvolution. *Tech. Rep*.

McGuire, J. P., and Chipman, R. A. (1990). Diffraction image formation in optical systems with polarization aberrations. i: formulation and example. *J. Opt. Soc. Am. A* 7, 1614–1626. doi:10.1364/JOSAA.7.001614

McGuire, J. P., and Chipman, R. A. (1991). Diffraction image formation in optical systems with polarization aberrations. ii: amplitude response matrices for rotationally symmetric systems. *J. Opt. Soc. Am. A* 8, 833–840. doi:10.1364/JOSAA.8.000833

Melchior, P., and Viola, M. (2012). Means of confusion: how pixel noise affects shear estimates for weak gravitational lensing. *Mon. Notices R. Astronomical Soc.* 424, 2757–2769. doi:10.1111/j.1365-2966.2012.21381.x

Melchior, P., Viola, M., Schäfer, B. M., and Bartelmann, M. (2011). Weak gravitational lensing with deimos. *MNRAS* 412, 1552–1558.

Merz, G., Liu, Y., Burke, C. J., Aleo, P. D., Liu, X., Kind, C., et al. (2023). Detection, instance segmentation, and classification for astronomical surveys with deep learning (DeepDISC): detectron2 implementation and demonstration with hyper suprime-cam data. *Monthly Notices of the Royal Astronomical Society*, stad2785. doi:10.1093/mnras/stad2785

Meyers, J. E., and Burchat, P. R. (2015a). "Chromatic ccd effects on weak lensing measurements for lsst," in *arXiv: instrumentation and methods for astrophysics*.

Meyers, J. E., and Burchat, P. R. (2015b). Impact of atmospheric chromatic effects on weak lensing measurements. *Astrophysical J.* 807, 182. doi:10.1088/0004-637x/807/2/182

Miller, L., Kitching, T. D., Heymans, C., Heavens, A. F., and Van Waerbeke, L. (2007). Bayesian galaxy shape measurement for weak lensing surveys – I. Methodology and a fast-fitting algorithm. *Mon. Notices R. Astronomical Soc.* 382, 315–324. doi:10.1111/j.1365-2966.2007.12363.x

Miller, L., Heymans, C., Kitching, T., Van Waerbeke, L., Erben, T., Hildebrandt, H., et al. (2013). Bayesian galaxy shape measurement for weak lensing surveys–iii. application to the canada–france–hawaii telescope lensing survey. *MNRAS* 429, 2858–2880.

Moffat, A. F. J. (1969). A theoretical investigation of focal stellar images in the photographic emulsion and application to photographic photometry. *A&A* 3, 455.

Nardiello, D., Bedin, L. R., Burgasser, A., Salaris, M., Cassisi, S., Griggio, M., et al. (2022). Photometry and astrometry with JWST – I. NIRCam point spread functions and the first JWST colour–magnitude diagrams of a globular cluster. *Mon. Notices R. Astronomical Soc.* 517, 484–497. doi:10.1093/mnras/stac2659

Neill, D., Angeli, G., Claver, C., Hileman, E., DeVries, J., Sebag, J., et al. (2014). "Overview of the LSST active optics system," in *Modeling, systems engineering, and project management for astronomy VI*. Editors G. Z. Angeli, and P. Dierickx (Society of Photo-Optical Instrumentation Engineers SPIE), 9150, 91500G. doi:10.1117/12.2056553

Ngolè, F., Starck, J.-L., Ronayette, S., Okumura, K., and Amiaux, J. (2015). Super-resolution method using sparse regularization for point-spread function recovery. *A&A* 575, A86.

Ngolè, F., Starck, J.-L., Okumura, K., Amiaux, J., and Hudelot, P. (2016). Constraint matrix factorization for space variant psfs field restoration. *Inverse Probl.* 32, 124001.

Ngolè, F., and Starck, J.-L. (2017). Point spread function field learning based on optimal transport distances. *SIAM J. Imaging Sci.* 10, 1549–1578.

Nie, L., Li, G., Peterson, J. R., and Wei, C. (2021a). The point spread function reconstruction – II. The smooth PCA. *Mon. Notices R. Astronomical Soc.* 503, 4436–4445. doi:10.1093/mnras/stab733

Nie, L., Li, G., Zhang, J., Fan, Z., and Peterson, J. R. (2021b). Constraining the point spread function by using galaxy images. *Mon. Notices R. Astronomical Soc.* 508 (3), 3785–3795. doi:10.1093/mnras/stab2824

Niemi, S.-M., Cropper, M., Szafraniec, M., and Kitching, T. (2015). Measuring a charge-coupled device point spread function. *Exp. Astron.* 39, 207–231.

Nino, D., Makidon, R., Lallo, M., Sahu, K., Sirianni, M., and Casertano, S. (2007). Hst focus variations with temperatures. *Instrum. Sci. Rep. ACS*, 2008–2103.

Noll, R. J. (1976). Zernike polynomials and atmospheric turbulence. *J. Opt. Soc. Am.* 66, 207–211. doi:10.1364/JOSA.66.000207

Nyquist, H. (1928). Thermal agitation of electric charge in conductors. *Phys. Rev.* 32, 110–113. doi:10.1103/PhysRev.32.110

Parikh, N., and Boyd, S. (2014). Proximal algorithms. *Found. Trends Optim.* 1, 127–239. doi:10.1561/2400000003

Paulin-Henriksson, S., Amara, A., Voigt, L., Refregier, A., and Bridle, S. (2008). Point spread function calibration requirements for dark energy from cosmic shear. *A&A* 484, 67–77.

Perrin, M. D., Sivaramakrishnan, A., Lajoie, C.-P., Elliott, E., Pueyo, L., Ravindranath, S., et al. (2014). "Updated point spread function simulations for JWST with WebbPSF," in *Space telescopes and instrumentation 2014: optical, infrared, and millimeter wave*. Editors J. Oschmann, M. Jacobus, M. Clampin, G. G. Fazio, and H. A. MacEwen (Society of Photo-Optical Instrumentation Engineers SPIE), 9143. 91433X. doi:10.1117/12.2056689

Perrin, M. D., Soummer, R., Elliott, E. M., Lallo, M. D., and Sivaramakrishnan, A. (2012). "Simulating point spread functions for the james webb space telescope with WebbPSF," in *Space telescopes and instrumentation 2012: optical, infrared, and millimeter wave*. Editors M. C. Clampin, G. G. Fazio, H. A. MacEwen, J. Oschmann, and M. Jacobus (Society of Photo-Optical Instrumentation Engineers SPIE), 8442, 84423D. doi:10.1117/12.925230

Peterson, J. R., Jernigan, J. G., Kahn, S. M., Rasmussen, A. P., Peng, E., Ahmad, Z., et al. (2015). Simulation of astronomical images from optical survey telescopes using a comprehensive photon monte carlo approach. *Astrophysical J. Suppl. Ser.* 218, 14. doi:10.1088/0067-0049/218/1/14

Peterson, J. R., Peng, E., Burke, C. J., Sembroski, G., and Cheng, J. (2019). Deformation of optics for photon Monte Carlo simulations. *Astrophysical J.* 873, 98. doi:10.3847/1538-4357/ab0418

Peterson, J. R., O'Connor, P., Nomerotski, A., Magnier, E., Jernigan, J. G., Cheng, J., et al. (2020). Sensor distortion effects in photon Monte Carlo simulations. *Astrophysical J.* 889, 182. doi:10.3847/1538-4357/ab64e0

Peyré, G., and Cuturi, M. (2019). Computational optimal transport: with applications to data science. *Found. Trends® Mach. Learn.* 11, 355–607. doi:10.1561/2200000073

Plazas, A., Bernstein, G., and Sheldon, E. (2014). On-sky measurements of the transverse electric fields' effects in the dark energy camera ccds. *Publ. Astronomical Soc. Pac.*

Plazas, A. A., Shapiro, C., Kannawadi, A., Mandelbaum, R., Rhodes, J., and Smith, R. (2016). The effect of detector nonlinearity on wfirst psf profiles for weak gravitational lensing measurements. *Publ. Astronomical Soc. Pac.* 128, 104001. doi:10.1088/1538-3873/128/968/104001

Prod'homme, T., Verhoeve, P., Kohley, R., Short, A., and Boudin, N. (2014a). "A comparative study of charge transfer inefficiency value and trap parameter determination techniques making use of an irradiated ESA-Euclid prototype CCD," in *High energy, optical, and infrared detectors for astronomy VI*. Editors A. D. Holland, and J. Beletic (International Society for Optics and Photonics SPIE), 9154. doi:10.1117/12.2054862915409

Prod'homme, T., Verhoeve, P., Oosterbroek, T., Boudin, N., Short, A., and Kohley, R. (2014b). "Laboratory simulation of Euclid-like sky images to study the impact of CCD radiation damage on weak gravitational lensing," in *High energy, optical, and infrared detectors for astronomy VI*. Editors A. D. Holland, and J. Beletic (International Society for Optics and Photonics SPIE), 9154. doi:10.1117/12.2054870915414

Racca, G. D., Laureijs, R., Stagnaro, L., Salvignol, J.-C., Alvarez, J. L., Criado, G. S., et al. (2016). "The Euclid mission design," in *Space telescopes and instrumentation 2016: optical, infrared, and millimeter wave*. Editors H. A. MacEwen, G. G. Fazio, M. Lystrup, N. Batalha, N. Siegler, and E. C. Tong (International Society for Optics and Photonics SPIE), 9904, 235–257. doi:10.1117/12.2230762

Rees, W. G. (1987). The validity of the fresnel approximation. *Eur. J. Phys.* 8, 44.

Refregier, A. (2003). Shapelets — I. A method for image analysis. *Mon. Notices R. Astronomical Soc.* 338, 35–47. doi:10.1046/j.1365-8711.2003.05901.x

Refregier, A., and Bacon, D. (2003). Shapelets — II. A method for weak lensing measurements. *Mon. Notices R. Astronomical Soc.* 338, 48–56. doi:10.1046/j.1365-8711.2003.05902.x

Rhodes, J. D., Massey, R. J., Albert, J., Collins, N., Ellis, R. S., Heymans, C., et al. (2007). The stability of the point-spread function of the advanced camera for surveys on the hubble space telescope and implications for weak gravitational lensing. *Astrophysical J. Suppl. Ser.* 172, 203–218. doi:10.1086/516592

Rhodes, J., Leauthaud, A., Stoughton, C., Massey, R., Dawson, K., Kolbe, W., et al. (2010). The effects of charge transfer inefficiency (CTI) on galaxy shape measurements. *Publ. Astronomical Soc. Pac.* 122, 439–450. doi:10.1086/651675

Rigby, J., Perrin, M., McElwain, M., Kimble, R., Friedman, S., Lallo, M., et al. (2022). *The science performance of JWST as characterized in commissioning*. arXiv e-prints, arXiv:2207.05632.

Roddier, C., and Roddier, F. (1993). Wave-front reconstruction from defocused images and the testing of ground-based optical telescopes. *J. Opt. Soc. Am. A* 10, 2277–2287. doi:10.1364/JOSAA.10.002277

Roddier, F. (1981). *V the effects of atmospheric turbulence in optical astronomy*, 19. Elsevier, 281–376. of *Progress in Optics*. doi:10.1016/S0079-6638(08)70204-X

Roodman, A. (2010). "Focus and alignment using out-of-focus stellar images at the Dark Energy Camera," in *Ground-based and airborne instrumentation for astronomy III*. Editors I. S. McLean, S. K. Ramsay, and H. Takami (International Society for Optics and Photonics SPIE), 7735. 77353T. doi:10.1117/12.857680

Roodman, A. (2012). "Focus and alignment of the Dark Energy Camera using out-of-focus stars," in *Ground-based and airborne instrumentation for astronomy IV*. Editors I. S. McLean, S. K. Ramsay, and H. Takami (International Society for Optics and Photonics SPIE), 8446. 84466O. doi:10.1117/12.926671

Roodman, A., Reil, K., and Davis, C. (2014). "Wavefront sensing and the active optics system of the dark energy camera," in *Ground-based and airborne telescopes V*. Editors L. M. Stepp, R. Gilmozzi, and H. J. Hall (International Society for Optics and Photonics SPIE), 9145. doi:10.1117/12.2056904914516

Rowe, B. (2010). Improving psf modelling for weak gravitational lensing using new methods in model selection. *MNRAS* 404, 350–366.

Rowe, B., Hirata, C., and Rhodes, J. (2011). Optimal linear image combination. *Astrophysical J.* 741, 46. doi:10.1088/0004-637X/741/1/46

Rowe, B., Jarvis, M., Mandelbaum, R., Bernstein, G. M., Bosch, J., Simet, M., et al. (2015). Galsim: the modular galaxy image simulation toolkit. *Astronomy Comput.* 10, 121–150.

Sahu, K. C., Lallo, M., and Makidon, R. (2007). ACS PSF variations with temperatures. *Tech. Rep*.

Salmon, D., Cuillandre, J.-C., Barrick, G., Thomas, J., Ho, K., Matsushige, G., et al. (2009). Cfht image quality and the observing environment. *Publ. Astronomical Soc. Pac.* 121, 905–921.

Sánchez, C., Walter, C. W., Awan, H., Chiang, J., Daniel, S. F., Gawiser, E., et al. (2020). The LSST DESC data challenge 1: generation and analysis of synthetic images for next-generation surveys. *Monthly Notices of the Royal Astronomical Society* 497 (07), 210–228. doi:10.1093/mnras/staa1957

Sandin, C. (2014). The influence of diffuse scattered light - i. the psf and its role in observations of the edge-on galaxy ngc 5907. *A&A* 567, A97. doi:10.1051/0004-6361/201423429

Sasiela, R. J. (1994). "Electromagnetic wave propagation in turbulence: evaluation and application of mellin transforms," in *Springer series on wave phenomena*. 1 edn (Springer Berlin, Heidelberg).

Schmidt, J. D. (2010). *Numerical simulation of optical wave propagation with examples in MATLAB*. Bellingham, WA: SPIE Bellingham, WA.

Schmitz, M. A. (2019). *Euclid weak lensing: PSF field estimation*. Saclay, France: Ph.D. thesis, Université Paris-Saclay. 2019SACLS359.

Schmitz, M. A., Heitz, M., Bonneel, N., Ngole, F., Coeurjolly, D., Cuturi, M., et al. (2018). Wasserstein dictionary learning: optimal transport-based unsupervised nonlinear dictionary learning. *SIAM J. Imaging Sci.* 11, 643–678.

Schmitz, M. A., Starck, J. L., Mboula, F. N., Auricchio, N., Brinchmann, J., Capobianco, R. I. V., et al. (2020). Euclid: nonparametric point spread function field recovery through interpolation on a graph laplacian. *A&A* 636, A78. doi:10.1051/0004-6361/201936094

Schrabback, T., Hartlap, J., Joachimi, B., Kilbinger, M., Simon, P., Benabed, K., et al. (2010). Evidence of the accelerated expansion of the universe from weak lensing tomography with cosmos. *A&A* 516, A63. doi:10.1051/0004-6361/200913577

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. (2015). Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Process. Mag.* 32, 87–109. doi:10.1109/MSP.2014.2352673

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. (2015). Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Process. Mag.* 32, 87–109. doi:10.1109/MSP.2014.2352673

Sheldon, E. S., and Huff, E. M. (2017). Practical weak-lensing shear measurement with metacalibration. *ApJ* 841, 24.

Smith, R. M., Zavodny, M., Rahmer, G., and Bonati, M. (2008a). "A theory for image persistence in HgCdTe photodiodes," in *High energy, optical, and infrared detectors for astronomy III*. Editors D. A. Dorn, and A. D. Holland (International Society for Optics and Photonics SPIE), 7021. 70210J. doi:10.1117/12.789372

Smith, R. M., Zavodny, M., Rahmer, G., and Bonati, M. (2008b). "Calibration of image persistence in HgCdTe photodiodes," in *High energy, optical, and infrared detectors for astronomy III*. Editors D. A. Dorn, and A. D. Holland (International Society for Optics and Photonics SPIE), 7021. 70210K. doi:10.1117/12.789619

Snyder, A., Srinath, S., Macintosh, B., and Roodman, A. (2016). "Temporal characterization of Zernike decomposition of atmospheric turbulence," in *Ground-based and airborne telescopes VI*. Editors H. J. Hall, R. Gilmozzi, and H. K. Marshall (International Society for Optics and Photonics SPIE), 9906. doi:10.1117/12.2234362990642

Soulez, F., Courbin, F., and Unser, M. (2016). "Back-propagating the light of field stars to probe telescope mirrors aberrations," in *Proc. SPIE 9912, advances in optical and mechanical technologies for telescopes and instrumentation II*, 991277. (July 22, 2016). doi:10.1117/12.2232702

Soulez, F., Denis, L., Tourneur, Y., and Thiébaut, É. (2012). "Blind deconvolution of 3d data in wide field fluorescence microscopy," in *2012 9th IEEE international symposium on biomedical imaging (ISBI)*. 1735–1738.

Southwell, W. H. (1981). Validity of the fresnel approximation in the near field. *J. Opt. Soc. Am.* 71, 7–14. doi:10.1364/JOSA.71.000007

Spergel, D., Gehrels, N., Baltay, C., Bennett, D., Breckinridge, J., Donahue, M., et al. (2015). Wide-Field InfrarRed survey telescope-astrophysics focused telescope assets WFIRST-AFTA 2015 report. *Tech. Rep*.

Starck, J.-L., Murtagh, F., and Fadili, J. (2015). *Sparse image and signal processing: wavelets and related geometric multiscale analysis*. Second Edition (Cambridge: Cambridge University Press). doi:10.1017/CBO9781316104514

Storkey, A. J., Hambly, N. C., Williams, C. K. I., and Mann, R. G. (2004). Cleaning sky survey data bases using Hough transform and renewal string approaches. *Mon. Notices R. Astronomical Soc.* 347, 36–51. doi:10.1111/j.1365-2966.2004.07211.x

Stubbs, C. W. (2014). Precision astronomy with imperfect fully depleted ccds — An introduction and a suggested lexicon. *J. Instrum.* 9, C03032. doi:10.1088/1748-0221/9/03/C03032

Suchkov, A., and Casertano, S. (1997). Impact of focus drift on aperture photometry. *Tech. Rep*.

Thomas, S. J., Chandrasekharan, S., Lotz, P., Xin, B., Claver, C., Angeli, G., et al. (2016). "LSST active optics system software architecture," in *Ground-based and airborne telescopes VI*. Editors H. J. Hall, R. Gilmozzi, and H. K. Marshall (Society of Photo-Optical Instrumentation Engineers SPIE), 9906. 99063B. doi:10.1117/12.2231798

Troxel, M. A., Lin, C., Park, A., Hirata, C., Mandelbaum, R., Jarvis, M., et al. (2023). A joint roman space telescope and rubin observatory synthetic wide-field imaging survey. *Mon. Notices R. Astronomical Soc.* doi:10.1093/mnras/stad664.Stad664

Venancio, L. M. G., Carminati, L., Alvarez, J. L., Amiaux, J., Bonino, L., Salvignol, J.-C., et al. (2016). "Coating induced phase shift and impact on Euclid imaging performance," in *Space telescopes and instrumentation 2016: optical, infrared, and millimeter wave*, Editors. H. A. MacEwen, G. G. Fazio, M. Lystrup, N. Batalha, N. Siegler, and E. C. Tong (International Society for Optics and Photonics SPIE), 9904,325–330.

Venancio, L. M. G., Carminati, L., Amiaux, J., Bonino, L., Racca, G., Vavrek, R., et al. (2020). "Status of the performance of the Euclid spacecraft," in *Space telescopes and instrumentation 2020: optical, infrared, and millimeter wave*. Editors M. Lystrup, M. D. Perrin, N. Batalha, N. Siegler, and E. C. Tong (International Society for Optics and Photonics SPIE), 11443, 45–60. doi:10.1117/12.2562490

Viola, M., Melchior, P., and Bartelmann, M. (2011). Biases in, and corrections to, KSB shear measurements. *Mon. Notices R. Astronomical Soc.* 410, 2156–2166. doi:10.1111/j.1365-2966.2010.17589.x

Walter, C. W. (2015). The Brighter-Fatter and other sensor effects in CCD simulations for precision astronomy. *J. Instrumentation* 10 (05), C05015. doi:10.1088/1748-0221/10/05/C05015

Wong, A., Pope, B., Desdoigts, L., Tuthill, P., Norris, B., and Betters, C. (2021). Phase retrieval and design with automatic differentiation. *J. Opt. Soc. Am. B* 38, 2465. doi:10.1364/josab.432723

Wright, A. H., Hildebrandt, H., Kuijken, K., Erben, T., Blake, R., Buddelmeijer, H., et al. (2019). Kids+viking-450: a new combined optical and near-infrared dataset for cosmology and astrophysics. *A&A* 632, A34. doi:10.1051/0004-6361/201834879

Xin, B., Claver, C., Liang, M., Chandrasekharan, S., Angeli, G., and Shipsey, I. (2015). Curvature wavefront sensing for the large synoptic survey telescope. *Appl. Opt.* 54, 9045–9054. doi:10.1364/AO.54.009045

Xin, B., Roodman, A., Angeli, G., Claver, C., and Thomas, S. (2016). "Comparison of LSST and DECam wavefront recovery algorithms," in *Ground-based and airborne telescopes VI*. Editors H. J. Hall, R. Gilmozzi, and H. K. Marshall (International Society for Optics and Photonics SPIE), 9906. 99064J. doi:10.1117/12.2234456

Xin, B., Ivezić, Ž., Lupton, R. H., Peterson, J. R., Yoachim, P., Jones, R. L., et al. (2018). A study of the point-spread function in sdss images. *Astronomical J.* 156, 222. doi:10.3847/1538-3881/aae316

Yamamoto, M., Laliotis, K., Macbeth, E., Zhang, T., Hirata, C. M., Troxel, M. A., et al. (2023). *Simulating image coaddition with the nancy grace roman space telescope: II. Analysis of the simulated images and implications for weak lensing*. arXiv e-prints, arXiv:2303.08750. doi:10.48550/arXiv.2303.08750

York, D. G., Adelman, J., JohnAnderson, E. J., Anderson, S. F., Annis, J., Bahcall, N. A., et al. (2000). The sloan digital sky survey: technical summary. *Astronomical J.* 120, 1579. doi:10.1086/301513

Yun, G., Crabtree, K., and Chipman, R. A. (2011). Three-dimensional polarization ray-tracing calculus i: definition and diattenuation. *Appl. Opt.* 50, 2855–2865. doi:10.1364/AO.50.002855

Zhang, J. (2007). Measuring the cosmic shear in Fourier space. *Mon. Notices R. Astronomical Soc.* 383, 113–118. doi:10.1111/j.1365-2966.2007.12585.x

Zhang, J. (2011). Measuring the reduced shear. *J. Cosmol. Astropart. Phys.* 2011, 041. doi:10.1088/1475-7516/2011/11/041

Zhang, J., Luo, W., and Foucaud, S. (2015). Accurate shear measurement with faint sources. *J. Cosmol. Astropart. Phys.* 2015, 024. doi:10.1088/1475-7516/2015/01/024

Zhang, J., Dong, F., Li, H., Li, X., Li, Y., Liu, D., et al. (2019). Testing shear recovery with field distortion. *Astrophysical J.* 875, 48. doi:10.3847/1538-4357/ab1080

Zhang, T., Mandelbaum, R., and Collaboration, T. L. D. E. S. (2021). Impact of point spread function higher moments error on weak gravitational lensing. *Mon. Notices R. Astronomical Soc.* 510, 1978–1993. doi:10.1093/mnras/stab3584

Zhang, T., Almoubayyed, H., Mandelbaum, R., Meyers, J. E., Jarvis, M., Kannawadi, A., et al. (2022). *Impact of point spread function higher moments error on weak gravitational lensing ii: a comprehensive study*. doi:10.48550/ARXIV.2205.07892

Zhuang, M.-Y., and Shen, Y. (2023). *Characterization of JWST NIRCam PSFs and implications for AGN+Host image decomposition*. arXiv e-prints, arXiv:2304.13776. doi:10.48550/arXiv.2304.13776

Zuntz, J., Sheldon, E., Samuroff, S., Troxel, M. A., Jarvis, M., MacCrann, N., et al. (2018). Dark energy survey year 1 results: weak lensing shape catalogues. *MNRAS* 481, 1149–1182.

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Original Research
PUBLISHED 18 December 2023
DOI 10.3389/fspas.2023.1322980

Check for updates

OPEN ACCESS

EDITED BY
Xueshang Feng,
Chinese Academy of Sciences (CAS), China

REVIEWED BY
Paolo Ventura,
Astronomical Observatory of Rome (INAF), Italy
Ding Yuan,
Harbin Institute of Technology, China

*CORRESPONDENCE
A. Ali,
afouad@sci.cu.edu.eg

RECEIVED 17 October 2023
ACCEPTED 27 November 2023
PUBLISHED 18 December 2023

CITATION
Mohery M, Ali A, Khames AA, Snaid S and Mindil A (2023), Exploring the high abundance discrepancy in the planetary nebula IC 4663.
Front. Astron. Space Sci. 10:1322980.
doi: 10.3389/fspas.2023.1322980

COPYRIGHT
© 2023 Mohery, Ali, Khames, Snaid and Mindil. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Exploring the high abundance discrepancy in the planetary nebula IC 4663

M. Mohery[1], A. Ali[2]*, Aya A. Khames[2], S. Snaid[3] and A. Mindil[1]

[1]Department of Physics, College of Science, University of Jeddah, Jeddah, Saudi Arabia, [2]Astronomy, Space Science and Meteorology Department, Faculty of Science, Cairo University, Giza, Egypt, [3]Department of Astronomy and Space Science, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

This article presents an in-depth analysis of integral field unit spectroscopy performed on the Galactic planetary nebula (PN) IC 4663, with a primary focus on investigating its physical characteristics, chemical composition, and morphological properties. The examination involved the utilization of a series of emission-line maps representing various elements and ionization zones, allowing for a comprehensive assessment of morphological and ionization structure of IC 4663. The majority of these maps depict an overall elliptical shape, featuring a relatively faint core at the center. Emanating from this core are two distinct and illuminated lobes extending in opposite directions, forming a conspicuous double-lobed configuration. The chemical analysis conducted on IC 4663 has revealed that the PN is rich in both helium and nitrogen, characterized by an N/O ratio exceeding 0.5, classifying it as a Peimbert type I nebula. Furthermore, the presence of a notably intense He II emission line at λ4686, in conjunction with the existence of high-excitation lines within the nebular spectrum, indicate its classification as a high-excitation class nebula. The computation of O^{2+}, N^{2+}, and N^{3+} through optical recombination lines has unveiled extreme abundance discrepancy factors, with values of 39.0 ± 6 for O^{2+} and 37.0 ± 9 for N^{2+} and N^{3+}. These findings suggest the possibility that the central star of IC 4663 may be part of a binary system with a sub-stellar companion, potentially engaged in a common envelope interaction. Recent astrometric data derived from the Gaia mission further indicate a significant likelihood of interaction between the asymptotic giant branch halo of the PN and its surrounding interstellar medium, particularly in the southeast direction.

KEYWORDS

Planetary nebulae: individual IC 4663 - Properties: physical, chemical, kinematic

1 Introduction

Following the depletion of hydrogen and helium as their nuclear fuel in the cores, low- and intermediate-mass stars (1–8 $M_{\odot}$) progress into the asymptotic giant branch (AGB) phase. Towards the tip of the AGB, these stars experience stellar winds, which cause them to shed a considerable amount of their outer layers. This mass loss can occur in the form of a slow, dense wind or more episodic events like stellar pulsations and thermal pulses. The mass loss is a crucial factor in shaping the subsequent evolution of the star and its surrounding environment. When the central star (CS) reaches a sufficiently high temperature

($T_{eff} \geq 25,000$ K), the surrounding circumstellar envelope becomes ionized, resulting in the formation of a planetary nebula (PN, plural PNe). At different spatial scales, PNe exhibits a variety of shapes and structural components, ranging from small-scale low-ionization structures (LISs), e.g., knots, filaments, and jets to large-scale structures, e.g., rims, shells, and haloes (Balick, 1987; Corradi et al., 1996; Balick et al., 1998; Gonçalves et al., 2001). Various theories have been proposed to explain the formation of LISs within PNe, but none of them can fully account for all the distinct forms (Akras and Gonçalves, 2016, & references therein).

The classical model of interacting stellar winds (ISW), proposed by Kwok et al. (1978), successfully explains the symmetrical and round morphology of PNe through the interaction of fast and tenuous post-AGB wind with slow and dense AGB wind. However, it does not provide a comprehensive explanation for the asymmetric, elliptical, and bipolar shapes observed in PNe. These complex shapes suggest the involvement of additional factors influencing the formation and shaping of PNe, including magnetic fields and CS rotation and binarity (Balick and Frank, 2002, & references therein). Extensive research on the interaction between slow and fast winds has demonstrated that non-spherical mass loss during the AGB phase can account for many of the observed PN morphologies and kinematics (Mellema and Frank, 1995, & references therein). Mellema (1996) has offered a concise overview of hydrodynamic modeling of PNe, highlighting the specific observations that these models aim to replicate and presenting the outcomes of various modeling approaches. Furthermore, García-Segura et al. (2022) have conducted a study utilizing 2D hydrodynamical simulations to explore the evolution from a protoplanetary (PPN) phase to a PN in binary systems that have undergone a common-envelope (CE) event. Their findings suggest that these models have the potential to explain a significant portion of PN shapes, including the elliptical and bipolar morphological classes.

The study of chemical enrichment in the universe through PNe offers valuable insights into several key aspects of astrophysics. It allows us to gain a deeper understanding of stellar nucleosynthesis, the origins and distribution of elements within the cosmos, and the tracing of galactic chemical evolution by examining the abundances of various elements across different regions of a galaxy (Faúndez-Abans and Maciel, 1986; Costa et al., 2004; Maciel et al., 2006). These gradients in elemental abundance provide crucial clues regarding the processes of galaxy formation and evolution, as well as the mechanisms responsible for mixing and transporting enriched materials. Furthermore, the study of PNe in distant galaxies provides unique insights into the chemical conditions of the early universe, shedding light on the composition and development of the cosmos during its formative stages (Magrini et al., 2012; Corradi et al., 2015).

In the framework of studying a sample of PNe located towards the bulge of our galaxy, Cavichia et al. (2010) provide a long-slit spectrum for IC 4663. They derived the physical properties and elemental abundances of IC 4663. Following that, a comprehensive analysis of IC 4663 and its associated CS, categorized as having the spectral type [WN3], was conducted by Miszalski et al. (2012). They employed a CMFGEN NLTE model atmosphere to characterize the CS, revealing a very high temperature of T = 140 kK and a fast stellar wind with $v_{\infty} = 1900$ km s^{-1}. To investigate the physical and chemical characteristics of the ionized nebula, they combined optical and UV spectra from archival low-dispersion *IUE* spectra (SWP33946 and LWP13707). The results displayed a chemical abundance pattern relatively similar to solar values, where the blue portion ($\lambda \leq 5,900$ Å) came from the Gemini Multi-Object Spectrograph (GMOS) spectrum, and the red portion was obtained from the work by Cavichia et al. (2010). Additionally, the shape of IC 4663 was explored using three images captured during the GMOS program, employing Hα and [O III] filters with central wavelengths/FWHMs of 656.0/7.2 nm, 499.0/4.5 nm, and 514.0/8.8 nm, respectively. These images revealed an ellipsoidal bubble surrounded by a fainter outer shell, along with the presence of an external AGB halo expanding at a relatively low velocity of 30 km s^{-1}. Notably, the [O III] image indicated a flux enhancement in the AGB halo towards the southeast (SE) of the CS, suggestive of potential interaction with the interstellar medium (ISM). Furthermore, Puspitaningrum and Malasan (2019) developed a photoionization model for IC 4663 using CLOUDY and observational data spanning from optical to far-infrared wavelengths. They determined the properties of the CS by employing the PoWR atmosphere model, which yielded values of $T_{eff} = 125$ kK and $\log g = 6.0$, aligning with the observed data. By comparing the nebular abundances obtained from the model with those predicted by an AGB model and a born-again scenario evolutionary track, they inferred that IC 4663 originated from a progenitor star with a mass of 3 $M_{\odot}$ and Z = 0.02.

Integral Field Unit (IFU) spectroscopy represents a cutting-edge technique in the examination of extended objects such as PNe, H II regions, and galaxies, offering distinct advantages over traditional long-slit spectroscopy methods. This approach allows researchers to study these extended objects in a 2D field of view, providing a comprehensive understanding of their properties. Establishing the correlation between spectra and spatial structures is vital for accurately measuring the physical conditions and understanding the distribution of different ions within PNe. Emission-line and line ratio maps generated through IFU spectroscopy of PNe contribute valuable insights into their ionization structure, morphology, and the underlying physical processes. Previous work by Ali et al. (2016) has highlighted the advantages of IFU spectroscopy in the field of PNe compared to long-slit spectroscopy. Over the last 2 decades, various authors have delved into IFU spectra of Galactic PNe. Notably, our research group has published a series of papers discussing IFU spectroscopy of PNe utilizing the Wide Field Spectrograph (WiFeS) instrument, with examples including Ali et al. (2015), Basurah et al. (2016), Dopita et al. (2017), Ali and Dopita (2019), and Snaid et al. (2023). A review of the existing literature reveals a substantial body of IFU studies conducted on numerous PNe, as demonstrated by the works of Monreal-Ibero et al. (2005), Tsamis et al. (2007), Danehkar et al. (2013), García-Rojas et al. (2022) and Mari et al. (2023).

Systematic investigations of ionized nebulae, such as H II regions and PNe, have revealed that the abundance of heavy element ions derived from optical recombination lines (ORLs) is higher compared to those deduced from collisionally excited lines (CELs) of identical species. The ratio of ORL to CEL abundances is referred to as the abundance discrepancy factor (ADF). In the majority of PNe, the

ADF typically ranges from 2 to 3. However, a small number of Pne exhibit ADF values that are one or two orders of magnitude higher. Wesson et al. (2018) categorize these variances, describing an ADF value below 5 as "normal," values spanning between 5 and 10 as "elevated," and those exceeding 10 as "extreme."

The major goal of this research is to conduct an extensive analysis of IC 4663 through spectroscopic, morphological, and kinematic investigations using IFU spectra. Furthermore, this research investigates the abundance discrepancy observed within PN, along with assessing the probability of the CS existing within a closely orbiting binary system. The article is structured as follows: Section 2 provides a discussion on the observations and data reduction techniques employed. Section 3 presents the morphology and ionization structures observed in the object. Section 4 provides an investigation into the plasma diagnostics, physical conditions, and chemical composition of the PN. Section 5 discusses the probable close binarity of the PN CS. The interaction between the object and the ISM is examined in Section 6. Finally, the conclusion drawn from the research are summarized in the last section.

2 Observations and data reduction

On 17 May 2018, observations of IC 4663 were carried out using the WiFeS instrument (Dopita et al., 2007) mounted on the 2.3-m telescope of the Australian National University (ANU) at Siding Spring Observatory. WiFeS is a robust integral field, dual-beam, concentric, and image-slicing spectrograph that provides a field of view measuring $25'' \times 38''$ with a spatial resolution of 1.0×1.0. The data were collected in low-resolution mode (R~3000), which corresponds to a velocity resolution of 100 km/s. During these observations, the CCD was set up in double binning mode, producing a $4,096 \times 2048$ data format with $1.0''$ pixels in the spatial direction. Observations were conducted simultaneously in two gratings. The B3000 and R3000 gratings used the RT560 dichroic that cuts at 5600 Å. This setup ensures a significant overlap in wavelength coverage between each of the gratings, resulting in continuous wavelength coverage from ~3400 to ~8950Å. The WiFeS observations were summarized in Table 1. The PyWiFeS pipeline (Childress et al., 2014) was used for data reduction. The wavelength scale was calibrated using nighttime observations of the Ne-Ar arc lamp, whereas the nebular fluxes were calibrated using the standard stars HD 111980 and HD 031128.

TABLE 1 The observing log.

Object	No. of frames	Exposure time (s)	Date	Airmass
IC 4663 (PN G346.2-08.2)				
B3000	3	10, 100, 800	05/15/2018	1.19, 1.19, 1.17
R3000	3	10, 100, 800	05/15/2018	1.19, 1.19, 1.17

3 Emission-line maps

In order to investigate the morphology and ionization structure of the PN, we generated a collection of emission-line maps that represent various elements and ionization states. The majority of these maps exhibit an overall elliptical shape, with a relatively faint core positioned in the central region. Emerging from this core, two distinct and prominently illuminated lobes extend in opposite directions, forming a clear double-lobed structure, and are enveloped by a relatively smooth and extended gas shell. Figure 1 displays three images of the PN, highlighting emission lines of Hα (left panel), He II (middle panel), and [O III] (right panel). Notably, the brightness of the pair of lobes varies across these images. The [O III] map shows brighter lobes compared to the Hα map, which, in turn, is brighter than the He II map. This variation in brightness is attributed to differences in the line fluxes of the three ions. Furthermore, the south lobe in the He II map appears brighter than the north one.

In Figure 2, we have constructed three emission-line maps: [Ar III] (left panel), [Ar IV] (middle panel), and [Ar V] (right panel), with the aim of examining the ionization stratification of the nebula. As expected, the lower ionization line map displays a larger nebular size. The order of the panels, from left to right, corresponds to displaying the argon line maps in ascending order of ionization levels, starting with [Ar III] and progressing to [Ar V]. Notably, the figure reveals that in the emission line of [Ar III], the north lobe appears brighter than the south lobe. However, in the emission lines of [Ar IV] and [Ar V], the north lobe appears relatively dimmer in comparison to the south lobe. This observed variation in brightness across different ionization levels shows the complex ionization structure of the nebula.

Figure 3 displays the nebula in ORLs of nitrogen and oxygen: N II (left panel), O II (middle panel), and O III (right panel). The ORLs lines on nitrogen and oxygen appear less distinct than their corresponding CELs due to their lower fluxes.

4 Physical and chemical analysis

4.1 Global spectra and line fluxes

The software QFitsView v3.3[1] was used to extract the complete spectrum of the PN from the data cubes. A circular aperture of radius $12''$ was used to match the apparent size of the PN. The spectra from the B and R gratings were combined using the scombine task of the IRAF software. The global spectrum of IC 4663 is displayed in Figure 4.

The ALFA code (Wesson, 2016) facilitated the identification and flux measurement of emission lines across the entire spectrum. To calculate the interstellar reddening coefficient, $c(H\beta)$, an iterative method was employed based on the average ratios of the Hydrogen Balmer lines Hα, Hβ, Hγ, and Hδ. This calculation assumed Case B conditions at $T_{eff} = 10,000$ K and relied on the

1 A FITS file viewer using the QT widget library developed at the Max Planck Institute for Extraterrestrial Physics by Thomas Ott.

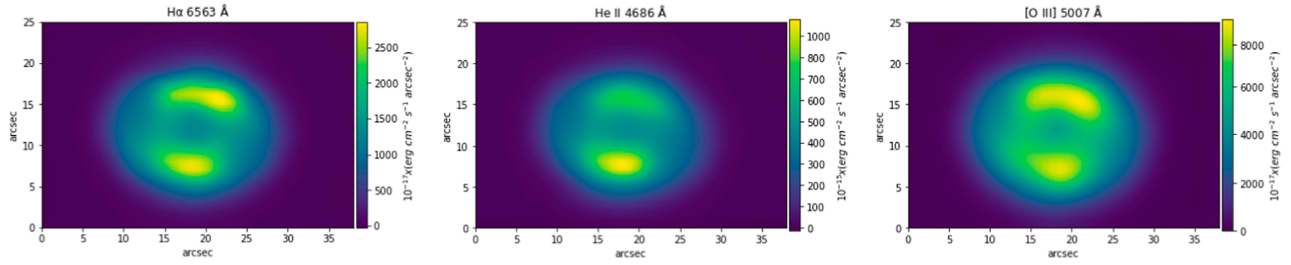


FIGURE 1
The emission-line maps of IC 4663 in three different ions: Hα at 6563 Å, He II at 4686 Å, and [O III] at 5007 Å. The elliptical morphology is seen in all maps. In all figures the North is to the top of the image and East is to the left. The colour bar displays the surface brightness variation of the nebula.

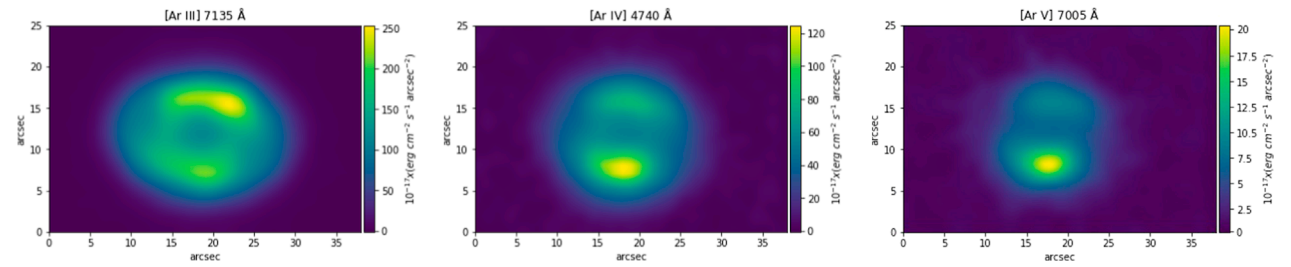


FIGURE 2
The purpose of the three emission-line maps of IC 4663 is to emphasize the ionization stratification of the nebula using different argon ions. The left, middle, and right panels highlight [Ar III] at 7135 Å, [Ar IV] at 4740 Å, and [Ar V] at 7005 Å, respectively. It is apparent that the size of the nebula decreases as we transition from the lower ionization represented by the [Ar III] emission-line map to the higher ionization indicated by the [Ar IV] emission-line map.

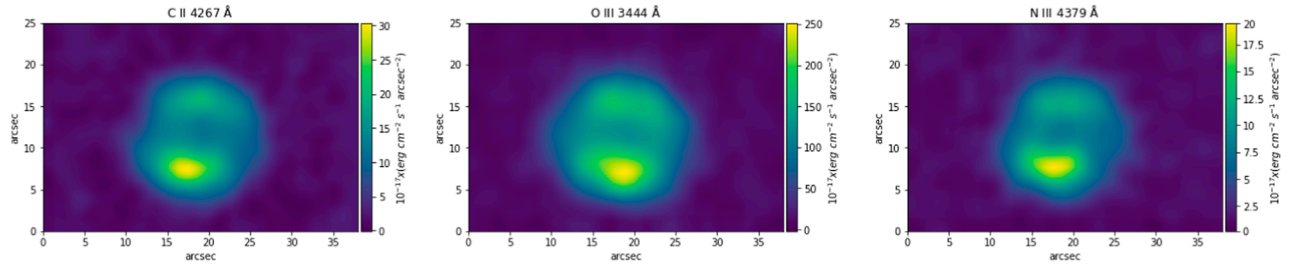


FIGURE 3
The emission-line maps of IC 4663 in three ORLs of C II at 4267 Å (left panel), O III at 3444 Å (middle panel), and N III at 4379 Å (right panel).

Nebular Empirical Abundance Tool [NEAT, Wesson et al. (2012)] The observed line fluxes in the given context were adjusted for interstellar reddening using the extinction law proposed by Howarth (1983). Table 2 provides values for both the observed line fluxes, $F(\lambda)$, and the de-reddened line fluxes, $I(\lambda)$, with respect to Hβ being set at 100. To address statistical uncertainties arising from line flux measurements, NEAT incorporated the Monte Carlo technique, which ensured accurate consideration of these uncertainties throughout subsequent steps of plasma diagnosis, including the determination of physical conditions and chemical abundances.

The line list presented in Table 2 demonstrates that IC 4663 exhibits a significant array of ORLs. Surprisingly, the flux measurements of these lines are observed to be substantially higher compared to analogous lines commonly observed in the majority of PNe. To illustrate this, the corrected fluxes for particular lines, such as O III lines at λ3444.1, N II at λ4026.1, O II at λ4641.8, C II at λ4267.2, and N III at λ4640.6, yield values of 33.3, 2.5, 8.5, 3.0, and 8.7, respectively, relative to F(Hβ) = 100. This finding serves as an indicator of the high ADF of this PN.

Additionally, the spectral analysis indicates that the intensities of low-ionization emission lines, including as [O I], [N I], [O II], and

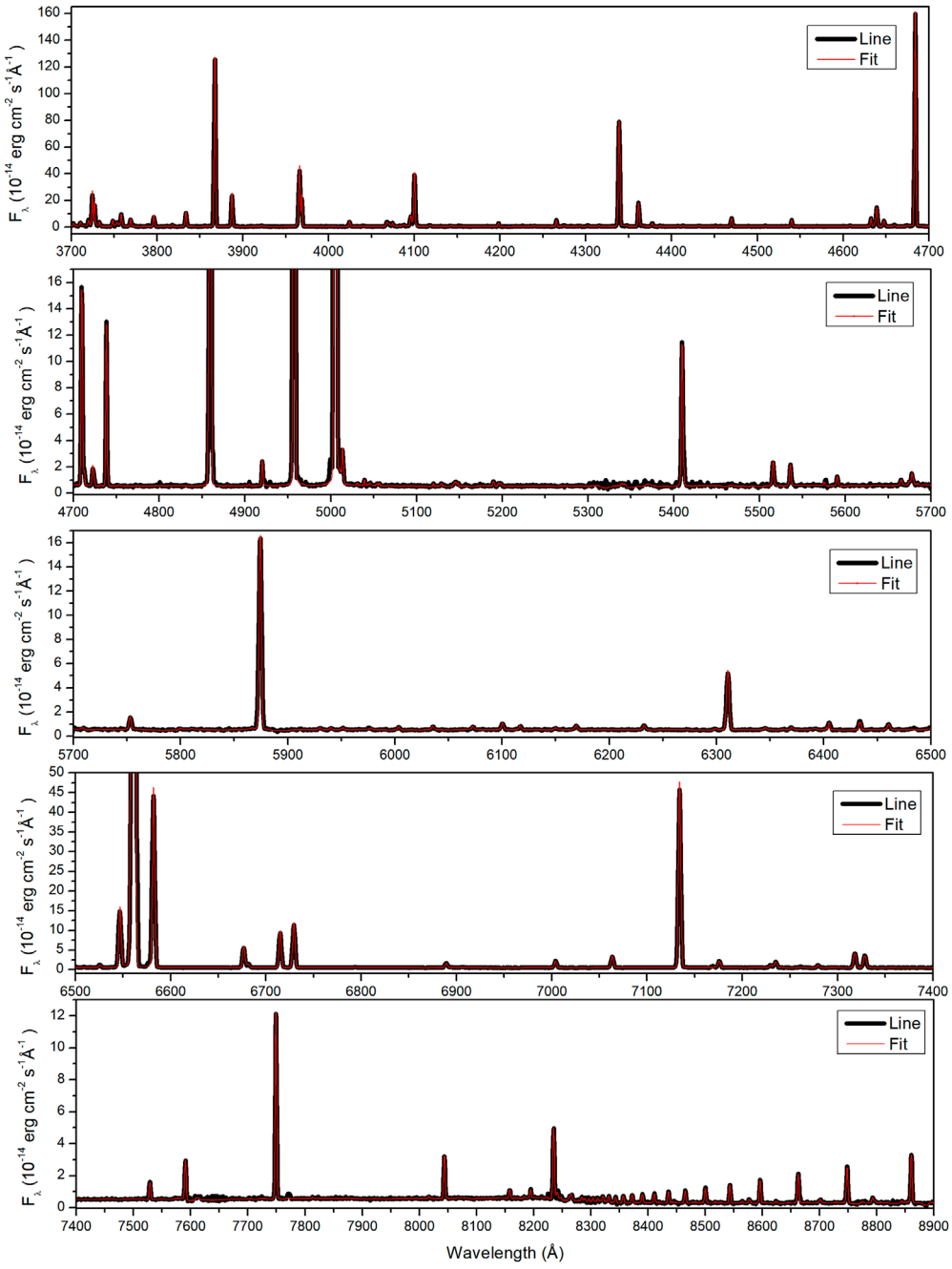


FIGURE 4
The overall spectrum of PN IC 4663. The observed spectrum was represented by the bold black line, while the fitted spectrum obtained using the ALFA code is represented by the thin red line.

[N II], are weak. Particularly, the flux ratio of [N II] λ6583 to Hα is below 0.5. This observation implies that the nebula is nearly optically thin.

The presence of the He II line at λ4686 in the nebular spectrum suggests that the nebula falls into the medium to high excitation class, and its CS has an effective temperature greater than 85 kK. By applying the procedure outlined by Reid and Parker (2010), a high EC value of 10.8 was derived. This value aligns with the presence of highly excited lines like [Ar V] and [Ne IV], which possess atomic ionization energies of 75.0 and 97.1 eV, respectively.

The systemic velocity RV_{sys} of IC 4663 was measured using the ALFA code, primarily from strong emission lines. The heliocentric radial velocity RV_{hel} was then calculated by correcting RV_{sys} for the effect of Earth's motion. The result is -80.0 ± 22.0 km/s, which is consistent with the value reported by Beaulieu et al. (1999). Table 3 provides a comparison, highlighting the differences between the values of $c(H\beta)$, Hα and Hβ fluxes, EC, and RV_{hel}

for IC 4663 and the corresponding values reported in the existing literature.

4.2 Temperatures and densities of IC 4663

The detected emission lines (CELs and ORLs) in the PN spectrum were utilized to determine electron densities and temperatures at various ionization levels. Nebular densities were calculated by employing line ratios: [O II] (λ3727/λ3729), [S II] (λ6716/λ6731), [Cl III] (λ5517/λ5537), and [Ar IV] (λ4711/λ4740). On the other hand, nebular temperatures were calculated using line ratios: [O III] (λ4959 + λ5007)/λ4363, [Ar III] (λ7135 + λ7751)/λ5192, [N II] (λ6548 + λ6584)/λ5754, [O II] (λ7319 + λ7330)/(λ3726 + λ3729), [S II] (λ6717 + λ6731)/(λ4068 + λ4076), and He I (λ7281/λ6678).

According to García-Rojas et al. (2022), in certain highly-excited PNe with high ADFs, the auroral [O III] λ4363 line may be significantly excited by recombination processes, which makes the T_e [O III] diagnostic unreliable (Gómez-Llanos et al., 2020). Additionally, they found that the recombination contribution to the auroral [N II] λ5755 line can strongly impacts the measured line flux in PNe with high ADF. To address these complexities, the analysis incorporated the adjustment of CELs by considering the contribution of ORLs using NEAT, which offers an option to subtract the effect of recombination contribution from key diagnostic CELs.

TABLE 2 Integrated $F(\lambda)$ and de-reddened $I(\lambda)$ line fluxes (relative to Hβ = 100) for IC4663. The complete table is available as **Supplementary Material**.

$\lambda_{Lab.}$(Å)	$\lambda_{Obs.}$(Å)	Ion	$F(\lambda)$	$I(\lambda)$
4711.37	4709.66	[Ar IV]	7.24 ± 0.108	$7.60^{+0.115}_{-0.115}$
4724.15	4722.43	[Ne IV]	1.00 ± 0.020	$1.03^{+0.021}_{-0.021}$
4725.62	4723.90	[Ne IV]	0.36 ± 0.020	$0.40^{+0.021}_{-0.021}$
4740.17	4738.45	[Ar IV]	5.73 ± 0.078	$6.11^{+0.082}_{-0.082}$
4859.32	4857.53	He II	4.17 ± 0.840	$4.62^{+0.847}_{-0.836}$
4861.33	4859.54	H I	100 ± 0.840	$100^{+0.845}_{-0.845}$
4921.93	4920.12	He I	0.83 ± 0.065	$0.92^{+0.064}_{-0.064}$
4958.91	4957.08	[O III]	367 ± 3.35	$354.0^{+3.00}_{-3.00}$
5006.84	5005.02	[O III]	1088 ± 6.98	$1030^{+10.00}_{-10.00}$
5015.68	5013.85	He I	1.36 ± 0.058	$1.24^{+0.055}_{-0.055}$

TABLE 3 The physical parameters of IC 4663, which are derived from its global spectrum.

Parameters	Ours	Literature
log F(Hβ)	−11.42	−11.47 ± 0.03 (1)
C(Hβ)	$0.54^{+0.031}_{-0.031}$	0.58 (2), 0.51 (3)
RV_{hel} (km/s)	−79.7 ± 21.9	−48.7 ± 6 (4), −52.1 ± 6.2 (5), −74 (6)
EC	10.8	7.0 (6)
Peimbert class	Type I	
Ionic density (N_e) in cm^{-3}		
[O II]	2410^{+570}_{-810}	
[S II]	1800^{+280}_{-340}	3810 (2), 1000 (3)
[Cl III]	712^{+325}_{-351}	
[Ar IV]	844^{+178}_{-183}	
Ionic temperature (T_e) in K		
[O II]	14900^{+1600}_{-1800}	12800 (2), 10300 (3)
[S II]	9370^{+1040}_{-1170}	
[N II]	12300^{+400}_{-400}	
[O III]	11400^{+100}_{-100}	11500 (2), 11500 (3)
[Ar III]	10700^{+300}_{-300}	
He I (λ7281/λ6678)	12300^{+900}_{-900}	11400 (3)

(1) Cahn et al. (1992), (2) Cavichia et al. (2010), (3) Miszalski et al. (2012), (4) Schneider and Terzian (1983), (5) Durand et al. (1998), (6) Beaulieu et al. (1999).

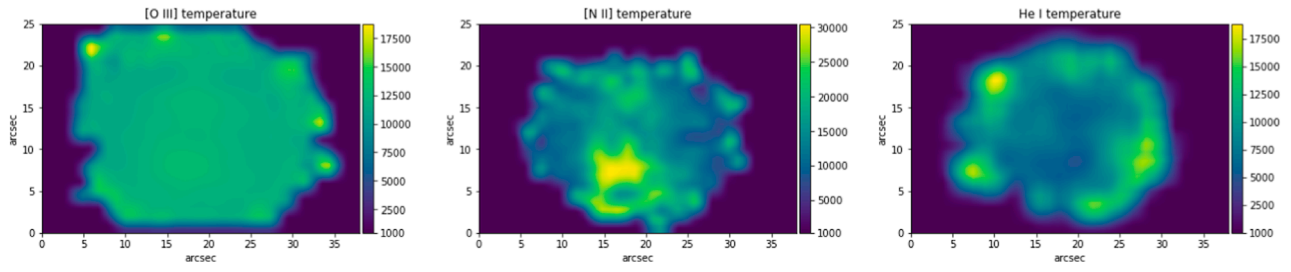


FIGURE 5
The electron temperature of IC 4663 was determined using three different diagnostic line ratios. The temperature map derived from the [O III] diagnostic line ratio is displayed in the left panel, the [N II] line ratio in the middle panel, and the He I line ratio in the right panel.

TABLE 4 Ionic and total abundances of IC 4663.

Ions	IC 4663
CEL abundances	
N^+/H	$(1.33^{+0.12}_{-1.04}) \times 10^{-6}$
ICF(N)	$(195.00^{+116.00}_{-42.00})$
N/H	$(2.55^{+0.66}_{-0.46}) \times 10^{-4}$
O^0/H	$(2.87^{+2.87}_{-2.80}) \times 10^{-9}$
O^+/H	$(1.77^{+0.25}_{-2.80}) \times 10^{-6}$
O^{2+}/H	$(2.86^{+0.27}_{-0.31}) \times 10^{-4}$
ICF(O)	$(1.56^{+0.02}_{-0.02})$
O/H	$(4.51^{+0.41}_{-0.48}) \times 10^{-4}$
Ne^{2+}/H	$(7.83^{+0.88}_{-1.04}) \times 10^{-5}$
Ne^{4+}/H	$(1.29^{+0.24}_{-0.27}) \times 10^{-5}$
ICF(Ne)	$(1.38^{+0.01}_{-0.01})$
Ne/H	$(1.26^{+0.14}_{-0.17}) \times 10^{-4}$
Ar^{2+}/H	$(1.82^{+0.15}_{-0.17}) \times 10^{-6}$
Ar^{3+}/H	$(1.60^{+0.15}_{-0.18}) \times 10^{-6}$
Ar^{4+}/H	$(1.52^{+0.14}_{-0.16}) \times 10^{-7}$
ICF(Ar)	$(1.38^{+0.04}_{-0.03})$
Ar/H	$(4.93^{+0.42}_{-0.47}) \times 10^{-6}$
S^+/H	$(9.16^{+0.91}_{-5.52}) \times 10^{-8}$
ICF(S)	$(289.00^{+172.00}_{-63.00})$
S/H	$(2.61^{+0.84}_{-0.44}) \times 10^{-5}$
Cl^+/H	$(2.44^{+0.19}_{-1.36}) \times 10^{-9}$
Cl^{2+}/H	$(9.72^{+0.90}_{-1.03}) \times 10^{-8}$
Cl^{3+}/H	$(7.83^{+0.66}_{-0.74}) \times 10^{-8}$
ICF(Cl)	$(1.00^{+0.00}_{-0.00})$
Cl/H	$(1.78^{+0.15}_{-0.17}) \times 10^{-7}$
RL abundances	
He^+/H	$(6.90^{+0.30}_{-0.30}) \times 10^{-2}$
He^{2+}/H	$(6.59^{+0.04}_{-0.04}) \times 10^{-2}$
He/H	$(1.35^{+0.03}_{-0.03}) \times 10^{-1}$
C^{2+}/H	$(2.74^{+0.07}_{-0.07}) \times 10^{-3}$
C^{3+}/H	$(4.97^{+0.67}_{-0.66}) \times 10^{-4}$
ICF(C)	$(1.62^{+0.04}_{-0.04})$
C/H	$(5.25^{+0.16}_{-0.16}) \times 10^{-3}$

(Continued on the following page)

Table 3 presents the electron densities and temperatures of IC 4663, comparing them to values documented in previous studies. The findings indicate that the density determined from the [S II] emission line is higher than the value reported by Miszalski et al. (2012), but lower than the value provided by Cavichia et al. (2010). In terms of temperature, the values obtained from the [O III] and He I line ratios are consistent with those reported in the literature, falling within the uncertainty limits. However, the temperature derived from the [O II] diagnostic line ratio is notably higher than the corresponding values found in previous studies.

Figure 5 illustrates the electron temperature distribution of IC 4663 using three different diagnostic line ratios. In the left panel, the temperature map derived from the [O III] diagnostic ratio is shown, revealing a uniform temperature distribution across the PN, ranging from ~9.0 kK to 18.0 kK. Some regions along the PN's border exhibit higher temperatures. In the middle panel, the temperature map derived from the [N II] diagnostic line ratio is presented, showing a slightly higher temperatures compared to the [O III] map, with a range of ~9.0 kK to 30.0 kK. Notably, the highest temperature region corresponds to the lower middle part of the PN, aligning with the location of the southern lobe. The right panel showcases the electron temperature derived from the He I line ratio. This map demonstrates a decrease in temperature from the outer to inner regions of the PN, with temperature values ranging from ~20.0 kK to 5.0 kK. This finding consistent with high-abundance discrepancy factor PNe, which can be attributed to the presence of two obvious gas phases. The inner region exhibits low temperatures and high metal content, while the outer region displays warmer temperatures and normal metal content, as highlighted by García-Rojas et al. (2022).

However, it is important to note that due to the poor signal-to-noise ratios, an accurate map of the distribution of interstellar extinction coefficients across the complete PN could not be generated.

4.3 Ionic and elemental abundances

Table 4 provides the calculated ionic and elemental abundances of IC 4663 using the NEAT code. The abundances of nitrogen, oxygen, neon, argon, sulfur, and chlorine were determined through the analysis of CELs, while the abundances of helium and carbon were derived from the ORLs. The ionic abundances were computed by considering the appropriate temperature and density values for their specific ionization zones. In situations where multiple lines were observed for the same ion, the average abundance was calculated. The elemental abundances were obtained by applying the ionization correction factors (ICFs) reported by Delgado-Inglada et al. (2014) to the calculated ionic abundances.

TABLE 5 Comparison between elemental abundances of IC 4663 and prior studies, Type I PNe, and Sun.

Element	Ours	Cavichia et al. (2010)	Miszalski et al. (2012)	Type I PN (8)	Sun (9)
He/H	0.136	0.111	0.111	0.111	0.1
C/H	5.25E-03		1.58E-04 (4)	2.69E-04	2.69E-04
N/H	2.55E-04	3.09E-04 (1)	1.82E-04 (5)	5.25E-04	6.76E-05
O/H	4.51E-04	6.92E-04 (2)	5.01E-04 (5)	4.47E-04	4.90E-04
Ne/H	1.26E-04	1.78E-04 (1)	1.29E-04 (4)	1.23E-04	8.51E-05
Ar/H	4.93E-06	7.41E-06 (3)	2.69E-06 (6)	2.63E-06	2.51E-06
S/H	2.61E-05	1.07E-05 (1)	1.05E-05 (7)	8.13E-06	1.32E-05
Cl/H	1.78E-07		3.16E-07 (7)		3.16E-07
N/O	0.57	0.45	0.36	1.17	0.14

References of ICFs, and the sun and Orion nebula abundances: (1) Kingsburgh and Barlow (1994), (2) Torres-Peimbert and Peimbert (1977), (3) de Freitas Pacheco et al. (1993), (4) Miszalski et al. (2012), (5) Aller (1984), (6) Barker (1980), (7) Koeppen et al. (1991), (8) Kingsburgh and Barlow (1994), (9) Asplund et al. (2009).

Table 5 presents a comparison of the total abundances of IC 4663 with Type I PNe, solar abundances, and values reported in the literature. The results show a noticeable increase in He, C, and N and slightly increase in Ne, Ar, and S abundances compared to solar values. However, these elevated values are in good agreement with the abundances observed in Type I PNe. The differences in abundance measurements between our analysis and those of Cavichia et al. (2010) and Miszalski et al. (2012) can be attributed to two factors. Firstly, we used ICFs provided by Delgado et al. (2014), while Cavichia et al. (2010) and Miszalski et al. (2012) have used ICFs that determined from many different sources (see Table 5). Secondly, our analysis was based on IFU spectra of IC 4663, while the previous both studies were relied on long-slit spectra.

The PN IC 4663 displays remarkable enrichments of helium (He/H ≥0.125) and nitrogen [log (N/H) + 12 ≥8.0], along with an N/O ratio ≥0.5. These characteristics classify IC 4663 as a Type I PN according to the classification scheme introduced by (Peimbert and Torres-Peimbert, 1983). This classification suggests that the progenitor star of IC 4663 had initial mass ≥3.0 M⊙ (Quireza et al., 2007). In addition to determining the nitrogen and oxygen abundances from the CELs, the analysis also included their calculations from ORLs, specifically O^{2+}/H^{+}, N^{2+}/H^{+}, and N^{3+}/H^{+}. The O^{2+}/H^{+} abundance was derived from the V1, V2, V5, V10, V19, V20, and V48b multiplets, which exhibited consistent agreement with each other. Similarly, the N^{2+}/H^{+} abundances were determined from the V3, V28, and V39b multiplets, while the N^{3+}/H^{+} abundances were obtained from the V2 multiplet.

Table 6 presents the fractional ionic abundances of O II, N II, and N III, along with the averages of each multiplet. To calculate the abundance of each multiplet, an abundance-weighted average of its components was computed, and the full abundances were determined as the average of the multiplet abundances. To assess the abundance discrepancy, two key factors were defined. The ADF (O^{2+}) was determined as the ratio of the ORL abundance of O^{2+} to the CEL abundance of O^{2+}. Similarly, the ADF (N) was calculated as the ratio of the ORL abundances of N^{2+} and N^{3+} to the total nitrogen abundance derived from the CELs. The resulting abundance discrepancy factors are presented in Table 6 for both oxygen and nitrogen. They reveal an extreme abundance discrepancy for both oxygen and nitrogen in IC 4663, with an ADF(O^{2+}) of approximately 39.0 ± 6 and an ADF (N) of approximately 37.0 ± 9. It is worth noting that the derived ADF for oxygen is considered more reliable than that of nitrogen. This is due to the fact that the ORL nitrogen abundance is mainly determined from the N^{2+} and N^{3+} abundances, while the nitrogen abundance from the CELs relies solely on N^{+} lines and heavily depends on the estimated ICF. The accuracy of the ICF becomes particularly uncertain for CELs where only a small fraction of nitrogen is in the form of N^{+}.

These high ADF values observed for oxygen and nitrogen place IC 4663 among the group of PNe with extreme ADFs (ADF > 10) such as A 46, Fg 1, Hf 2-2, and Pe 1–9 (Wesson et al., 2018). According to the findings of Wesson et al. (2018), extreme abundance discrepancy in any PN is indicative of a binary CS. However, this hypothesis also suggests that the presence of a binary CS is insufficient to account for extreme ADFs.

5 Is the CS of IC 4663 part of a close binary system?

The CS of IC 4663 has been classified as a nitrogen Wolf-Rayet [WN] type, making it the first star of this particular class to be identified within PNe. The CS's spectrum exhibits broad emission lines associated with He II and NV, while lacking prominent lines from He I and C IV, placing it with the [WN3] subtype. Observations conducted by Miszalski et al. (2012) revealed radial velocity (RV) variability in the range of 3–4 km/s^{-1}. They concluded that the CS is not in a binary system with a main-sequence or white dwarf companion because in such a binary system the RV variability must be ≳ 5–10 km/s^{-1}, corresponding to a period ≲ 1 day. Additionally, they attributed this minor variability to either noise or wind turbulence affecting the line profiles in the CS spectrum.

TABLE 6 Fractional ionic abundances of O^{2+}, N^{2+}, and N^{3+} lines in IC 4663.

λ(Å)	Multiplet	$\frac{X(line)}{H^+}$	λ(Å)	Multiplet	$\frac{X(line)}{H^+}$
O^{2+}/H^+			N^{2+}/H^+		
4641.81	V1	2.94E-02 ± 5.50E-04	5666.63	V3	1.54E-03 ± 1.25E-04
4649.13	V1	1.30E-02 ± 1.00E-03	5676.02	V3	1.55E-03 ± 5.10E-04
4661.63	V1	3.52E-03 ± 2.70E-04	5679.56	V3	1.51E-03 ± 1.35E-04
4676.23	V1	4.50E-03 ± 5.65E-04	5686.21	V3	1.16E-03 ± 5.35E-04
		1.26E-02 ± 3.25E-04	5710.77	V3	2.48E-03 ± 6.05E-04
4345.55	V2	1.60E-02 ± 1.20E-02			1.65E-03 ± 1.95E-04
4349.43	V2	3.67E-03 ± 4.10E-04	5931.78	V28	2.66E-03 ± 2.45E-04
		9.84E-03 ± 6.00E-03	5941.65	V28	7.68E-04 ± 7.55E-05
4414.9	V5	1.40E-02 ± 1.00E-03			1.71E-03 ± 1.28E-04
4416.97	V5	1.99E-03 ± 4.65E-04	4026.08	V39b	5.80E-02 ± 4.00E-03
4452.38	V5	2.20E-02 ± 9.00E-03	4041.31	V39b	1.40E-02 ± 4.00E-03
		1.27E-02 ± 3.02E-03			3.60E-02 ± 2.83E-03
4069.62	V10	1.30E-02 ± 1.00E-03			
4072.15	V10	4.34E-03 ± 5.00E-04	N^{3+}/H^+		
4075.86	V10	1.10E-02 ± 1.00E-03	4366.89	V2	1.70E-02 ± 2.00E-03
4078.84	V10	1.00E-03 ± 1.00E-03	4640.64	V2	2.25E-03 ± 3.00E-05
		7.34E-03 ± 4.51E-04			9.63E-03 ± 1.00E-03
4153.3	V19	1.90E-03 ± 3.20E-04			
4156.53	V19	6.00E-03 ± 4.00E-03			
		3.95E-03 ± 2.01E-03			
4110.79	V20	5.00E-03 ± 2.00E-03			
4119.22	V20	9.00E-03 ± 2.00E-03			
		7.00E-03 ± 1.41E-03			
4083.9	V48b	2.00E-03 ± 2.00E-03			
4097.26	V48b	5.30E-02 ± 1.00E-02			
		2.75E-02 ± 5.10E-03			
O^{2+}/H^+		**1.13E-02 ± 1.26E-03**	N^{2+}/H^+ & N^{3+}/H^+		**9.36E-03 ± 7.52E-04**
ADF		**39.44 ± 5.94**	ADF		**36.69 ± 8.58**

The weighted mean values were denoted using boldface formatting.

Soker (1997) proposed that the axisymmetric shapes of PNe, such as bipolar and elliptical morphologies, are a consequence of the axisymmetric mass loss experienced by progenitor stars during their interaction with binary companions. These companions can take the form of either stellar or sub-stellar objects, including planets and brown dwarfs. Based on this hypothesis, PNe were categorized into four distinct classes: 1) Progenitors that did not undergo any interaction with their companions; 2) Progenitors that interacted with stellar companions but did not go through a CE phase; 3) Progenitors that interacted with stellar companions through a CE phase; 4) Progenitors that interacted with sub-stellar companions through a CE phase (56% of the PN sample). In accordance with this classification, IC 4663 was assigned to class 4), implying that the morphology of IC 4663 is a result of its progenitor star being part of a close binary system with a sub-stellar companion. The criteria proposed this class also provide an explanation for the presence of spherical halos observed around certain elliptical PNe, as observed in the case of IC 4663 (see Figure 5 in Miszalski et al. (2012)). In their analysis of TESS light curves, Aller et al. (2020) examined eight PNCSs to uncover potential binary systems. Except for one case, they identified evident signs of variability in all of them. The light curves displayed a pattern of variability consistent with an irradiation effect, and two of the targets exhibited ellipsoidal variability. Their findings indicate that the observed variability limits the potential companion to either a very low-mass main-sequence star or a sub-stellar object. In the case where the companion to the CS is a sub-stellar object, it is conceivable for it to display slight variability, as reported in the study by Miszalski et al. (2012).

Sub-stellar objects, such as Jupiter-type exoplanets, are known to have RV variations that range from a few meters per second to considerably higher values. Particularly, in systems where massive gas giants orbit their host stars at close distances, these RV variations can span from a few meters per second up to hundreds of meters per second (Tala Pinto et al., 2020; Hara and Ford, 2023).

In the context of IC 4663, the detection of minor RV variability identified by Miszalski et al. (2012), together with its classification as a class (4) PN by Soker (1997), and the correlation established by Wesson et al. (2018) linking high-ADF PNe to CS binarity, all support the hypothesis of a sub-stellar companion in a common envelope phase with the CS of IC 4663.

To investigate the potential binarity of the PNCS, we thoroughly examined various photometric databases, including the Gaia, Kepler, and TESS missions. Unfortunately, our search did not yield any positive results confirming the presence of a companion. Consequently, further monitoring of the CS is imperative to conclusively determine whether a companion to the PNCS exists or not.

6 IC 4663 interaction with ambient ISM

A spherical AGB halo surrounding IC 4663 was detected in the [O III] image taken by the Gemini South GMOS. The halo displayed an enhancement positioned approximately 22″ SE of the CS. The origin of this enhancement is attributed to the interaction between the PN and the ISM. Wareing et al. (2007) characterized this specific case of interaction as the initial evolutionary phase of interaction between PNe and ISM employing the triple wind model for this description. It should be noted that this phase, the ionised PN remains unaffected by the interaction. Notably, IC 4663 exhibits similar characteristics to the cases of NGC 2438, NGC 2440, NGC 3242, and IC 4593 (Ali et al., 2012).

Following the procedure provided by Mohery et al. (2022), we determined a position angle (PA) for the proper motion of the CS, measuring 171 ± 45. Despite the slightly high uncertainty, primarily caused by a significant error in the μ_α component, this discovery strongly supports the possibility that the observed enhancement in the AGB halo results from its interaction with the ISM. To calculate the PA, we utilized proper motion measurements obtained from the Gaia Data Release 3 (Gaia DR3), which included $\mu_\alpha = -0.0787 \pm -0.0787$ and $\mu_\delta = -0.0787 \pm 0.052469$, along with a photo-geometric distance of 2400^{+557}_{-320} pc determined by Bailer-Jones et al. (2021).

7 Conclusion

In this study, we conducted the first integral field spectroscopy of the southern PN IC 4663, spanning the optical range of 3400–8950 Å. The analysis of the emission-line maps, encompassing low, intermediate, and high ionisation lines, consistently revealed an elliptical shape with two inner illuminated lobes extending in opposite directions from the faint core. This formation created a distinct double-lobed structure. The emission maps, whether from CELs or ORLs, shared a similar overall morphology. However, disparities were noted in the brightness of the lobes and the apparent size of the PN in the CEL and ORL emission-line maps. Moreover, our study confirmed the interaction between IC 4663 and its surrounding ISM by calculating the position angle of CS's motion towards the interaction region.

Spectral analysis of the PN has identified it is a high excitation class object that is nearly optically thin. Furthermore, the Doppler shift observed in the emission lines yielded a heliocentric radial velocity of −80.0 ± 22.0 km/s, which aligns with the values reported in literature. Our measurements of the electron temperature and density of the PN, derived from different diagnostic lines across various ionization stages, are consistent with the previous studies. The electron temperature maps generated for [O III] and [N II] reveal a uniform distribution across the PN, with the exception of slightly higher temperatures observed in the lower middle section of the PN for [N II]. In contrast, the constructed He I map demonstrates a temperature gradient, declined from the outer to inner regions of the PN, thus suggesting the presence of two different gas phases: a cooler inner region and a comparatively warmer outer region. This temperature profile is consistent with what is typically observed in PNe with high ADFs.

Based on the derived elemental abundances of helium, nitrogen, and oxygen, we classified IC 4663 as a Peimbert type I PN, suggesting a progenitor star with a mass larger than 3.0 M⊙. The elemental abundances of the PN show good agreement with those of Type I PNe, exceeding solar abundances. Additionally, there were slight variations in these elemental abundances compared to values reported in the literature. Furthermore, a substantial discrepancy was observed between the oxygen and nitrogen abundances calculated from ORLs versus those from CELs, with a factor of

approximately 40. This categorizes IC 4663 as a PN with extreme ADFs.

The hypothesis of a close binary system within IC 4663, comprising the CS and a sub-stellar companion, is supported by several key observations: (1) the detection of minor radial velocity variability in the CS of IC 4663 CS, (2) the progenitor star of IC 4663 has been categorized as part of a close binary system that underwent interaction with sub-stellar companions during a common envelope phase, and (3) the correlation between high ADFs in PNe and close binary CS systems.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

MM: Formal Analysis, Writing–original draft, Funding acquisition, Methodology. AA: Formal Analysis, Writing–original draft, Data curation, Investigation, Writing-review and editing. AAK: Formal Analysis, Investigation, Data curation, Methodology. SS: Formal Analysis, Investigation, Writing–original draft. AM: Formal Analysis, Funding acquisition.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the University of Jeddah, Jeddah, Saudi Arabia under grant No. (UJ-23-DR-249). Therefore, the authors thank the University of Jeddah for its technical and financial support.

Acknowledgments

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia under grant No. (UJ-23-DR-249). Therefore, the authors thank the University of Jeddah for its technical and financial support. We express our gratitude to the referees for their careful examination of our work and valuable feedback and suggestions, which have significantly enhanced the quality of the paper. This work presents results from the European Space Agency (ESA) space mission Gaia. Gaia data are being processed by the Gaia Data Processing and Analysis Consortium (DPAC).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2023.1322980/full#supplementary-material

References

Akras, S., and Gonçalves, D. R. (2016). Low-ionization structures in planetary nebulae - I. Physical, kinematic and excitation properties. *MNRAS* 455, 930–961. doi:10.1093/mnras/stv2139

Ali, A., Amer, M. A., Dopita, M. A., Vogt, F. P. A., and Basurah, H. M. (2015). The physics and kinematics of the evolved, interacting planetary nebula PN G342.0-01.7. *MNRAS* 583, A83. doi:10.1051/0004-6361/201526223

Ali, A., and Dopita, M. A. (2019). IFU spectroscopy of southern PNe - VII. Photoionization modelling of intermediate excitation class objects. *MNRAS* 484, 3251–3266. doi:10.1093/mnras/stz201

Ali, A., Dopita, M. A., Basurah, H. M., Amer, M. A., Alsulami, R., and Alruhaili, A. (2016). IFU spectroscopy of southern planetary nebulae - III. *MNRAS* 462, 1393–1404. doi:10.1093/mnras/stw1744

Ali, A., Sabin, L., Snaid, S., and Basurah, H. M. (2012). Interacting planetary nebulae: I. Classification and orientation⋆. *A&A* 541, A98. doi:10.1051/0004-6361/201118389

Aller, A., Lillo-Box, J., Jones, D., Miranda, L. F., and Barceló Forteza, S. (2020). Planetary nebulae seen with TESS: discovery of new binary central star candidates from Cycle 1. *A&A* 635, A128. doi:10.1051/0004-6361/201937118

Aller, L. H. (1984). *Physics of thermal gaseous nebulae*. Berlin, Germany: Springer. doi:10.1007/978-94-010-9639-3

Asplund, M., Grevesse, N., Sauval, A. J., and Scott, P. (2009). The chemical composition of the sun. *ARA&A* 47, 481–522. doi:10.1146/annurev.astro.46.060407.145222

Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., and Andrae, R. (2021). Estimating distances from parallaxes. V. Geometric and photogeometric distances to 1.47 billion stars in Gaia early data Release 3. *AJ* 161, 147. doi:10.3847/1538-3881/abd806

Balick, B. (1987). The evolution of planetary nebulae. I. Structures, ionizations, and morphological sequences. *AJ* 94, 671. doi:10.1086/114504

Balick, B., Alexander, J., Hajian, A. R., Terzian, Y., Perinotto, M., and Patriarchi, P. (1998). FLIERs and other microstructures in planetary nebulae. IV. Images of elliptical PNs from the [ITAL]hubble[/ITAL] [ITAL]space[/ITAL] [ITAL]T[/ITAL] [ITAL]elescope[/ITAL]. *AJ* 116, 360–371. doi:10.1086/300429

Balick, B., and Frank, A. (2002). Shapes and shaping of planetary nebulae. *ARA&A* 40, 439–486. doi:10.1146/annurev.astro.40.060401.093849

Barker, T. (1980). The ionization structure of the ring nebula. I. Sulfur and argon. *ApJ* 240, 99–104. doi:10.1086/158210

Basurah, H. M., Ali, A., Dopita, M. A., Alsulami, R., Amer, M. A., and Alruhaili, A. (2016). Problems for the WELS classification of planetary nebula central stars: self-consistent nebular modelling of four candidates. *MNRAS* 458, 2694–2709. doi:10.1093/mnras/stw468

Beaulieu, S. F., Dopita, M. A., and Freeman, K. C. (1999). A survey of planetary nebulae in the southern galactic bulge. *ApJ* 515, 610–632. doi:10.1086/307052

Cahn, J. H., Kaler, J. B., and Stanghellini, L. (1992). A catalogue of absolute fluxes and distances of planetary nebulae. *A&AS* 94, 399–452.

Cavichia, O., Costa, R. D. D., and Maciel, W. J. (2010). Planetary nebulae in the inner Milky Way: new abundances. *Rev. Mex. Astron. Astrofis.* 46, 159–177. doi:10.48550/arXiv.1003.0416

Childress, M. J., Vogt, F. P. A., Nielsen, J., and Sharp, R. G. (2014). PyWiFeS: a rapid data reduction pipeline for the Wide field spectrograph (WiFeS). *Ap&SS* 349, 617–636. doi:10.1007/s10509-013-1682-0

Corradi, R. L. M., Kwitter, K. B., Balick, B., Henry, R. B. C., and Hensley, K. (2015). The chemistry of planetary nebulae in the outer regions of M31. *ApJ* 807, 181. doi:10.1088/0004-637X/807/2/181

Corradi, R. L. M., Manso, R., Mampaso, A., and Schwarz, H. E. (1996). Unveiling low-ionization microstructures in planetary nebulae. *A&A* 313, 913–923.

Costa, R. D. D., Uchida, M. M. M., and Maciel, W. J. (2004). Chemical abundances of planetary nebulae towards the Galactic anticenter. *A&A* 423, 199–207. doi:10.1051/0004-6361:20034539

Danehkar, A., Parker, Q. A., and Ercolano, B. (2013). Observations and three-dimensional ionization structure of the planetary nebula SuWt 2. *MNRAS* 434, 1513–1530. doi:10.1093/mnras/stt1116

de Freitas Pacheco, J. A., Barbuy, B., Costa, R. D. D., and Idiart, T. E. P. (1993). Type I planetary nebulae in the Large Magellanic Cloud: oxygen, sulphur and argon abundances as tracers of chemical enrichment. *A&A* 271, 429–434.

Delgado-Inglada, G., Morisset, C., and Stasińska, G. (2014). *In revista Mexicana de astronomia y astrofisica conference series*, 17–17.

Dopita, M. A., Ali, A., Sutherland, R. S., Nicholls, D. C., and Amer, M. A. (2017). IFU spectroscopy of southern planetary nebulae IV: a physical model for IC 418. *MNRAS* 470, 839–864. doi:10.1093/mnras/stx1166

Dopita, M., Hart, J., McGregor, P., Oates, P., Bloxham, G., and Jones, D. (2007). Astrophys. *Space Sci.* 310, 255.

Durand, S., Acker, A., and Zijlstra, A. (1998). The kinematics of 867 galactic planetary nebulae. *A&AS* 132, 13–20. doi:10.1051/aas:1998356

Faúndez-Abans, M., and Maciel, W. J. (1986). Galactic abundance gradients from planetary nebulae. *Rev. Mex. Astron. Astrofis.* 12, 240–248.

García-Rojas, J., Morisset, C., Jones, D., Wesson, R., Boffin, H. M. J., Monteiro, H., et al. (2022). MUSE spectroscopy of planetary nebulae with high abundance discrepancies. *MNRAS* 510, 5444–5463. doi:10.1093/mnras/stab3523

García-Segura, G., Taam, R. E., and Ricker, P. M. (2022). Common-envelope shaping of planetary nebulae - IV. From protoplanetary to planetary nebula. *MNRAS* 517, 3822–3831. doi:10.1093/mnras/stac2824

Gómez-Llanos, V., Morisset, C., García-Rojas, J., Jones, D., Wesson, R., Corradi, R. L. M., et al. (2020). The impact of strong recombination on temperature determination in planetary nebulae. *MNRAS* 498, L82–L86. doi:10.1093/mnrasl/slaa131

Gonçalves, D. R., Corradi, R. L. M., and Mampaso, A. (2001). Low-ionization structures in planetary nebulae: confronting models with observations. *ApJ* 547, 302–310. doi:10.1086/318364

Hara, N. C., and Ford, E. B. (2023). Statistical methods for exoplanet detection with radial velocities. *Annu. Rev. Statistics Its Appl.* 10, 623–649. doi:10.1146/annurev-statistics-033021-012225

Howarth, I. D. (1983). LMC and galactic extinction. *MNRAS* 203, 301–304. doi:10.1093/mnras/203.2.301

Kingsburgh, R. L., and Barlow, M. J. (1994). Elemental abundances for a sample of southern galctic planetary nebulae. *MNRAS* 271, 257–299. doi:10.1093/mnras/271.2.257

Koeppen, J., Acker, A., and Stenholm, B. (1991). Spectrophotometric survey of southern planetary nebulae. II. Chemical compositions. *A&A* 248, 197.

Kwok, S., Purton, C. R., and Fitzgerald, P. M. (1978). On the origin of planetary nebulae. *ApJ* 219, L125–L127. doi:10.1086/182621

Maciel, W. J., Lago, L. G., and Costa, R. D. D. (2006). New results on the time variation of the radial abundance gradients from planetary nebulae. *Proc. Int. Astronomical Union* 2 (S234), 453–454. doi:10.1017/S174392130600370X

Magrini, L., Stanghellini, L., and Gonçalves, D. R. (2012). Extragalactic planetary nebulae: tracers of the chemical evolution of nearby galaxies. *Planet. Nebul. Eye Future* 283, 251–258. doi:10.1017/S1743921312011040

Mari, M. B., Akras, S., and Gonçalves, D. R. (2023). Low-ionization structures in planetary nebulae - III. The statistical analysis of physico-chemical parameters and excitation mechanisms. *MNRAS* 525, 1998–2014. doi:10.1093/mnras/stad2256

Mellema, G. (1996). Hydrodynamic models of planetary nebulae. *Ap&SS* 245, 239–253. doi:10.1007/BF00642229

Mellema, G., and Frank, A. (1995). Radiation gasdynamics of planetary nebulae - V. Hot bubble and slow wind dynamics. *MNRAS* 273, 401–410. doi:10.1093/mnras/273.2.401

Miszalski, B., Crowther, P. A., De Marco, O., Köppen, J., Moffat, A. F. J., Acker, A., et al. (2012). IC 4663: the first unambiguous [WN] Wolf-Rayet central star of a planetary nebula. *MNRAS* 423, 934–947. doi:10.1111/j.1365-2966.2012.20929.x

Mohery, M., Ali, A., Mindil, A., and Alghamdi, S. A. (2022). Interacting planetary nebulae III: verification and galactic population based on the measurements of Gaia EDR3. *Ap&SS* 367, 63. doi:10.1007/s10509-022-04085-y

Monreal-Ibero, A., Roth, M. M., Schönberner, D., Steffen, M., and Böhm, P. (2005). Integral field spectroscopy of faint halos of planetary nebulae. *ApJ* 628, L139–L142. doi:10.1086/432664

Peimbert, M., and Torres-Peimbert, S. (1983). "Type I planetary nebulae,". *Planetary nebulae*. Editor D. R. Flower (London, U.K: IAU Symposium), 233–241.

Puspitaningrum, E., and Malasan, H. L. (2019). Photoionization modelling of the Wolf-Rayet planetary nebulae IC 4663 based on multiwavelength observations. *J. Phys. Conf. Ser.* 1231, 012031. doi:10.1088/1742-6596/1231/1/012031

Quireza, C., Rocha-Pinto, H. J., and Maciel, W. J. (2007). Bayesian posterior classification of planetary nebulae according to the Peimbert types. *A&A* 475, 217–231. doi:10.1051/0004-6361:20078087

Reid, W. A., and Parker, Q. A. (2010). An evaluation of the excitation-class parameter for the central stars of planetary nebulae. *PASA* 27, 187–198. doi:10.1071/AS09055

Schneider, S. E., and Terzian, Y. (1983). Planetary nebulae and the galactic rotation curve. *ApJ* 274, L61–L64. doi:10.1086/184151

Snaid, S., Ali, A., and Alruhaili, A. (2023). Comprehensive study of the galactic planetary nebula NGC 2792. *Ap&SS* 368, 62. doi:10.1007/s10509-023-04218-x

Soker, N. (1997). Properties that cannot Be explained by the progenitors of planetary nebulae. *ApJS* 112, 487–505. doi:10.1086/313040

Tala Pinto, M., Reffert, S., Quirrenbach, A., Stock, S., Trifonov, T., and Mitchell, D. S. (2020). Precise radial velocities of giant stars. XIV. Evidence of planetary companions around HD 25723, 17 Sco, 3 Cnc, and 44 UMa. *A&A* 644, A1. doi:10.1051/0004-6361/202038285

Torres-Peimbert, S., and Peimbert, M. (1977). Photoelectric photometry and physical conditions of planetary nebulae. *Rev. Mex. Astron. Astrofis.* 2, 181–207.

Tsamis, Y. G., Walsh, J. R., Pequignot, D., Barlow, M. J., Liu, X.-W., and Danziger, I. J. (2007). Integral-field spectroscopy of galactic planetary nebulae with VLT FLAMES. *Messenger* 127, 53.

Wareing, C. J., Zijlstra, A. A., and O'Brien, T. J. (2007). The interaction of planetary nebulae and their asymptotic giant branch progenitors with the interstellar medium. *MNRAS* 382, 1233–1245. doi:10.1111/j.1365-2966.2007.12459.x

Wesson, R. (2016). ALFA: an automated line fitting algorithm. *MNRAS* 456, 3774–3781. doi:10.1093/mnras/stv2946

Wesson, R., Jones, D., García-Rojas, J., Boffin, H. M. J., and Corradi, R. L. M. (2018). Confirmation of the link between central star binarity and extreme abundance discrepancy factors in planetary nebulae. *MNRAS* 480, 4589–4613. doi:10.1093/mnras/sty1871

Wesson, R., Stock, D. J., and Scicluna, P. (2012). Understanding and reducing statistical uncertainties in nebular abundance determinations. *MNRAS* 422, 3516–3526. doi:10.1111/j.1365-2966.2012.20863.x

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 15 February 2024
DOI 10.3389/fspas.2023.1177097

Check for updates

OPEN ACCESS

EDITED BY
Maria Ann Weber,
Delta State University, United States

REVIEWED BY
Robert James Leamon,
University of Maryland, United States
Bidya Bina Karak,
Indian Institute of Technology (BHU), India

*CORRESPONDENCE
Sara F. Martin,
sara@helioresearch.org

RECEIVED 01 March 2023
ACCEPTED 13 December 2023
PUBLISHED 15 February 2024

CITATION
Martin SF (2024), Observations key to understanding solar cycles: a review. *Front. Astron. Space Sci.* 10:1177097. doi: 10.3389/fspas.2023.1177097

COPYRIGHT
© 2024 Martin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Observations key to understanding solar cycles: a review

Sara F. Martin*

Helio Research, Glendale, CA, United States

A paradigm shift is taking place in the conception of solar cycles. In the previous conception, the changing numbers of sunspots over intervals of 9–14 years have been regarded as the fundamental solar cycle although two average 11-year cycles were necessary to account for the complete magnetic cycle. In the revised picture, sunspots are a phase in the middle of two 22-year overlapping solar cycles that operate continuously with clock-like precision. More than 20 researchers have contributed to the initial research articles from 2014 through 2021 which are dramatically altering the perception of solar cycles. The two 22-year cycles overlap in time by 11 years. This overlap is coincidentally the same average duration as the sunspot phase in each 22-year cycle. This coincidence and the relative lack of knowledge of the large numbers of small active regions without sunspots is what led to the previous paradigm in which the 11-year sunspot phases were misinterpreted as a single fundamental solar cycle. The combination of the two 22-year solar cycles, with their large numbers of short-lived active regions and ephemeral active regions are now understood to be the fundamental cycle with the proposed name "The Hale Solar Cycle." The two 22-year solar cycles each occupy separate but adjacent bands in latitude. The orientations of the majority of bipolar magnetic regions in the two adjacent bands differ from each other by ~180°. Both bands continuously drift from higher to lower latitudes as has been known for sunspot cycles. However, the polarity reversal occurs at the start of each 22-year cycle and at higher latitudes than it does for the sunspot cycles. This paradigm shift in the concept of solar cycles has resulted in major reconsiderations of additional topics on solar cycles in this review. These are 1) the large role of ephemeral active regions in the origin of solar cycles, 2) the depth of the origin of active regions and sunspots, 3) the mechanisms of how areas of unipolar magnetic network migrate to the solar poles every 11 years, and 4) the nature of the polarity reversal in alternate 22-year cycles rather than 11-year cycles.

KEYWORDS

solar cycles, magnetic fields, sunspots, filaments, paradigm

1 Introduction

A paradigm is "A set of assumptions, concepts, values, and practices that constitutes a way of viewing reality for the community that shares them, especially in an intellectual discipline." This review discusses solar cycles from the perspective of a paradigm shift on solar cycles now taking place. The evidence is from observations. A few solar cycle theories are mentioned in this review but are not discussed in detail. The reader is referred to a complementary review on theories of solar cycles by

Weber et al. (2023) and other review articles referenced therein. In this limited review, solar observations are first discussed as occurring in eras that span decades of solar research and are identified as paradigms.

I suggest that the previous paradigm was stimulated by new magnetic field observations during the early days of the Mt. Wilson Observatory using spectroheliographs that were new at that time. Paramount among the early discoveries of the unique magnetic properties of sunspot cycles are 1) the discovery of the bipolar nature of sunspots and magnetic fields of their active regions, 2) the finding of opposite orientations in the northern and southern hemispheres with west-to-east orientations in the northern hemisphere in odd numbered sunspot cycles and east-to-west orientations in the northern hemisphere in even numbered cycles, with few exceptions, 3) the reversal of their dominant orientations at the beginning of each new sunspot cycle, and 4) the discovery of the two polar magnetic fields and their reversal in polarity during each 11-year solar maximum.

These magnetic properties of solar cycles found at the Mt. Wilson Observatory added to the properties learned from the previous 400 years of sunspot observations prior to our knowledge of their magnetic fields. The average sunspot cycle duration was 11 years. The number of sunspot groups increased to solar maximum over 3–4 years after the first sunspots appeared. The number declined over 5–6 years until few or no sunspots occurred during solar minimum. They occurred in two bands initially approximately 10–15° in width and increased to approximately 30° in latitude before decreasing again after solar maximum in both hemispheres. The initial sunspots were small, short-lived, and began approximately 30–40° in latitude. As the population of sunspots increased, a broadening of their distributions in size and latitude increased until solar maximum. After the maximum, the range in latitude decreased along with the decreasing population of sunspots until none remained near the equator at the end of the approximate 11 years of the cycle.

A puzzling difference between successive sunspot cycles was the complete change in the orientation of the magnetic fields of the sunspots and their surrounding bipolar magnetic fields by 180° with each new sunspot cycle (Hale et al., 2019; Hale and Nicholson, 1925). This meant the basic sunspot cycle length was approximately 11 years, but the complete magnetic cycle was two 11-year sunspot cycles.

The first articles to explicitly say that a new paradigm for solar cycles had emerged are by Srivastava et al. (2018) and McIntosh et al. (2021). The individual 20 researchers contributing to these articles recognize that the new paradigm was initiated in the article of McIntosh et al. (2014). Using proxies of EUV bright points for the large population of small active regions without sunspots, McIntosh et al. (2014) presented diagrams called "band-o-grams" showing two 22-year overlapping solar cycles present on the Sun continuously. In brief, this is a radically different perception from the 11-year average solar cycles described above. This paradigm shift is the finding of the true meaning of the 22-year magnetic cycle and re-identifying sunspots as phases in the 22-year cycles rather than independent cycles.

This change in perception of solar cycles from a single linear series of 11-year cycles to the realization that two complete overlapping 22-year cycles are always present on the Sun, is the substance of this review. Readers already familiar with solar cycles might choose to first read Section 4 on the extended solar cycle or Section 5 on the existence and ramifications of two 22-year, overlapping solar cycles. Sections 6–9 describes changed views on other aspects of solar cycles that could become a part of the new paradigm.

Section 2 briefly reviews and illustrates details learned about solar cycles from observing only sunspots.

Section 3 describes what was learned about sunspots and solar cycles when magnetic field measurements first became possible using the Mt. Wilson spectrographs and later the Babcock magnetograph.

Section 4 summarizes key details in the emergence of the concept of an extended solar cycle from magnetographs and instruments at various solar observatories.

Section 5 examines the existence and ramifications of two 22-year solar cycles continuously on the Sun.

Section 6 emphasizes the fundamental role of ephemeral active regions in understanding the 22-year Hale Solar Cycle. Section 6 also cites articles that conclude: the smallest bipolar structures of ephemeral region and larger active regions are the same features. They are named "elementary bipoles" because they appear to be the building blocks of all active regions.

Section 7 discusses observations relevant to identifying the source of the magnetic fields that migrate toward the poles and replace the polar fields around the maximum of the 11-year sunspot cycle.

Section 8 analyzes the evolution of magnetic fields with the potential for reversing the polarity orientation of the active regions at the beginning of every 22-year solar cycle.

Section 9 is the Discussion section; it presents a brief overview envisioning future solar cycle studies.

2 The solar cycle, as known from sunspots

Sunspots were observed by ancient cultures, and no individual can be given credit for their discovery. The historical record of sunspot cycles includes drawings made from naked-eye observations. An example is shown in Figure 1 made in the year 1618 and rediscovered by Wang and Li (2022). From this rediscovery, the authors could conclude a solar minimum, thought to have occurred in 1618, instead occurred in 1620. Such seemingly small details can be important; the cause of the 22-year solar cycle and longer periods in the variation of solar cycles remains one of the biggest unsolved puzzles in solar physics (Hathaway, 2015, a review).

Sunspots large enough to be seen by the naked eye are approximately one-hundredth of the solar diameter. Sunspots, this large and larger, often appear as pairs of slightly distorted, dark disks with surrounding smaller sunspots. Their frequency of occurrence from solar maximum to solar minimum over the 11-year solar cycle is illustrated in Figure 2. These images from the Solar Dynamics Observatory were recorded during the solar maximum of solar cycle 23 on 19 July 2000 (Figure 2A) and during the next solar minimum on 18 March 2009 (Figure 2B).

Figure 1B on the right shows a similar later series of drawings by Christoph Scheiner in Germany; a pair of sunspots in the northern

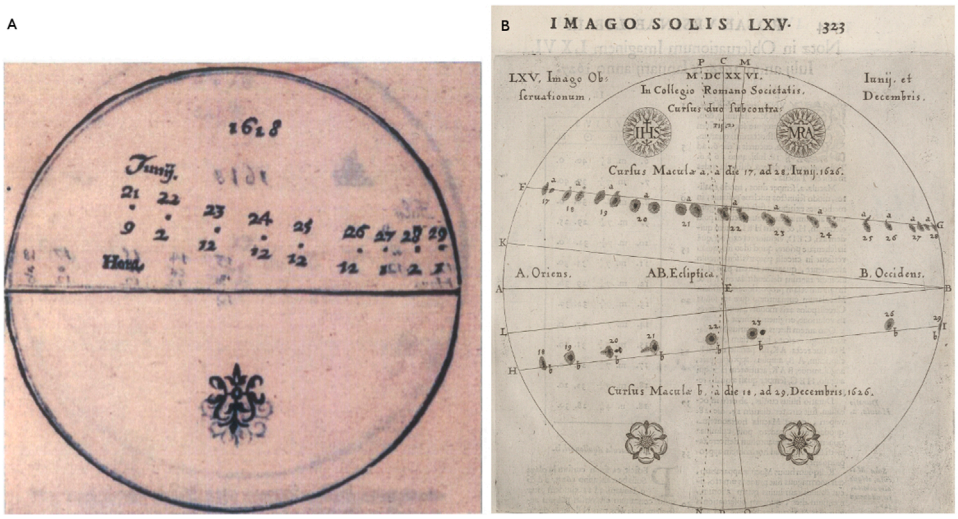


FIGURE 1
(A) Malapert drawing from the article "Rediscovery of 23 Naked-eye Sunspot Records." As originally written on the drawing, the dates are 21 to 29 July 1618 as published by Wang and Li (2022). The drawing of a dot represents the daily progress of a sunspot as it crosses the solar disk at a rate of approximately 13° per day. Sunspots remain at relatively fixed locations on the Sun. The curved path of the sunspots reveals that the north solar pole is tilted toward Earth. The horizontal line is not the equator. One can assume this sunspot stays at the same latitude and distance from the equator within a degree or less. Therefore, in this drawing, the rotational axis of the Sun is tilted clockwise approximately 20°. **(B)** On the right is seen a similar, later series of drawings by Christoph Scheiner in Germany; a pair of sunspots in the northern hemisphere was traced in 1626 from June 1 through June 28. The straight line K at the left side of the solar disk shows where a line of constant latitude intersects the east limb of the Sun. In the southern hemisphere, another sunspot was drawn in similar detail from 18 to 29 December 1626. The tilt of the rotational axis of the Sun in the sky is in the opposite direction from the drawing in the northern hemisphere. Credit for **(A)**: Wang and Li (2022). Credit for **(B)**: the library at the University of Oslo which owns an original of the Book by Scheiner, "Rosa Ursina sive Sol." The copied image was provided by the courtesy of Prof. Oddbjorn Engvold.

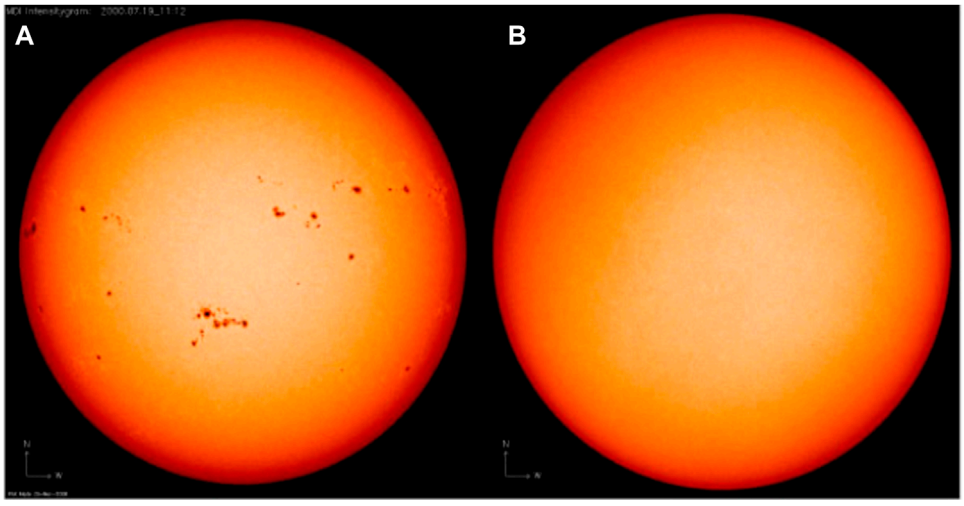


FIGURE 2
The image of **(A)** was recorded on 19 July 2000 during the maximum of solar cycle 23 and the image of **(B)** on 18 March 2009 during the following solar minimum. Credit: The Solar Dynamics Observatory and NASA.

hemisphere was traced in 1626 from June 1 through June 28. The straight line, K at the left side of the solar disk, shows where a line of constant latitude intersects the east limb of the Sun. In the southern hemisphere, another sunspot was drawn in similar detail from 18 to 29 December 1626. The tilt of the rotational axis of the Sun in the sky is in the opposite direction from the drawing in the northern hemisphere (credit for Figure 1A: Wang and Li, 2022 and credit for Figure 1B: to the library at the University of Oslo, which owns an original of the Book by Scheiner, "Rosa Ursina sive Sol." The copied image was provided by courtesy of Prof. Oddbjorn Engvold).

The discovery of the 11-year sunspot cycle in Western recorded history is usually credited to Schwabe (1844). However, the older

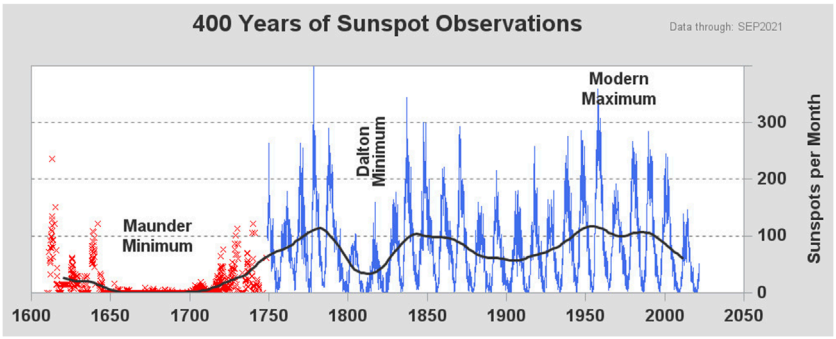


FIGURE 3
Sunspot number versus time that shows the variations from cycle to cycle over more than 400 years. The deep part of the Maunder Minimum lasted 70 years. The early part of current solar cycle 25 is shown at the right end of the graph. Credit: Space Weather Prediction Center.

sunspot drawings in China showed that sunspots were sufficiently long in duration that it is possible and also likely that earlier observers in cultures living close to nature would have discovered the sunspot cycle.

The sunspot cycle is commonly represented as a graph of sunspot number versus time as illustrated in Figure 3. To date, the observed sunspot numbers have exceeded the predictions for the first years of solar cycle 25. This difference between the observed and predicted sunspot numbers is sufficient reason for continuing research on solar cycles of our closest star.

Except for small motions in their growth phase, sunspots are relatively fixed in location. They frequently develop in pairs or small groups, as seen in Figure 1B by Scheiner, a co-temporary of Galileo (Engvold and Zirker, 2016). Their lifetimes are from a fraction of a day for small ones to several weeks for large ones, as shown in Figure 2. The longer apparent lifetimes are often observed in closely spaced groups of sunspots, repeatedly occurring at nearly the same latitudes and longitudes.

Changes in sunspots are readily seen daily as the Sun rotates around its central axis on average once every 27 days or approximately 13.5° per day. One can track some large sunspots across the 180° of the visible Sun realistically for up to 13 days. Figure 1 represents an example of three long-lived sunspots or sunspot groups. The tracking of sunspots like this one, four hundred or more years ago led astronomers to discover that the Sun's speed of rotation is fastest at its equator and slowest at its highest latitudes, a phenomenon known as differential rotation. They found that one rotation of the Sun at its equator requires 25 days but increases to 29 days at 60° north and south latitudes.

The "butterfly" diagram in Figure 4 illustrates the gradual drift of the formation sites of sunspots from high to low latitudes over 13 complete sunspot cycles and two partial cycles. Although the average duration of a sunspot cycle is approximately 11 years, they can be as short as 9 years or as long as 14 years. Sunspots of a single cycle are first seen in both solar hemispheres at relatively high latitudes, up to approximately 40°. Throughout each solar cycle, successive sunspots gradually appear at lower and lower latitudes until they are close to the solar equator. A few appear to cross the equator as the ~11-year sunspot cycle comes to an end. In Figure 4, one can see the band of latitudes, occupied by sunspots, widens until solar maximum and then narrow as the cycle declines. Sunspot cycles with the largest number also have the broadest distribution of sunspots in latitude. Some consecutive sunspot cycles overlap by 1–2 years but others do not overlap.

The above information about sunspot cycles was known before the year 1900. This was before Hale embarked successively on the construction of three of the largest solar spectroheliographs in the world at Mt. Wilson, California, United States, early in the 20th century. When sunspots were observed with the spectroheliographs, Hale and his colleagues found Zeeman splitting in solar spectral lines. This splitting was known to only occur in the presence of a magnetic field. From observing the Zeeman splitting, Hale and his colleagues could directly measure the strength and polarity of the magnetic fields of sunspots. The degree of Zeeman splitting is directly proportional to the magnetic field strength at the source on the spectrograph slit. The opposite polarizations, of the two split components of magnetically sensitive spectral lines, enabled the detection of sunspot polarity as either positive or negative. These measurements provided definitive evidence that sunspots are intense magnetic cores within the surrounding, larger magnetic fields in which they form. Hale and his colleagues measured magnetic fields ranging from 100 Gauss up to 2,300 Gauss (Hale and Nicholson, 1938). For more detailed information on sunspots, also see two books on sunspots by Bray and Loughhead (1964) and Loughhead and Bray (1979).

From all published information about sunspots, Babcock (1961) summarized five discoveries relevant to solar cycles. He presented them in the following order with credit to those who first discovered them. These discoveries are abbreviated here:

1. Differential rotation of the solar atmosphere, 25 days at the equator and 29.3 days at 60° north and south latitudes (Carrington, 1858).
2. An average 11-year rise and decline in the numbers of sunspots (Schwabe, 1844).
3. Limitation of sunspots to one band of latitudes in each hemisphere parallel to the solar equator (Carrington, 1858) and the slow progression of the bands toward the equator throughout the solar cycle (Sporer, 1894).
4. The bipolar character of sunspots. An imaginary line, from the centroid of sunspots of one polarity to the centroid of

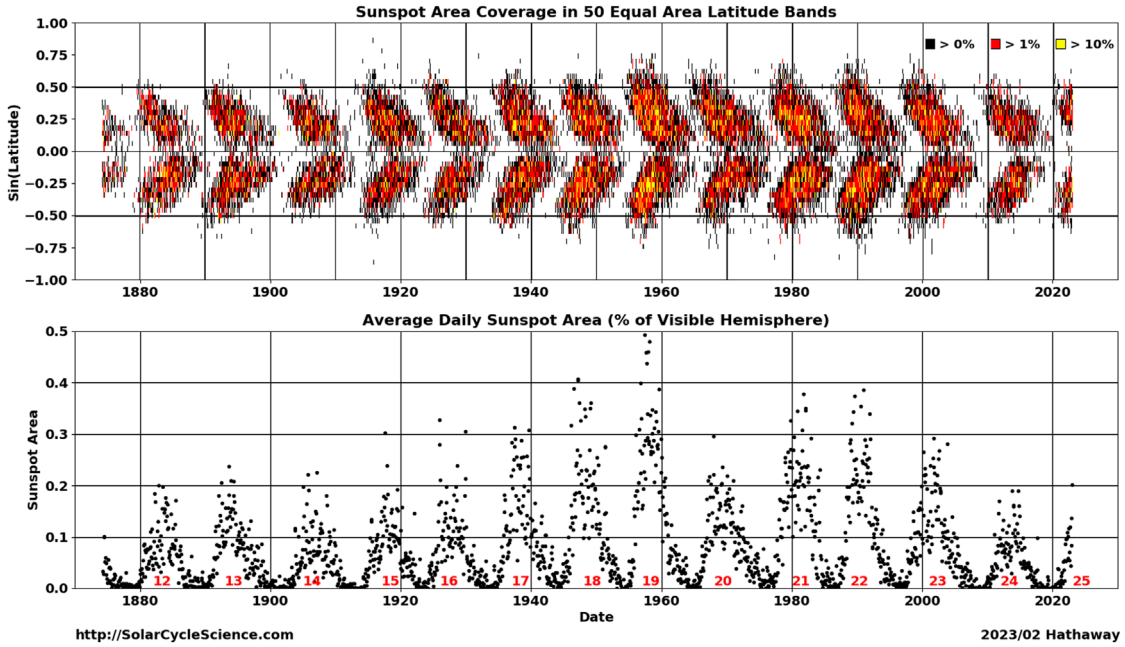


FIGURE 4
This "butterfly" diagram of 28 continuous 11-year sunspot cycles and two partial cycles reveals the range of latitudes of sunspots with time for both the northern and southern hemispheres. Sunspot cycles with the largest number of sunspots also have the broadest distribution of sunspots in latitude. Credit: David Hathaway, NASA Marshall Space Flight Center, Alabama, United States.

the opposite polarity, is aligned nearly parallel with the solar equator (Hale, 1913); the relative east–west orientations of the positive and negative polarities are opposite in the northern and southern hemispheres (Hale and Ellerman, 1918).

5. The reversal of the east–west polarity pattern of sunspots in both solar hemispheres at the beginning of every sunspot cycle. This led to the recognition of the Sun's 22-year magnetic cycle consisting of two 11-year sunspot cycles (Hale et al., 1919; Hale and Nicholson, 1925).

From the images of the Sun recorded in various spectral lines, called spectroheliograms, early scientists studying the Sun could readily see sunspots were embedded in areas with bright and dark intricate structure. These larger areas bear the name "active regions." This name is a catch-all for many types of dynamic solar events that occur in active regions, such as solar flares, surges, erupting prominences, and coronal mass ejections. Nearly all cyclical properties of sunspots are related to dynamic features produced in active regions.

In the early 1900s, daily observations of the Sun and its active regions and the associated features were also initiated at observatories around the world. Important, long-term, historic observations were taken at the Potsdam Observatory in Germany, the Kodaikanal Observatory in India, the Meudon Solar Observatory and Pic du Midi Observatory in France, and the Climax Observatory and Sacramento Peak Observatory in the United States. These major observatories have also recorded solar prominences above the solar limb and their counterpart as filaments on the solar disk. The reader is referred to other reviews and books for more historic details on solar cycle records (review on Solar Cycles by Hathaway, 2015; Solar Prominences by Tandberg-Hanssen, 1974, 1995; Sunspots by Loughhead and Bray, 1979; and Solar Prominences with editors Vial and Engvold, 2015).

With the construction of spectroheliographs, filtergraphs, specialized narrow-band filters, and other instruments around the world, the search for the cause of sunspots became a broader search for the cause of the magnetic fields of active regions. The search for the cause of the magnetic fields of active regions remains a goal fundamental to understanding solar cycles.

3 The solar cycle, as known from early magnetograms

Continuing the picture of the previous paradigm are two important findings about the solar cycle that emerged from data taken with the new Mt. Wilson magnetograph designed and built by H. D. and H. W. Babcock. Following the numbering of key observations in Section 2, these findings are

6. The existence of a main dipolar field (Babcock and Babcock, 1952).
7. The reversal of the polarity of the solar polar fields at the maximum of each solar cycle (Babcock and Livingston, 1958; Babcock, 1959).

The polar field was detected with certainty in 1951 (Babcock and Babcock, 1952) using the first Mt. Wilson magnetograph. Earlier attempts to detect the global magnetic field of the Sun had been made with other instruments but definitive observations were still required. The polar fields were found to be irregular in shape. Their

overall area changed slowly over months, but their fine structure varied from day-to-day (Babcock and Babcock, 1952).

Large areas of magnetic fields of opposite polarity at the north and south poles slowly disappeared and were replaced by large areas of the opposite polarity magnetic field that had been migrating poleward. A surprising but slow reversal of the polarity of the polar magnetic fields took many months. The first observed reversal of the north and south polar fields was not simultaneous. The reversal of the north polar fields took place 1.1 years before the reversal occurred in the southern polar region (Babcock and Livingston, 1958; Babcock, 1959).

H. W. Babcock was well versed on the early knowledge of stellar magnetic fields as well as solar magnetic fields. Polarity reversals of the magnetic fields on stars had been found previously by Babcock (1947). His accumulation of knowledge became sufficient for him to confidently propose the first descriptive and semi-quantitative model of solar cycles (Babcock, 1961).

Babcock (1961) emphasized that only 1/2 of the 22-year solar cycle had been observed at that time. He said, "This attempt at a synthesis may be premature." Babcock knew that changes had to be taking place on scales smaller than what the magnetograph could detect. That idea led him to work on instrument changes that would allow him to view the Sun in higher resolution. He also knew tiny sunspots in small bipolar regions outnumbered those in large active regions. He observed that solar cycles began with small bipolar regions and that their numbers increased with time as did the occurrence of larger regions. Looking ahead, he stated, "Of particular value would be more and better measurements of the distribution and quantity of magnetic flux in BMR (bipolar magnetic regions) as a function of age."

In his original effort to explain the solar cycle, Babcock (1961) opens with, "shallow submerged lines of force of an initial axisymmetric dipolar field. . ." In his article, Babcock concluded that the active region magnetic fields could not be very deep for three reasons: a) the presence of more smaller, short-lived active regions than large long-lived ones, b) the migratory patterns of sunspot fields, and c) the replacement of fields of one polarity by opposite polarity fields. Babcock concluded that active region magnetic fields were not tied very deeply below the photosphere. He believed that they could float on the solar surface until opposite polarities merged or were "neutralized" and disappeared.

Since the time of Hale's observations of solar magnetic fields, research on solar cycles has spread as new solar telescopes and spectroheliographs were developed at solar observatories around the world. Scientists at Mt. Wilson continued to be the leaders. Below are additional research findings, largely by Mt. Wilson scientists, taking observations with the Mt. Wilson spectrographs and magnetograph. In addition to the numbered points above, Babcock (1961) listed more findings that he thought required explanation in any future model of the 11-year solar cycle. This list is in the same order as in Babcock's article. This list remains significant today. With more than 60 years of studies reviewed elsewhere (Petrovay, 2010, living review, on predicting solar cycles; Pesnell, 2012, solar cycle predictions, review; Cliver, 2014, extended solar cycle; Hathaway, 2015, review on solar cycles; and Martin, 2018, review on observational evidence of the depth at origin of active regions), these topics remain without definitive answers and important for further research:

8. The larger size and longer lifetime in the preceding (western) polarity sunspots than in the trailing (eastern) polarity sunspots in any solar cycle.
9. A forward tilt of the radial field of sunspots varying from 0.44° to 7.6° for sunspots of the oldest age.
10. The orientation of sunspot groups such that, on average, the leading (western) polarity in each hemisphere is closer to the equator than the trailing polarity.
11. The common recurrence of new sunspot groups at longitudes close to those of previous sunspots.
12. The bipolar magnetic regions underlying the development of sunspot groups and centers of activity (active regions).
13. The existence of hydrogen whirls which are groups of superpenumbral fibrils with a partial or complete spiral pattern around the periphery of many sunspots.
14. An ordered appearance for calcium flocculi, hydrogen flocculi, faculae, and sunspots. This sequence is the early evidence for concluding that active regions emerge from below the photosphere. It is largely supplanted now by more fundamental magnetic field data.

Item 13 warrants further explanation. Because sunspots were the first evidence of the existence of solar cycles, one would think that any feature closely related to sunspots might be a key to understanding solar cycles. Hale (1908) and Hale et al. (1919) were fascinated with H-alpha (hydrogen) images because the patterns of fine structure in active regions and around sunspots so closely resembled the lines of force around magnets on Earth. Daily observations were deemed essential for studying both short- and long-term changes in these solar structures. The images were recorded on glass plates using the Mt. Wilson spectroheliograph.

Richardson's (1941) long-term study of the hydrogen whirls from the daily observations turned up surprising results that seemed to belie the expected solar cycle properties specifically associated with hydrogen whirls. As first discussed by Hale (1908), the degrees of curvature in the clockwise and counterclockwise directions of the hydrogen whirls were not proportional to the magnetic fields of the associated sunspots nor was the curvature associated with sunspot polarity.

Both clockwise and counterclockwise patterns were sometimes observed on a different part of the same sunspot of single polarity. An example is shown in the middle panel of Figure 5 from Balasubramaniam et al. (2004). There is a low degree of opposite curvature on both the east (left) side and west (right) side of the sunspot. The pattern on the right side of the sunspot in Figure 5B corresponds to the curvature of the schematic whirl in the left panel in Figure 5A from Martin (1998). Conversely, the pattern on the left side of the sunspot in Figure 5B corresponds to the schematic whirl in the panel on the right side of Figure 5A. This correspondence, of the curvature of the whirls around the sunspot to the schematic drawing of the chirality of filaments, is better seen if one mentally rotates the schematic drawing by 180° relative to the sunspot.

Another new finding was a strong tendency for counterclockwise to be dominant in the northern hemisphere and clockwise in the southern hemisphere (Richardson, 1941). This finding of a hemispheric preference for hydrogen whirls is directly correlated with filaments as shown in Figure 5 and verified by Rust and Martin (1994), and Balasubramaniam et al. (2004).

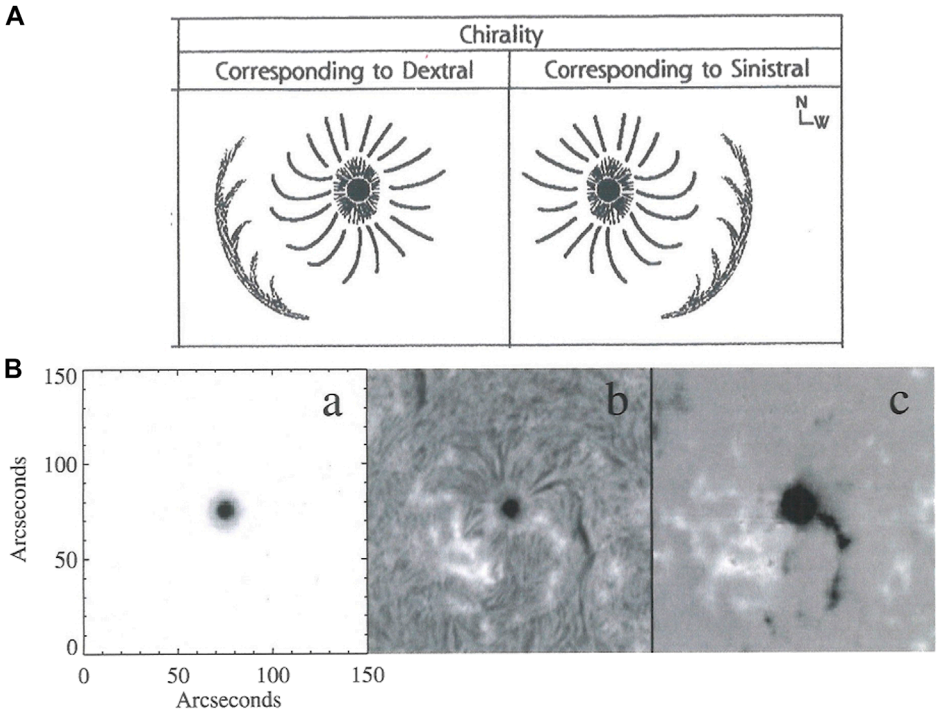


FIGURE 5
(A) Schematic drawing of hydrogen whirls revealed by sets of fibrils around sunspots. Dextral filaments are associated only with counterclockwise whirls as measured from the outer end of a fibril to the inner end of a sunspot. Sinistral filaments are associated only with clockwise whirls. The sign of whirls is independent of solar cycles (Richardson, 1941). Outside of the whirls around the sunspots, a dextral filament is schematically represented on the left and a sinistral filament on the right. Dextral and sinistral filaments and counterclockwise and clockwise whirls are related patterns of handedness known as chirality. Chiral patterns are mirror images of each other as illustrated here. **(B)** The dark feature to the right of the sunspot in the middle panel is a filament. Filaments closely follow the boundaries between areas of positive and negative polarity magnetic fields. White areas in the magnetogram on the right are positive polarity and black areas are negative polarity. Credits: **(A)** Martin, (1998, Figure 1). **(B)** Balasubramaniam et al. (2004, Figure 1).

The chirality of the whirls around sunspots is determined by the horizontal component of the magnetic fields of filaments not the strong vertical magnetic field of sunspots. This is why Richardson found no relationship between the whirls and the magnitude of the magnetic fields in sunspots. Because of this lack of relationship of the hydrogen whirls and the magnitude of sunspots, one might think that there is no relationship between chirality and the solar cycle. However, there is a correlation between the hemispheric preference of chirality and the hemispheric preference of the hydrogen whirls. Dextral filaments and filament channels are most common in the northern hemisphere as are counterclockwise hydrogen whirls.

The definition of counterclockwise and clockwise given here for sunspots is from the outer end of each superpenumbral fibril furthest from the sunspot to the end closest to the sunspot (Richardson, 1941; Rust and Martin, 1994). This definition is consistent with the direction of the mass flows in the superpenumbral fibrils, which are toward the associated sunspot. The superpenumbral fibrils within a whirl of fibrils appear to originate in the chromosphere 10,000–30,000 km beyond the outer boundaries of sunspots.

In some articles, the whirls around a sunspot are called "superpenumbral fibrils." The definition of the counterclockwise and clockwise direction along a fibril in a whirl in Balasubramaniam et al. (2004) is opposite to that by Richardson (1941). It is suggested that the Richardson's definition become the standard definition as applied herein. The majority of authors, to date, have chosen to use the Richardson definition.

Hydrogen whirls are one of the several patterns of chirality (handedness) in solar features that do not reverse their hemispheric pattern every 11 years (Martin, 1998). At the present time, there is only one chiral pattern among solar features that appears to be related to the origin of solar cycles. That is the chirality of filaments and filament channels discussed in Section 8.

What should be clarified here is that the property of chirality has a physical significance that is rare. Features that possess the magnetic property of chirality have strict one-to-one relationships to each other but not necessarily to other solar conditions or properties. Claims of the lack of a one-to-one relationship have been proven to be incorrect due to a lack of sufficiently high spatial resolution (Martin, 2015). Chiral relationships have no exceptions. However, high-resolution observations are required for filament threads to be clearly seen. For example, counterclockwise whirls are always related to dextral filaments as illustrated by Rust and Martin (1994). This is always true and independent of other relationships, such as to solar hemisphere. By contrast, chiral relationships to hemisphere are statistical. There are exceptions. For example, counterclockwise hydrogen whirls were found by Richardson (1941) to be most common in the northern hemisphere while clockwise hydrogen whirls were most common in the southern hemisphere.

Exceptions have consistently been found to this hemispheric trend, first by Richardson (1941), and then by Martin

(1998) and Balasubramaniam et al. (2004). We can identify relationships with exceptions as statistical relationships as distinct from one-to-one relationships which are invariable and linked to physical laws. This distinction is important because reliable predictions can be based on one-to-one relationships, whereas statistical relationships are inherently unreliable to some degree.

The reversal of the relative orientations of the positive and negative polarities of active regions with every new solar cycle is a statistical relationship. The polarities of individual active regions are sometimes reversed from the statistical tendency for most active regions to be either oriented east-to-west or west-to-east. There is a statistical tendency for up to 10% of individual active regions to be exceptions (Munoz-Jaramillo et al., 2022).

Babcock (1961) also tried to account for sufficient magnetic flux for large active regions. He followed Parker's (1955) idea of the possible existence of submerged, but buoyant, toroidal, magnetic flux ropes below the photosphere. Babcock proposed the wrapping of such flux ropes by differential rotation in a shallow layer beneath the photosphere. To him, this was the best possible mechanism to bring rising loops of the submerged flux ropes to the photosphere where they would present as the bipolar magnetic fields of active regions. An important element in Babcock's (1961) theory was finding the means of amplifying relatively weak magnetic fields of tens to hundreds of Gauss into fields of up to several thousand Gauss. Parker (1955) offered an amplification process that should occur in sunspots and Babcock made use of it in his grand scheme. This amplification is convective collapse, further investigated by Zwaan (1978) and others.

Multiple wrapping of subsurface magnetic fields around the Sun below the photosphere was originally proposed by Babcock (1961). His theoretical idea was to build subsurface flux ropes. The buoyancy of the flux ropes would allow small loops to rise and break through the photosphere to become active regions. It is now becoming evident that the processes of convection and reconnection, respectively, in the photosphere and outer layers of the solar atmosphere, are continuously reorganizing to achieve the simplest version of their dynamic state. The finding of supergranule-sized convection cells (Simon and Leighton, 1964; Leighton et al., 1962) in the outer one-third of the solar radius, along with increased knowledge of magnetic reconnection in different layers of the solar atmosphere, raises questions about the compatibility of these processes with the proposed long-term existence of the subsurface flux ropes (Martin, 2018).

Another process that locally amplifies magnetic flux has recently been identified by Chian et al. (2023). In observations from Hinode continuum intensity images and longitudinal magnetograms, it was seen that the magnetic field is intensified at the centers of two merging magnetic flux tubes trapped inside a vortex. The intensification had occurred during a 30-min interval during the vortex lifetime.

4 The extended solar cycle

At least three categories of observational studies are included in the concept of the extended solar cycle:

(a) Direct observations of small bipolar magnetic fields known as "ephemeral active regions" and
(b) EUV bright point data used as proxies for ephemeral active regions.

The second and third categories are only briefly mentioned in this review because an excellent review by Cliver (2014) on the extended solar cycle thoroughly discusses these important observations. These categories were first described by Wilson et al. (1988), who first used the term "extended solar cycle."

The extended solar cycle also harks back to, and further extends, Babcock's original lists of observations and results relevant to understanding solar cycles. Continuing with the numbering of these lists in Sections 2 and 3, I add more recent observations as an update to those cited by Babcock. The next item (15 below) is a fulfillment of Babcock's stated desire for "more and better measurements of the distribution and quantity of BMR (bipolar magnetic regions) as a function of age:"

In the first category (direct observations of new magnetic flux):

15. Magnetic field observations of small active regions, usually without sunspots, and called "ephemeral active regions" (Harvey and Martin, 1973; Harvey et al., 1975; Martin and Harvey, 1979).
16. EUV bright points used as proxies for ephemeral regions at high latitudes (McIntosh et al., 2014; McIntosh and Leamon, 2017).

In the secondary category (related long-term changes in magnetic flux previously emerged):

17. Long-term records of filaments from Meudon and Pic du Midi Solar Observatories (Mazumder et al., 2021), Kodaikanal Observatory, (Chaterjee et al., 2020; Priyadarchi et al., 2022), and the McIntosh database (Mazumder, 2019; Mazumder et al., 2021).
18. Vector magnetic field information on solar filaments (Leroy et al., 1984; Bommier et al., 1994).
19. Coronal features observed in the green coronal line (Altrock, 1988; Altrock, 1997; Robbrecht et al., 2010).
20. Variations (also known as torsional oscillations) in zonal flows, found in exacting measurements of differential rotation (Snodgrass, 1987).

In hindsight, a door to the new paradigm on solar cycles was first opened when a magnetograph was constructed in the early 1970s at Kitt Peak. It was designed to produce daily magnetograms of the full solar disk at a spatial resolution that exceeded other full-disk magnetographs of that time. Harvey and Martin (1973) initiated systematic studies of the surprisingly large numbers of very small bipolar regions that were revealed in the new magnetograms.

For their studies of ephemeral active regions, Harvey and Martin (1973) chose all new bipolar regions which were a) not present on the previous day, b) did not develop obvious sunspots, and c) could still be observed on the following day. The last criterion allows the observer to be certain that happenstance encounters of opposite polarity fields are not being included. They adopted the name "ephemeral active regions" first used by Dodson in the early 1950s (Harvey, 2000) for the small bipolar active regions. A few

such "ephemeral" regions had been designated among lists of larger active regions published in Solar Geophysical Data by the National Oceanic and Atmospheric Administration in the United States. Most ephemeral active regions survive 1 or 2 days. The designation "ephemeral", however, implied that these active regions were of no importance because they were too small and short-lived to develop significant sunspots or solar flares.

Harvey and Martin (1973), Harvey et al. (1975), and Martin and Harvey (1979), however, found that ephemeral regions provided information about the solar cycle not yet known from either sunspots or larger and long-lived active regions. Harvey and Martin (1973) found that the ephemeral regions a) were more broadly distributed in latitude than larger active regions, b) displayed a much broader range of orientations of the axes of their bipolar magnetic fields than larger active regions, and c) had a majority that possessed the west-to-east or east-to-west orientation of the solar cycle to which they belonged.

The latitude distribution of ephemeral active regions is so broad that the question has been often raised: do all ephemeral active regions belong to solar cycles, or is there a component of small bipolar regions that are independent of solar cycles? Harvey et al. (1975) answered this question by systematically studying every observable property of ephemeral regions without sunspots to see if they shared the same characteristics as the larger active regions. And they did. Ephemeral regions develop the same way as larger active regions with sunspots. The opposite polarities are first seen close together. The two polarities separate and move in opposite directions until they either merge with a magnetic field in the background with the same polarity or encounter a magnetic element of opposite polarity. If the latter occurs, small filaments can form, and erupt complete with miniature flares. Harvey et al. (1975) concluded that in every respect, ephemeral active regions are simply small active regions without enough magnetic flux to form sunspots.

The preferred orientation of ephemeral regions in belts of latitude alerted Martin and Harvey (1979) to recognize the presence of a new solar cycle 3–4 years before the traditional beginning of its sunspot phase and its continuation 1–2 years after the end of the sunspot phase. This extended the known duration of a complete solar cycle by 5–6 years; thus, the extended cycles showed that adjacent solar cycles overlapped in time. The overlapping years were around the sunspot minimum. Not surprising then, Martin and Harvey (1979) also found a secondary maximum in the number of ephemeral regions during the solar minimum.

Identifying ephemeral regions associated with two adjacent active region cycles was possible due to two findings:

(1) There were diminished numbers of ephemeral regions in latitudes between the current cycle and the new solar cycle.
(2) The orientation of a majority of the ephemeral regions, in the band at higher latitudes, differed by 180° on average from the band at lower latitudes, thereby matching the pattern of larger active regions expected in the same solar cycles.

This difference in the orientation at higher latitudes was recognizable for approximately 60% of the ephemeral regions. This bias was subsequently confirmed with the large statistical sample analyzed by Harvey (1993a).

The orientations of ephemeral regions have a large scatter, large enough to readily mask whether they follow Joy's law, the trend for the orientations of active regions to have greater tilts with increasing latitude and with smaller areas or magnetic flux. Qualitatively, ephemeral active regions are sometimes described as having a random component. This is an expected property for the smallest regions. It fits the statistical trend of the orientations versus areas of small newly emerging magnetic flux regions shown by Harvey (1993a).

Subsequently, however, Tlatov et al. (2010) isolated the subset of ephemeral regions that had the same east-to-west or west-to-east orientation as the cycle in which they occurred. Tlatov et al. (2010) could then show that ephemeral regions followed Joy's law in their study of cycles 21–23. The tilts of the ephemeral regions, and active regions, tend to decrease as their areas increase (or magnetic flux decreases) if they change at all (Tlatov et al., 2010).

Harvey (1993a), in her Figure 3, had also shown that the range of orientations for all active regions decreases with increasing areas of the regions. Ephemeral active regions also have been found in the polar regions (Jin et al., 2020). They have random orientations in these polar areas of unipolar network magnetic fields.

Ephemeral region studies were continued by Harvey (1993a), Harvey, (1994) by documenting impressively large statistical samples of over 10,000 ephemeral active regions. The ephemeral regions were recorded over parts of four 11-year solar cycles from magnetograms from the National Solar Observatory at Kitt Peak. This gigantic work yielded three types of definitive evidence that ephemeral regions belong to the same solar cycles as active regions with sunspots. These three types of evidence were

(1) The smooth extension of butterfly diagrams to higher latitudes as much as 3 years before the solar minimum. This is shown in Figures 6, 7A for solar cycles 19–22.
(2) The smooth fit of ephemeral regions within the power law distribution of active regions shown in Figure 7A [including new data from Hinode/SOT, combined with other data sets, the power law distribution in the size distribution of active regions was later extended over 5 decades (Parnell et al., 2009)].
(3) The centering of all sizes of active regions over a solar cycle around the same average latitude as shown in Figure 7B.

The graph in Figure 7B (top), shows that the population of ephemeral active regions (5–120 uH), at all times, is greater than the populations of active regions with small areas (120–315 uH) and active regions with large areas (315–3,550 uH). Conversely, the overall trend is for the active regions with the largest area to have the least numbers. This trend was observed visually by Babcock (1961) as mentioned in Section 3. One might then wonder, had Babcock known the strength of this trend established by Harvey (1993a), would he have proposed toroidal flux ropes for the origin of such a large population of small active regions?

Further thought should also be given to the pyramid-shaped graphs in Figure 7B. Harvey (1993a) shows that the populations of the active regions of all sizes are centered around the same average latitudes. It is surprising that there is little evidence of a second peak at the higher latitudes although the shape of the distribution is asymmetric in the sense of having higher numbers at

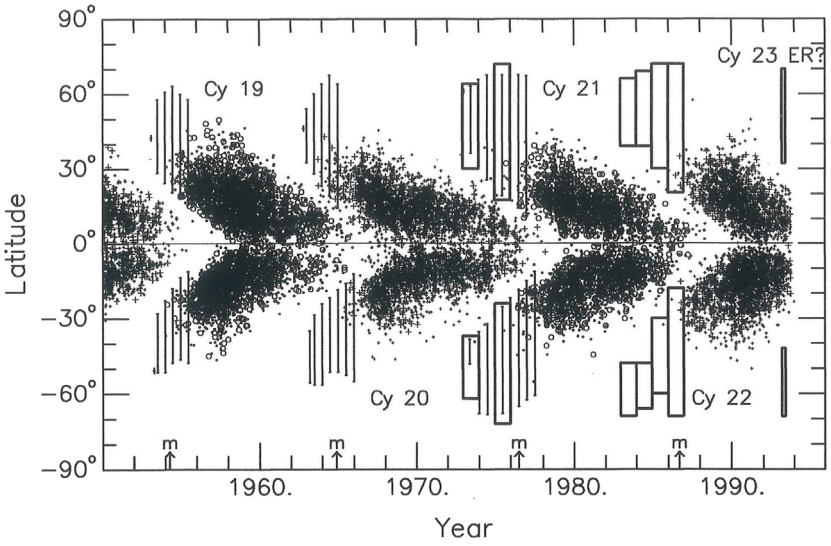


FIGURE 6
Butterfly diagrams of active regions with sunspots. Regions with sunspot areas >100 millionths of a hemisphere are shown alternately for successive cycles by + signs and small circles. The rectangular boxes indicate the latitude range of ephemeral regions with a preferential orientation reversed from lower-latitude regions. The light vertical bars represent small Ca II plage regions which are ephemeral regions seen in that spectral line. The last sunspot region belonging to a cycle is shown by a small circle superposed on a + sign. The times of solar minima (m) are shown as vertical arrows on the bottom line of the horizontal axis (credit: Figure 9 in Harvey, K. L. (1994) with permission to reproduce from Springer-Verlag).

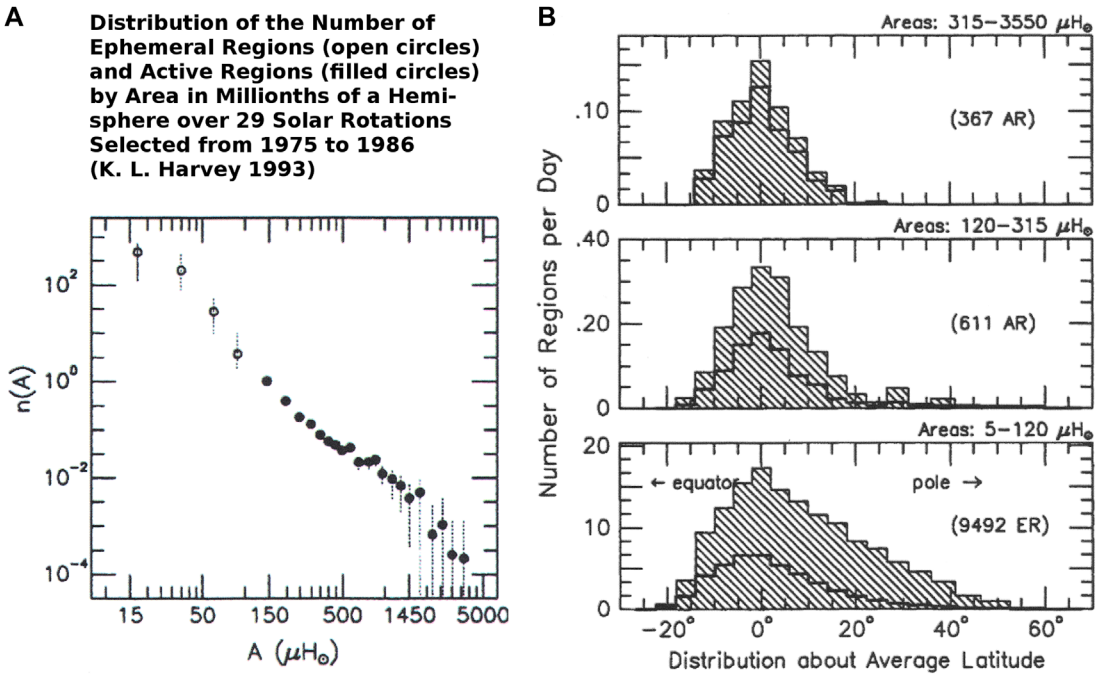


FIGURE 7
(A) This distribution shows that ephemeral active regions (open circles) fit smoothly into the same power law distribution as active regions (filled circles). **(B)** These graphs compare the populations of three sizes of active regions as shown in the upper right in millionths of a solar hemisphere. This panel illustrates the inverse relationship between the numbers of active regions and their areas; the larger the area, the smaller are their numbers, and conversely, the smaller the areas, the larger are their numbers. It also shows that all sizes of active regions over a solar cycle have their maximum number at the same average latitude. It confirms that the ephemeral regions constitute the largest fraction of the population of active regions (credit: Harvey, 1993a; Figure 5, with permission to reproduce from Springer-Verlag).

the higher latitudes where one might expect to see more evidence of the high latitude band of active regions with opposite east-to-west or west-to east bipolar magnetic orientations. Future studies of ephemeral active regions should seek to verify whether there is or is not a distinct minimum in the latitude distribution of the active regions whose average orientations differ by approximately 180°. The suggestion is to test whether the polarity reversal is inherent in the population of the apparent bands of active regions (as now assumed), or whether the polarity reversal could be acquired as active regions develop.

It should be noticed that the scales of the three graphs in Figure 7B are each different, in order to show them in an illustration of convenient size for publication. The scale of the bottom graph, with the ephemeral active region population, is two orders of magnitude greater than the top graph showing the population of large active regions. A graph on a linear instead of a logarithmic scale would be more effective in showing the importance of the large population of ephemeral active regions compared to the small population of active regions with sunspots.

As a loose analogy, the population of ephemeral active regions is like a mountain of sand and little rocks. The variable population of sunspots is like the seasonal snows on the peak of the mountain. Sunspots, like snow, have different physical characteristics than the whole mountain; they only appear when the conditions are favorable. Before and after the cool sunspots begin to appear, only the "sand and rocks" of the active regions are present—but, it is the snow that first captures our attention.

More than any other individual's research work, K. L. Harvey's research for her PhD thesis (1993 a,b) was the unseen harbinger of the paradigm shift discussed in detail in the next section.

In further investigating the extended solar cycle, McIntosh et al. (2014) recognized that EUV bright points could be useful proxies for ephemeral regions. This was reasonable because many EUV bright points occur when one magnetic pole of an ephemeral region collides with a patch of opposite polarity.

It is not expected that all observable small brightenings match one-to-one with ephemeral regions. Some of the brightenings might or might not be miniature solar flares. Like tiny flares, the EUV bright points occur where magnetic features of opposite polarity collide but there could be more than one bright point or flare per ephemeral region, and ephemeral regions that are not associated with any bright points. Furthermore, ephemeral regions are not the only possible source of new magnetic fields. Little clumps of intranetwork magnetic flux can be another source of opposite polarity magnetic field which collides with network fields and result in bright points. Thus, there should be no expectation for an exact one-to-one relationship between the numbers of ephemeral regions and bright points. Despite these questions about how well EUV bright points serve as proxies for ephemeral regions, the high latitude band of tiny EUV bright events appears to be a practical and meaningful representation of ephemeral regions at high latitudes.

We also should keep in mind there is no evidence that the smallest of the active region population has been detected to date. It is a reasonable speculation that magnetograms of higher spatial resolution over the whole solar disk will become available in the future.

5 The recognition and ramifications of two complete 22-year Hale Solar Cycles

Using the EUV bright points as proxies for ephemeral regions, McIntosh et al. (2014) showed, the Sun has two complete 22-year solar cycles present continuously. Each 22-year cycle exists in its own separate band of latitudes and has oppositely oriented bipolar magnetic fields from the previous cycle. Each cycle is a full 22 years, and a fraction more, in duration.

Arriving at 22 years via observations was significant because any shorter duration does not complete a cycle in time or space according to the definition of a cycle. As shown by Leamon et al. (2019), a true cycle is continuous; it does not have stops and starts. The beginning of every 22-year cycle is the same point in time as the end of the previous cycle. In this strict sense, the "sunspot cycle" is not a cycle. The occurrence of sunspots is a phase within a cycle rather than a true longer cycle by itself.

The findings of McIntosh et al. (2014) were amplified and extended in a series of articles (McIntosh and Leamon, 2015; Srivastava et al., 2018; Leamon et al., 2019; Leamon et al. 2020a; Leamon et al. 2020b; Leamon et al. 2020b; McIntosh et al., 2021; McIntosh et al., 2022). Together these research results dramatically revise our conception of solar cycles. Their work establishes the existence of at least six distinct properties of the two 22-year cycles that together overturn and replace the former concept of 11-year solar cycles. These properties, each seen in Figure 8, are stated below in more detail:

(1) Two 22-year solar cycles are present on the Sun continuously, instead of a single series of approximate 11-year cycles; this means there is no 11-year solar cycle; there is only a phase within each 22-year cycle in which a small fraction of active regions develop sufficient concentrations of magnetic flux to form sunspots.
(2) Each 22-year cycle consistently has its own west-to-east or east-to-west orientation of most of its bipolar active regions; there is no flipping of the dominant orientation of its active regions by 180° after the 22-year cycle begins.
(3) Each 22-year cycle begins at approximately 55° in latitude, not at lower latitudes.
(4) Each 22-year cycle continuously drifts toward the equator (as in the previous paradigm) but over 22 years instead of 11 years in each hemisphere.
(5) Every 11 years, a new 22-year solar cycle begins at high latitudes with the opposite west-to-east or east-to-west orientations of its active regions from those in the preceding cycle.
(6) When one cycle of active regions reaches the equator, the next cycle of active regions of the same polarity orientations of its active regions begins in a band of latitudes centered at approximately 55° that might be as wide as 30°. It is surmised that there is no gap in time between the end of a 22-year cycle of one polarity orientation close to the equator and the beginning of the next cycle with the same polarity orientation at approximately 55° in latitude. In other words, there is no breaking or interruption in successive cycles.

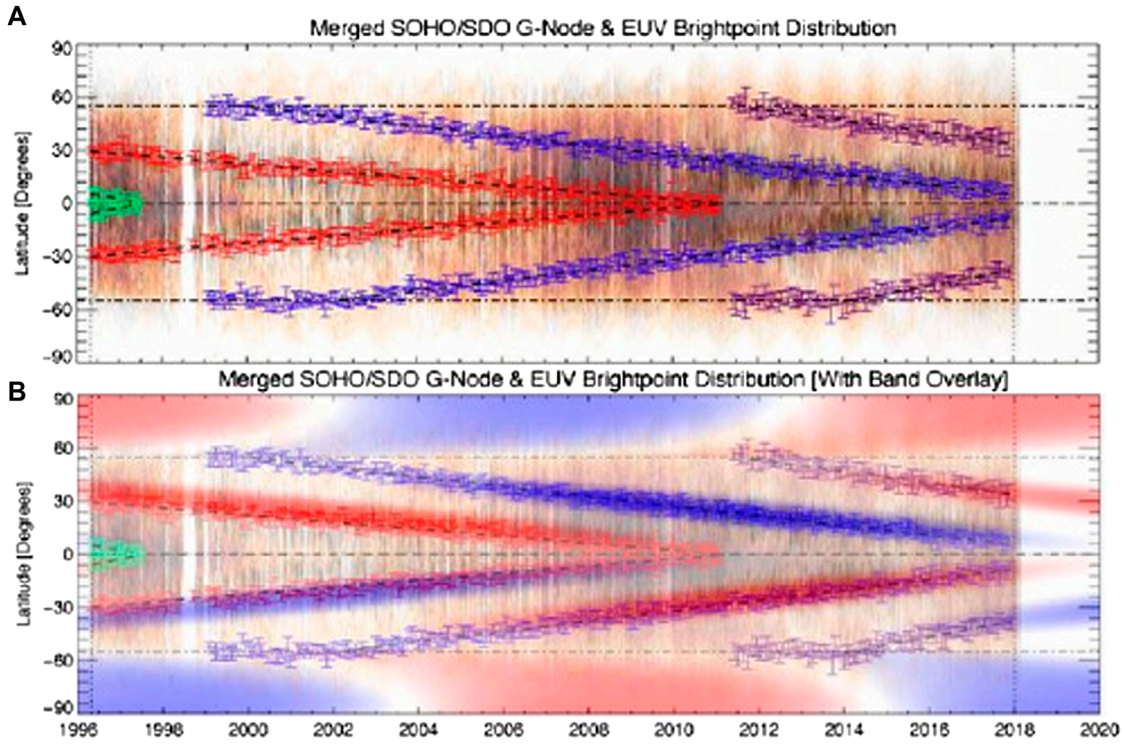


FIGURE 8
(A) Merged components of proxies, ephemeral regions, and active regions for solar cycle 22 in green, 23 in red, 24 in blue, and nearly half of cycle 25 in purple. Every solar cycle of active regions is represented by a chevron of 22 years duration and overlaps the preceding and following cycles by 11 years. Due to this overlap, consistent evidence of some part of the two separate cycles is present on the Sun at all times. This figure from Srivastava et al. (2018) adds color to chevrons whose width is a Gaussian distribution 10° wide in latitude. This approach to displaying the solar cycle data was used for showing the coronal bright points originally studied by McIntosh et al. (2014). Narrow bands of color are superposed to distinguish separate solar cycles. **(B)** A smoothed version of **(A)** with the dominant polarities of the polar regions also shown in alternate red and blue representing the positive and negative fields, respectively. The average polar fields alternate in polarity every 11 years consistent with a new solar cycle beginning in a new high latitude band every 11 years. In the active region latitudes, the northern hemispheric half of a chevron in the odd-numbered cycles are red in the sunspot zone; in the southern hemispheric half of the same chevron, the sunspot zone is blue. The main characteristics of the new paradigm: a) every solar cycle, represented by a chevron, is 22 years in duration; b) every 22-year solar cycle overlaps the preceding and following cycles by 11 years; c) every solar cycle begins at high latitudes in both hemispheres when the previous solar cycle, of the same polarity orientation, ends at the solar equator; d) successive new solar active regions, south of 55° in latitude, gradually appear closer to the solar equator over a 22-year interval (credit: Srivastava et al., 2018, Figure 3).

Properties in common with the former concept of 11-year solar cycles are

(a) The active regions of solar cycles exist in separate bands of latitudes not shared with the previous or following cycles.
(b) The change in the polarity of the network magnetic fields in the whole solar polar regions occurs around the time of the sunspot maximum, and 22 years are required for a complete cycle of magnetic fields in the polar regions.
(c) The change in the polarity of the polar fields is related to the largest active regions, not the whole population of active regions (Howard, 1992; Harvey, 1994). Multiple polarity changes can occur, but a main reversal in the dominant polarity occurs around the time of the solar maximum.
(d) The statistical properties of active regions within their belts of latitude are the same as previously known.

Figure 8A shows bright point proxies, ephemeral regions, and active regions for solar cycles 23 and 24 and nearly half of cycle 25. Cycle 23 is shown in red, cycle 24 in blue, and cycle 25 in a mixture of red and blue representing purple. The current solar cycle (25) has been present for 10 years at the time of this writing.

Figure 8B is a smoothed version of Figure 8A. The color-coding is changed to show polarity by color for positive-to-negative (west-to-east) polarities of active regions in the northern hemisphere (red) and negative-to-positive (east-to-west) polarities in the southern hemisphere (blue). This figure from Srivastava et al. (2018) superposes narrow bands of color on top of real data points. This attempt to emphasize the chevron patterns that represent each solar cycle tends to obscure the real data.

In Figure 8, at high latitudes, the polar fields are schematically represented as blue for the dominant positive polarity and red for the dominant negative polarity. Above approximately 55°, the background network magnetic fields gradually drift poleward. The continued action of differential rotation on filaments and filament channels gradually turns them nearly parallel with lines of constant latitude above approximately 55°. Some of the trailing polarity magnetic fields of the active regions move poleward faster than

equatorward. At the same time magnetic flux above 55° gradually disappears due to photospheric magnetic reconnection. As it moves poleward. Eventually all of the previous polarity disappears in the polar region and is replaced by along the polarity boundary the opposite polarity field moving poleward. This change of polarity has been interpreted as signaling the beginning of "the next solar cycle." However, in the new paradigm, the new cycle begins closely in time to the end of the previous 22-year cycle having the same polarity (Leamon et al., 2020a; Leamon et al., 2020b).

The lower panel is a smoothed version of the upper panel with the dominant polarities of the polar regions also shown in alternate red and blue, representing the positive and negative fields, respectively. The average polar fields alternate in polarity every 11 years consistent with a new solar cycle, having the opposite polarity, and beginning in a new high latitude band every 11 years. In the active region latitudes, the northern hemispheric half of a chevron in the odd-numbered cycles are red in the sunspot zone; in the southern hemispheric half of the same chevron, the sunspot zone is blue. This pattern is opposite in the even-numbered cycles. The main characteristics of the new paradigm are: a) Every solar cycle, represented by a chevron, is 22 years in duration. b) Every 22-year solar cycle overlaps the preceding and following cycles by 11 years. c) Every solar cycle begins at high latitudes in both hemispheres when the previous solar cycle of the same polarity orientation ends at the solar equator. d) Successive new solar active regions, south of 55° in latitude, gradually appear closer to the solar equator over a 22-year interval (credit: Srivastava et al., 2018, Figure 3).

Figure 9 (Altrock, 1997) has a similar chevron pattern for the green coronal line at 5303 A but is for cycles 20, 21, and 22. The green line serves as a proxy for the early phase of the 22-year cycles at high latitudes where ephemeral regions are difficult to detect. The end of cycle 20 occurs at approximately the end of 1978 with the last ephemeral regions of that cycle close to the equator. This same year

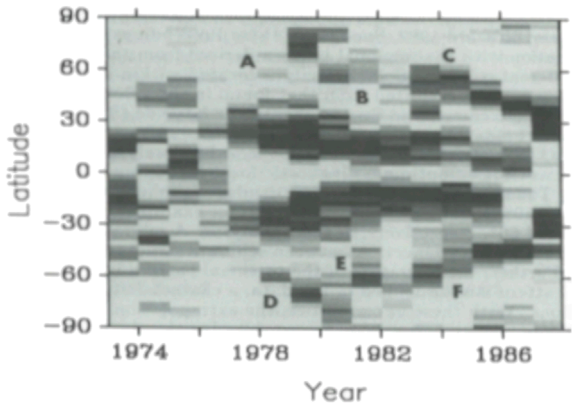


FIGURE 9
These chevrons represent the green coronal line emission for the end of solar cycle 20, most of cycle 21, a part of cycle 22, and the start of cycle 23 at high latitudes in 1985. The green line emission is a clearer signal of the beginning of the 22-year cycles than the tiny ephemeral region at high latitudes. However, the ephemeral regions near the equator are the clearest signal of the end of solar cycles. Combining information from both sets of data confirms the duration of 22 years as the fundamental length of solar cycles (credit: Altrock, 1988; also republished in Figure 13 by Harvey, 1992).

is the beginning of cycle 22 at high latitudes as revealed by the green line emission. This provides an excellent example of the concept of terminators analyzed by Leamon et al. (2018), Leamon et al. (2019), Leamon et al. (2020b).

Active regions that developed sunspots were 10% of the total population of active regions and ephemeral active regions analyzed by Harvey (1994). Without the ability to recognize a lower limit for the size or duration of ephemeral regions in current-day magnetograms, it is only known that the percentage of the population of all active regions with sunspots is much less than 10%.

The reduced emphasis on the sunspot phase that comes with more complete knowledge of 22-year solar cycles, also affects how we see the Maunder Minimum and other long-term Grand Solar Minima. The Maunder Minimum of approximately 70 years with few or no consecutive sunspots is only approximately three and a quarter 22-year cycles. The greater precision of 22-year solar cycles, over the concept of variable 11-year solar cycles, is a strong confirmation of the deductions of Beer et al. (1998) and McIntosh et al. (2014) who have suggested that solar cycles continued with few or no sunspots during the Grand Solar Minima. McIntosh et al. (2014) further proposed that evolution into and out of a Grand Minimum is feasible with changes in the spatial overlapping of cycles; the more the overlap, the less the spots and *vice versa*. This could happen if the migration rates to lower latitudes increase and produce less overlap or decrease and result in more overlap of the bands of activity, without changing the overall 22-year rhythm of the solar cycle.

Twenty-two years is again emphasized here because it is a significant threshold to meet for the theory of the duration of solar cycles. At this limit, more precisely 22.2 years, solar cycles are continuous in both polarity bands in each hemisphere without any gap between the end of one cycle at the equator and the beginning of the next cycle of the same polarity at approximately 55° in latitude. Successive cycles broken by gaps imply a stop and start mechanism while continuous repeating cycles indicate a continuous mechanism behind solar cycles. The overlapping of two cycles out of phase by half a cycle reinforces the concept of continuity of the 22-year cycles. If any longer than 22.2 years, the diagrams in Figure 8 show that there would be spatial interference with the next cycle's of magnetic fields with opposite polarity. In principle, the extended cycle concept comes to an end at 22.2 years. While this seems logical, it should be more thoroughly tested with observational data.

Leamon et al. (2018, Leamon et al., 2019, Leamon et al. 2020a, Leamon et al. 2020b) introduce the concept of a "terminator event," a time when solar cycles disappear at the equator concurrent with the appearance of the next 22-year solar cycle at approximately 55°. This concept meshes usefully into the magnetic cycle (Hale, Ellerman, Nicholson and Joy, 1919; Hale and Nicholson, 1925). It confirms the duration of 22 years as the fundamental length of solar cycles instead of approximately 11 years.

As discussed in McIntosh and Leamon (2017), the existence of two solar cycles on the Sun for a long duration does not preclude the possibility of the Sun (or a star) in the past or future from having three solar cycles or more concurrently present over a longer or shorter interval of time than its present cycle.

Leamon et al. (2019) and Srivastava et al. (2018) further proposed that the end of a solar cycle, a termination point at the equator, and the beginning of the next solar cycle at high latitudes

are causally linked in time. They showed that this linkage of the end and beginning of successive 22-year solar cycles could be as short as a single average 27-day rotation of the Sun. The finding of substantially greater precision in the recurrence of the 22-year cycle offers unprecedented possibilities for predicting future solar cycles and harmonics of the 22-year cycles.

The findings on solar cycles described herein have already been said to establish a new paradigm in the understanding of solar cycles (Srivastava et al., 2018). Previously, the 11-year solar cycle was regarded as the most fundamental cycle of all solar cycles. The longer 22-year cycle was thought to be due only to the 180° change in magnetic orientation of active regions and their sunspots near the beginning of each 11-year sunspot cycle. The new paradigm is consistent with seeing that the 22-year solar cycle is a more precise and fundamental "Hale Solar Cycle."

In hindsight, we can see that the former concept of 11-year solar cycles is an illusion that was created by thinking that active regions with sunspots were the only significant active regions on the Sun. A single linear series of these sunspots was inadvertently created by plotting only the sunspot phase of every 22-year solar cycles as a single linear series in both butterfly diagrams and linear plots of sunspot number versus time. This was done under the assumption that only a linear series existed. The vast numbers of active regions without sunspots were not known, surmised, or theoretically predicted. Four decades (1973–2014) of studies identified as "the extended solar cycle" gradually brought increasing evidence that these two assumptions were not true as demonstrated by McIntosh et al. (2014).

Again, in hindsight, the apparent change in magnetic polarity between each approximate 11-year sunspot cycle is due to the lack of recognizing the full, 11-year overlap in time of any two successive 22-year solar cycles that are always simultaneously present. In the old paradigm, we skipped back and forth between the sunspot phase of the two 22-year cycles and called that "the solar cycle." By inadvertently skipping from the sunspot half of a 22-year cycle to the sunspot half of the next 22-year cycle, it appeared as if an abrupt reversal of the orientation of active regions had taken place. Now, it is clear there is no abrupt change in polarity at the beginning of the sunspot cycle because the previous solar cycle has been present for 6–8 years. The previous paradigm was a type of illusion resulting from not yet including ephemeral active regions as playing a significant role in solar cycles. This, in turn, should not be regarded as human error; it was a natural conclusion from not yet having magnetograms of sufficiently high resolution to adequately detect the ephemeral active regions.

A more subtle point to mention is that 22-year cycles are more symmetric in time around the solar maximum than mentioned in the previous paradigm when the sunspot phase was synonymous with the concept of solar cycles. The rise of the sunspot phase of the cycle is well known to occur 3–4 years before the solar maximum and the decay phase to last 7–8 years after the solar maximum. By contrast, the whole solar cycle that includes the ephemeral regions tends to begin closer to 11 years before the solar maximum and to continue for 11 years after the solar maximum. This difference in the timing of the sunspot phase might be related to the observed fact that sunspots tend to grow larger and last longer after the solar maximum than before it. This is when the active region belt is centered closer to the equator. In addition, only a few complete 22-year cycles have been documented. Uncertainty remains about whether the sunspot number, the traditional way of determining the solar maximum, coincides with the maximum in the total magnetic flux of solar cycles.

6 The roles of ephemeral active regions and elementary bipoles in solar cycles

The change in the role of ephemeral active regions from insignificant to significant is one of the biggest upside–down turns in the new paradigm for solar cycles (Srivastava et al., 2018). For the understanding of solar cycles, ephemeral regions have emerged from trivial features to essential. As illustrated in Figure 8, ephemeral regions define the beginning, middle, and ending phases of the 22-year solar cycle (McIntosh et al., 2021). They are the primary means of tracing the early phase of the Hale Solar Cycle for at least 7–8 years before sunspots appear, and to continue during the late phase for 3–4 years after sunspots are no longer present (McIntosh et al., 2014). Furthermore, the greater the number of ephemeral regions, the greater the sunspot numbers (Harvey, 1994). Their vast numbers during the sunspot phase of the solar cycle (Harvey and Harvey, 1973; Harvey, 1993a) indicate they could become the primary means of predicting the amplitude and timing of future solar cycles (Leamon et al., 2019; McIntosh et al., 2021). In addition, they might be useful for predicting the amplitude and timing of the sunspot phase within the 22-year solar cycles (Srivastava et al. (2018).

The question "are ephemeral regions a separate population of phenomena from active regions that form sunspots?" has already been addressed in Section 4 by citing the studies of Harvey (1993a), Harvey and Zwaan (1993b). The fact that this question is still being asked warrants emphasizing it a second time. The three answers shown by Harvey are among the first observations that should be brought to the attention of anyone learning about solar cycles. These questions and their answers are of sufficient importance to be included in proposals for a new generation of higher quality and higher resolution magnetograms from space. The requirement has not gone away for the "newer and better observations of small bipolar magnetic regions" that Babcock sought in his pursuit of the theory of solar cycles. Recognition of the previously unknown properties of the two 22-year solar cycles is now justification for continuing to pursue Babcock's dream shared by the worldwide community of solar researchers.

With increasing observations of magnetic spot cycles on other stars, the understanding of solar cycles is more relevant than ever to the stellar cycle research community (Bondar and Katsova, 2020). It should be highly important to know that spot cycles on other stars might be accompanied by magnetic field components far beneath the highest resolution observations obtainable now. This would have bearing on interpreting Maunder Minimum–like intervals on Sun-like stars such as Eta Eridani (Metcalf et al., 2012). At the same time, the shallow depth of ephemeral regions makes observations of stellar surfaces more promising and intriguing.

A recent study of Kutsenko (2020) compares the rotation rates of ephemeral active regions and active regions. Kutsenko verifies the common understanding that the depths of ephemeral regions are

shallower than larger active regions and originate within 0.98 solar radii of the photosphere.

The graphs of Harvey as shown in Figures 7A, B demonstrate the major point made in an article by Beer et al. (1998). During the "grand Solar Minima," such as the famous Maunder Minimum (Eddy, 1976; Eddy, 1983), solar cycles are still expected to be present in the form of ephemeral regions when few or no sunspots are observed during these solar cycles. For further discussion on this topic, the readers are referred to McIntosh and Leamon (2015).

Although ephemeral regions are a major part of the solar cycles before, during, and after active regions develop sunspots, in another aspect of the solar cycle, they play no role. They do not play even a minor role in determining the polarity of the magnetic flux which reaches the poles. The articles of Harvey (1992, Harvey, 1994) call attention to the knowledge that ephemeral active regions and small active regions make no significant, long-term, net contribution to the network magnetic flux anywhere on the Sun. For small active regions, equal quantities of positive and negative magnetic flux disappear from the photosphere nearly *in situ*. Harvey concludes "… the net contribution of an ephemeral region to the unipolar fields is 0."

Harvey (1994) along with Webb et al. (1984) and Makarov and Sivaraman 1989a, Makarov and Sivaraman 1989b) show that the well-known rapid migration of the polar crown filaments toward the solar poles begins in concert with the appearance of sunspots. The migration does not start earlier in the solar cycle. Harvey's (1994) analysis reveals that the magnetic flux which reaches the polar regions, originates entirely or almost entirely from the largest active regions. Howard (1992) also argues that all unipolar large-scale magnetic patterns can be traced to the emergence of magnetic flux in active regions and nests of active regions. These articles also confirm that ephemeral regions have little or no importance in which magnetic flux reaches the poles.

These results, in the two paragraphs above, have led to a sharp contrast in how ephemeral regions and large active region contribute to the observed properties of solar cycles in very different ways.

A recent article on small bipoles in the polar regions by Jin et al. (2020) describes events they called "bipolar magnetic emergence." They were found at latitudes above 60° north and south of the equator. I treat these as ephemeral active regions because the authors have provided no information that justifies any separate designation. The significance of their study was finding over 300 such ephemeral active regions from the combination of observations of both the north and south polar regions. Some of the ephemeral active regions produced brightening events at 211 Å from the Atmospheric Imaging Assembly (AIA) during the year from June 2010 to May 2011. An important result was finding that the ephemeral regions in the polar areas have random orientations. The randomness of their orientations is not evidence that these small bipoles in the polar regions are not a component of solar cycles. A relatively large fraction of ephemeral regions in the active region belts also do not develop the west-to-east or east-to-west orientations of major active regions (Harvey, 1992; 1994).

The above result of Jin et al. (2020) is complementary to the findings of Tlatov et al. (2010) who separated ephemeral active regions into two groups. They called the smaller group Quiet Sun Regions (QSRs) and the larger ones Ephemeral Regions (ERs). Because their study relied on their program to identify bipoles, the smallest category was subject to including random background noise. However, Tlatov et al. (2010) found that the ER showed a preference in orientation depending on the polarity of the large-scale magnetic fields around and overlying the ephemeral regions. This statistical result is complementary to that of Jin et al. (2020). In the polar regions above 60° in latitude, there are no closed large-scale fields for ephemeral regions to align with. All magnetic reconnections send particles straight into the solar wind. Together, these two separate studies are consistent with the idea that most ephemeral regions have an overlying large-scale field to keep them from having random orientations. In any case, we should expect continuous interaction between newly emerging magnetic fields and any source of large-scale magnetic fields in which they emerge. An exception might be rare ephemeral regions that emerge already aligned with their local environmental magnetic field.

Chae et al. (2001) studied ephemeral regions around a large quiescent filament with the usual and essential overlying canopy of coronal loops. The spine of the tall filament that they studied is almost parallel with the polarity boundary at the photosphere except at the ends where the filament threads can splay to a wider pattern. The tall barbs curve from the spine to the chromosphere but are not resolved except near their footpoints. Because this is a tall, dextral, quiescent filament in the northern hemisphere, the angle of the barbs along the sides of the spine would be large and close to perpendicular (approximately 75–85°) to the long axis of the filament spine.

The filament erupted during the second day of the 5 days of observations and was recorded and analyzed. The filament spine, its barbs, and the overlying loops might have been already slowly rising and changing for hours to days preceding the eruption.

Chae et al. (2001) found a dominant direction for the ephemeral regions that did not apply to all of the ephemeral regions in the environment of the large quiescent filament. The dominant direction was similar to that of a cluster of ephemeral regions near the spine of the filament. That dominant direction did not match any specific features or direction that they thought could be related to ephemeral regions. However, in their Figure 1, as they had noted, the dominant orientation of the ephemeral regions was closer to being opposite to that of the magnetic field of the coronal loops overlying the filament than to any other feature of the whole erupting system. Being opposite in direction to the overlying loops means that the direction of the magnetic field in the barbs was inverse. It is now well known that the field direction in all barbs is inverse because the barb footpoints along the sides of the spine are small magnetic features opposite in polarity to the network magnetic fields on each side of the filament spine (Martin et al., 1994). The inverse magnetic field of the component of the filament magnetic fields perpendicular to the spine of the filament was first established for the vast majority of filaments by Leroy (1978). In their early observations of the filaments, the threads of the spine and barbs were not spatially resolved. Hence, the reason for the inverse magnetic field component perpendicular to the spine was a major puzzle for at least 1.5 decades.

Chae et al. (2001) did not offer any definitive explanation for this preferred direction of the bipoles. However, they did not consider whether the ephemeral regions could be aligned with the barbs of the filament before or during their eruption. The magnetic field in the barbs of filaments are along their threads and that direction is close

to being opposite to the direction of the overlying loops. Therefore, the dominant direction of the ephemeral region bipoles would be more closely aligned with the inverse direction of the filament barbs than any other feature. It should be remembered that this filament erupted after the first day of their observations when the magnetic field environment of the new ephemeral regions might have been in a major state of change.

More studies like that of Chae et al. (2001) should be done to see if in other cases, there is any evidence that ephemeral regions close to the spine could align with the inverse magnetic field of the barbs.

To date, the spatial resolution of full-disk magnetograms has not yet been high enough to determine if there is a lower limit to the area or magnetic flux of ephemeral active regions. The graphs in Section 4 by Harvey (1993a), conservatively reveal that the population of ephemeral regions is 90% of the total population of all active regions. However, the real percentage of the total population must be more than 90% because many ephemeral regions are more short-lived than 1 day and therefore not seen and counted in daily magnetograms. Magnetograms of high spatial resolution also reveal many new active regions forming in existing active regions.

The pattern of formation of ephemeral regions has been compared by Harvey et al. (1975) and Martin (1990) with the way the larger active regions develop. They were found to be the same. This is further substantial evidence that ephemeral active regions and larger active regions with sunspots belong to the same population of solar features. Both are formed from pairs of tiny knots of magnetic field herein called "elementary bipoles." The remainder of this section is devoted to clarifying the observed properties of elementary bipoles. They are the basic units of magnetic flux from which the active regions of solar cycles are observed to originate. At this point in our studies, ephemeral regions and elementary bipoles are the same manifestation of new emerging magnetic flux. An example of a large number elementary bipoles spread unusually broadly is shown in Figure 10 from Strous and Zwaan (2001). Their bipolar nature is better seen in movies than in single images as shown in Figure 10.

Vrabec (1974) was the first to acquire and show observations of sufficiently high quality, during intervals of good seeing, to clearly see streams of tiny bipoles, no larger than 2 arc seconds in diameter. He displayed them in movies made from a series of magnetograms taken at the San Fernando Observatory. He observed tiny bipoles successively separating into knots of opposite polarity. These knots of single polarity flowed in opposite directions to the sites where they coalesced and formed sunspots. He called the pattern of magnetic knots "MFIs" for "moving flows inward" meaning "into sunspots." He also saw moving flows outward, MFOs, from sunspots. The latter were the well-known radial outflows in the moats around sunspots first studied by Harvey and Harvey (1973) from the magnetograms taken at the Kitt Peak National Observatory. The speeds of the MFIs are comparable to MFOs. However, in MFIs, the dot-like positive and negative poles move directly away from each other, the same as in ephemeral active regions. In MFOs, the pairs of opposite polarity dots move in the same radial direction outward from the sunspot penumbrae without separating widely from each other.

Vrabec's observations revealed that sunspots were a second stage of magnetic flux increase initiated separate from the sites of emergence of the stream of very small-scale bipoles at the photosphere.

In H-alpha images, elementary bipoles are identifiable by the arch filament systems that are most clearly seen in H-alpha or calcium images (Bruzek, 1969; Frazier, 1972a).

Several high-resolution observations (of MFIs) were published after those of Vrabec (1974). However, the authors and peer reviewers did not initially recall or realize they were the same features that Vrabec had identified as MFIs. Martin (1990) called them "elementary bipoles," Strous and Zwaan (1999) identified them as "flux emergence events" and Bernasconi et al. (2002) used the name "moving dipolar features (MDFs)." All of these high-resolution observations revealed their role in forming sunspots in the same way as those recorded by Vrabec (1974).

Sunspot formation required exceeding a threshold of magnetic flux in the coalescing bipoles. This was evident from the number of elementary bipoles that streamed from the site of emergence because they appeared to have the same small size of 1–2 arc seconds in diameter found by Vrabec (1974). This separate process of sunspot formation was identified as "convective collapse" described by Zwaan (1978) and Spruit (1979). If there are not enough elementary bipoles, or they are not driven together with enough force, sunspots do not form. Instead, these elementary bipoles either cancel and disappear, when their separate poles collide with magnetic fields of opposite polarity, or they merge to form the remaining parts of the magnetic fields of active regions, or both at the same time (Martin, 1990). The evidence to date is that all active regions are built from one or more elementary bipoles, which is consistent with Martin (1990) calling them "the building blocks of active regions."

Martin (1990) made a selective study of elementary bipoles observed from birth at the Big Bear Solar Observatory using the former video magnetograph and H-alpha images under conditions of good seeing. Martin chose examples observed from the smallest active regions with a few elementary bipoles to successively larger ones with tens of elementary bipoles from three different sites. The following properties were reported in Martin (1990):

(1) Elementary bipoles are distinct and a characteristic sign of new magnetic flux appearing in both ephemeral active regions and active regions.
(2) The individual positive and negative elementary poles are usually detectable for tens of minutes to a few hours but there are large variations in their observable lifetimes due to interactions with neighboring elementary bipoles as well as to observing conditions and to instrumental sensitivity.
(3) In its simplest, earliest, and smallest form, a single active region is an elementary bipole that consists of only one positive and one negative juxtaposed magnetic knots that continuously move in opposite directions from their initial site. The separation speed is higher than other motions on the quiet Sun. Their separation speed is an essential identifier. Otherwise, basic elementary bipoles would be endlessly confused with small canceling magnetic features.
(4) If one of the poles of an elementary bipole collides with another magnetic feature of opposite polarity, it becomes a part of a canceling magnetic feature. A canceling magnetic feature is

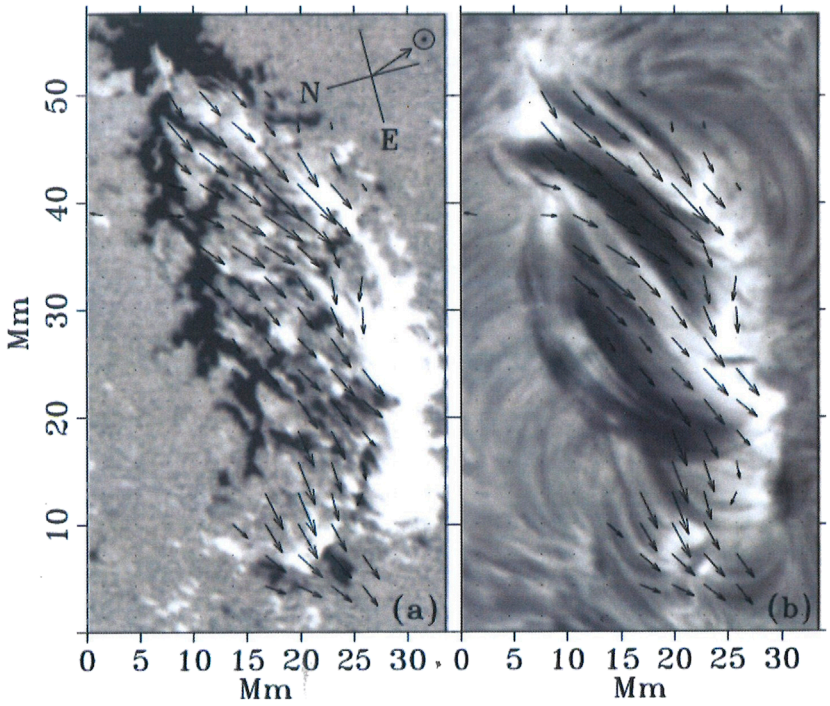


FIGURE 10
The image on the left is a line-of-sight magnetogram showing tiny white (positive) and black (negative) elementary bipoles between the main opposite polarity magnetic fields of a large, newly emerging active region. The initial pairing of the positive and negative elementary bipoles is not evident in a single image but can be seen in a time-lapse movie of this active region. The overlaid arrows show the direction of the magnetic field from a vector magnetogram. The image on the right reveals dark arch filaments from an H-alpha image. The arches connect the little bipoles that have separated from each other and exist at the footpoints of the arches. In the image on the right, the dark spot in the lower right is a sunspot being created by the merging of positive elementary bipoles (credit: Strous and Zwaan, 1999, Figure 6; by courtesy of L. Strous).

one in which equal and opposite polarity magnetic fields are disappearing along their common boundary. Due to canceling and merging, elementary bipoles have relatively short lifetimes from minutes to a few hours.

(5) Ephemeral regions that are larger than one elementary bipole often consist of a cluster of elementary bipoles developing within a few minutes of one another. Their motions can be parallel or in random directions from one another. Ones that are parallel or nearly parallel to each other will merge, combine their fluxes, and will then no longer be identifiable as elementary bipoles.

(6) The elementary bipoles can be seen before the arch filaments form above them in the chromosphere and low corona.

No fundamental difference has been found between elementary poles of short-lived ephemeral regions and active regions with sunspots (Martin, 1990). Many elementary bipoles cancel or merge with other elementary bipoles and become the other parts of active regions around their sunspots. The magnetic flux of elementary bipoles was estimated to be in the range of 200–600 Gauss (Martinez-Pillet et al. (1998).

Large active regions often form from multiple sites of elementary bipoles within or adjacent to a new active region (Martin, 1990). An example of a large number of elementary bipoles in a major new active region is shown in Figure 9 from Strous and Zwaan (1999). The spread of the elementary bipoles is much more extensive than in most active regions studied by Martin (1990). This is not surprising because many different centers of elementary bipoles can develop in medium and large active regions. For examples published to date, there is large variation in the number of clusters that form and in the direction they spread.

High-resolution, limited-field observations show that all active regions and their sunspots are built from elementary bipoles. Because the bipoles are not individually larger than a single ephemeral active region, Martin (2018) suggested all active regions might be surface phenomena which develop increased depth where their magnetic flux is highly concentrated.

The idea that sunspots are shallow phenomena forming at the photosphere and then extending to greater depths is consistent with the toroidal flows that develop as sunspots grow. They tend to become nearly round and symmetric if the magnetic fields of their associated active regions are simple and bipolar rather than complex. Such a large number of research articles have been written about sunspot formation, it is common knowledge that sunspots have downflows in their umbrae and upflows near the edges of their penumbrae throughout their lifetimes. The developed pattern of flows becomes like a toroid completely inside the surrounding sunspot moat where moving magnetic features flow outward. This inner toroidal part of the flows within sunspots is incompatible with merging into the pattern of rising intranetwork magnetic fields where Frazier (1972a, Frazier (1972b) suggested they form. Instead, the pattern of inner, downward toroidal flows is consistent with

their formation at the boundaries of and vertices of supergranules where downflows occur as observed and studied by Born (1974) and subsequently by many other authors.

Martin (2018) concluded that this cited collection of observations was strong evidence for sunspots growing wider and deeper as long as their site was fed by elementary bipoles, and convective collapse could continue. The observations are consistent and there is no known contradictory observational evidence to concluding that sunspots are shallow and develop from the photosphere and downward to greater depths (Martin, 2018).

Substantial evidence has been cited that ephemeral regions are shallow (Kutsenko (2020), that most sunspots are shallow (Martin, 2018), and that active regions form from tiny elementary bipoles no larger than ephemeral active regions (Vrabec 1974; Martin 1990, Strous and Zwaan (1999); Bernasconi et al., 2002). Because these observations show that essentially all active regions originate from small scale bipoles, it is reasonable to anticipate further findings confirming that the primary sources of new magnetic flux seen at the photosphere are shallow. Will we be looking more diligently in the future to the near-surface shear layer for a near-surface dynamo (Brandenburg, 2006), or even closer to the photosphere (Jarboe et al., 2017)? This shallow focus (Kutsenko, 2020) seems inescapable for theory to be relevant to the active region population consisting of two fundamental 22-year Hale Solar Cycles.

7 Candidate magnetic fields to replace the polar fields every 11 years

Observations over the whole Sun have shown that large areas of unipolar magnetic fields develop mostly from the spreading of the magnetic fields of active regions. Observations by Babcock and Babcock (1955, 1958) show that this diffusion of the magnetic fields results in systematically putting opposite polarities of magnetic fields in the two polar regions. Most fascinating has been Babcock's discovery that the polar fields gradually reverse polarity during the maximum of each sunspot cycle. The polar field disappears and is replaced by unipolar fields of the opposite polarity. Learning how these reversals take place has been a long-enduring observational and modeling problem. Significant results have been achieved.

Differential rotation and the random walk of the network of magnetic fields are two of the main processes that have been proven to result in the redistribution of the magnetic fields of active regions over the whole Sun (Wang et al., 1989; 1991). Meridional circulation, although slow, was shown to be another relevant factor (Wang and Sheeley, 1994). The role of the random walk caused by solar convection was modeled by Leighton (1964).

The solar cycle theories of Babcock (1961) and Leighton (1969) initiate and develop the concept that the tilt of active regions with respect to the solar equator is the main factor in determining the magnetic fields that reach the poles. For this reason, numerous articles have been written on the statistical properties of the tilt of active regions. This role has been included in recent reviews (Cliver, 2014; Hathaway, 2015; Weber et al., 2023). There is no need to reiterate the role of tilt in this review.

It should be mentioned here that the topic of flux cancellation has been introduced into the modeling of the transport of magnetic flux as an assumption. Cancellation was not included in the solar cycle theories of Babcock (1961) or Leighton (1969) except it was assumed that was how the polar fields disappeared. Now cancellation is mentioned in an off-hand way as a means of also eliminating magnetic flux from opposite hemispheres at the solar equator (Petrie, 2022 and references therein). With the thoroughness of the modeling of strong and weak magnetic flux (Petrie, 2022), the global aspects of the solar cycle seem to be thoroughly observed and modeled.

Observationally, however, there are still questions about the details of how the transfer of magnetic flux takes place from the active region belt to the polar regions. The merging and cancellation, discussed in Section 3, obscures the identity of specific sources of magnetic fields in relatively short intervals of time from hours to a few days. Under these circumstances, tracking and maintaining the identity of all sources of change is exceedingly difficult.

A current requirement is to recognize additional important observational factors, probably some not known to Babcock or Leighton, which affect magnetic field migration. A question not often addressed sufficiently is, *why does the random walk not result in a continuous mixing of opposite polarity network and intranetwork magnetic fields?* Differential rotation and meridional circulation would contribute only a bit more, if any, to the mixing of opposite polarities. As the active regions visibly spread over time, new active regions continuously develop within the older spread-out remnants. This factor contributes significantly to the mixing of opposite polarity magnetic fields. Additionally, the population of the small ephemeral active regions is large and broadly distributed over the whole Sun, which includes the polar regions (Jin et al., 2020). Furthermore, tiny intranetwork magnetic fields that depart from their convection cells by flowing into the lanes between supergranules. They are yet another source of mixing of opposite polarity magnetic fields. In the narrow lanes between supergranules, small intranetwork magnetic fields are observed to interact with any encountered magnetic feature of opposite polarity occupying the same lane. With all three of these large factors contributing simultaneously to the mixing of magnetic fields of opposite polarity (Martin, 1988). What are the opposing factors that create the *unipolar* network magnetic fields? How do unipolar areas of network magnetic fields sustain and increase in area over time?

Observationally, the answer is straightforward. Two solar plasma rules that apply to the line-of-sight component of magnetic fields on the Sun are now well known. These are:

(a) When opposite polarities come in contact, they are observed to cancel (Livi et al., 1985; Martin et al., 1985; Martin, 1998, and hundreds of subsequent articles by other authors). "Cancel" in the cited articles means only that equal amounts of positive and negative magnetic fields are observed to disappear at a common boundary between them (Martin, 1988; Martin, 1990). This is an observational definition; not a proposed mechanism on how magnetic flux disappears.

(b) When the same polarity features migrate into contact, they are observed to merge without any loss in magnetic flux (Martin et al., 1985; Livi et al., 1985, and numerous other previous and subsequent articles that mention this factor).

The ubiquity of canceling magnetic features was recognized first from video magnetograms taken at the Big Bear Solar Observatory. This does not imply that the observed flux loss from the photosphere

was not previously observed and considered seriously, especially at sites where solar flares occurred.

The Big Bear video magnetograms were unique at that time because the observer could select the integration time of pairs of images acquired at the video rate (15 frames per second for each pair of the two polarities acquired sequentially). The integration of 1,024 video pairs was especially good for enhancing the visibility of the magnetic fields on the "quiet" Sun, which was then revealed to be highly dynamic. These observations were not very popular because the active region magnetic fields were saturated beyond any practical use. After magnetograms from the Solar Dynamics Observatory became available, numerous studies included observations of canceling magnetic fields.

The interpretation of cancellation, that best fits the observations, was established to be magnetic reconnection at or very close to the photosphere (Litvinenko and Martin, 1999; Litvinenko and Martin, 1999; Litvinenko et al., 2007). There are other interpretations, such as submergence of the canceling magnetic fields from the top of the photosphere to an unknown depth beneath the photosphere. Differing interpretations are the reason for keeping observational definitions and their interpretations separate and clear.

Evidence that magnetic reconnection in canceling magnetic features occurs very close to or at the photosphere is:

(1) the lack of any sudden brightening in the chromosphere or corona concurrent with the disappearance of both polarities of the magnetic field
(2) downflows in the photosphere observed at cancellation sites
(3) concurrent transverse flows observed in the chromosphere simultaneously with the downflows in (2) above.

The disappearance of the line-of-sight component of the magnetic field is evidence that the magnetic field has changed configuration to such an extent that its line-of-sight component is no longer being observed. Where does the disappearing magnetic field find a new residence? Some authors have provided evidence that all or part of the magnetic field becomes submerged below the photosphere (Harvey et al. XXXX). However, concurrent transverse flows (3 above) are also invariably found along the polarity boundaries in the chromosphere emanating from the sites of reconnection at the photosphere. These simultaneous downflows and outflows (2 and 3), away from observable sites of cancellation, are strong evidence of magnetic reconnection versus complete submergence. These flows are the initial dynamics of filament channel formation along boundaries between opposite polarity magnetic field components (Livi et al., 1985; Martin et al., 2012; Martin, 2015). Filaments are not sustained when the cancellation ceases (example in Litvinenko and Martin, 1999). Further confirming the association between canceling magnetic fields and filament formation, (Wood and Martens, 2003) have shown that the magnetic flux that developed during the formation of a filament is roughly equivalent to the magnetic flux that disappeared during the associated cancellation beneath the forming filament.

While cancellation is correlated with many types of solar features and events in different ways, the relationship of cancellation to magnetic field migration is the aspect relevant to this review.

When cancellation occurs [magnetic field rule (a) mentioned above], the migration of the magnetic fields, due to the random walk, is quickly reduced to the slow speed of the observed disappearance.

Filament channels and filaments occur only along boundaries between areas of opposite polarities where magnetic fields are canceling (Martin et al., 1985; Martin et al., 1994). They are major markers of where the local slowing of the random walk occurs, but not the reason why the migration nearly stops.

Anderson and Martin (2005) have shown canceling magnetic fields along filament channels effectively retard the intermingling of magnetic network flux of opposite polarity. The network magnetic fields of either polarity can only move along narrow boundaries between supergranules (Martin, 1994). A single cancellation of two fragments of opposite polarity magnetic fields is sufficient to block one lane. The lanes on average are roughly 30 Mm apart, the average diameter of supergranules. They encircle all convection cells but the flows along the lanes are temporarily interrupted at the vertices between cells. A few cancellation sites are sufficient to block migration across a filament channel 100 Mm in length. Tens of canceling magnetic fields are all that is required to block all intermingling of opposite polarity magnetic fields over a filament channel several hundreds of megameters in length. This is why cancellation is the dominant observed factor in keeping areas of opposite polarity network magnetic fields from mixing across filament channels throughout the quiet Sun and in active regions.

Cancellation along polarity boundaries is the primary dynamic that increases and sustains areas of unipolar network magnetic fields that originate from new bipolar active regions. In these ways, filament channels are the gate keepers of flux migration. The emphasis here is on filament channels rather than filaments because filament channel formation, related to cancellation, does not stop when the filaments, above their filament channels, erupt with CMEs.

The plasma magnetic field rule (b), merging of same polarity features, plays a role more often recognized as assisting in the sustaining and increasing of the area of unipolar magnetic fields. Mergers, like canceling magnetic fields, occur between magnetic fields of the same polarity irrespective of their origin (Martin, 1988). Common examples of mergers are observed among the elementary bipoles within growing active regions between one pole of ephemeral regions and network magnetic fields. They are observed in moving magnetic features after they flow away from developed sunspots (Rust, 1968; Harvey and Harvey, 1973). The dynamics of merging is simply the joining of areas of the same polarity until the initially separate fields of the same polarity are no longer identifiable as separate features. Quantitatively, a merger is the sum of two or more observed fragments of magnetic field before they merge. Areas of single polarity network magnetic fields away from filaments are places where the magnetic field migration, due to transport by random walk, is operating at its best.

Although mergers help maintain and increase areas of unipolar network magnetic fields, they are not permanent. If the splitting apart of same polarity network fields did not also occur, the random walk, driven mostly by convection, might not occur. It is observed that clumpiness of network magnetic fields is related to network fields dwelling temporarily at the vertices between supergranules. These bits of magnetic flux are caused to move again, by forces associated with growth of new convection cells or other dynamics of new or old active regions of all sizes.

While canceling sites nearly stop the local migration due to the random walk and block the intermingling of opposite polarity

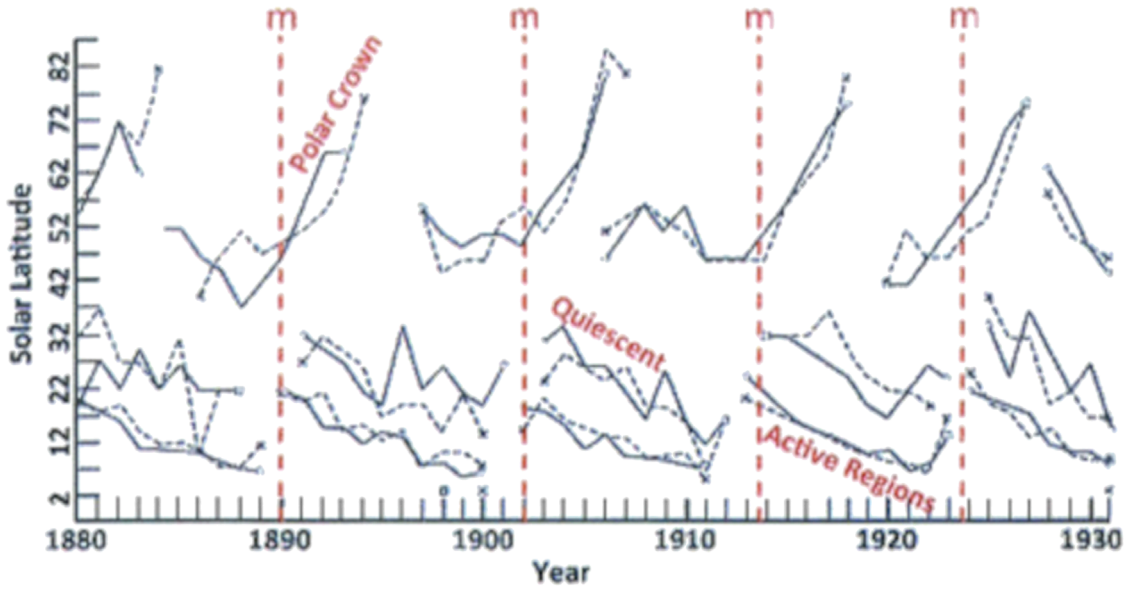


FIGURE 11
Two 22-year solar cycles are continuously evident in this presentation of filament data by Bocchino (1933). The 22-year patterns become clearer in mind if one overlooks the gaps representing the 11-year intervals that were formerly thought to be the duration of solar cycles. Two sets of filament tracks move equatorward every 11 years. The lower latitude track is an apparent continuation of the quiescent filament track during the preceding 11 years. This is evidence that filaments have retained the evidence of the two simultaneous solar cycles in separate latitude bands (Credit: from Cliver (2014) that was adapted from a figure in Bocchino (1933).

network magnetic fields, they do not stop differential rotation from acting on filament channels.

In active regions with east-to-west or west-to-east orientations, filaments and their channels are initially aligned nearly north-to-south or south-to-north but with local changes in the direction when new active regions form near them. Differential rotation gradually decreases the north-to-south and south-to-north component and increases the east-to-west or west-to-east components such that these filaments become diagonal features. These filaments, but not all filaments, develop southwest to northeast orientations in the northern hemisphere, and northwest to southeast orientations in the southern hemisphere.

If the filament channels survive the diagonal phase without disappearing or merging with other filament channels, they eventually become oriented nearly west-to-east or east-to-west, almost parallel with lines of constant latitude. This is most clearly seen on the poleward side of the active region belts. Filament channels and their canceling magnetic fields serve like shepherd dogs of differential rotation. As the channels move, they put up filaments as convenient flags of where magnetic fields are canceling along boundaries between opposite polarity magnetic fields. This enables us to record the higher contrast filaments, which over days to months and years, provide a record of where the channels have been. Filaments are not always benign flags revealing the changes ordered by differential rotation. They erupt frequently but new ones form in nearly the same locations as those that have erupted from their filament channels. Using the observations of filaments collected together from several observatories, Bocchini created an amazingly clear diagram revealing the two 22-year solar cycles on the Sun. This diagram adapted by Cliver (2014) is shown in Figure 11.

Solar cycles are more often represented by newly emerging flux rather than by filaments or filament channels that are commonly thought to be only related to decaying magnetic fields. Yet, the changing polar magnetic fields, related to the migration of filament channels, have long been known to have an important role in solar cycles (Leroy et al., 1983; Leroy, and Noens, 1983; Leroy, 1989).

Active regions show a tendency for their leading polarity to be closer to the equator than to their trailing polarity. This is known as "tilt." Joy's law is the tendency for tilt to increase with latitude and with decreasing magnetic flux in active regions. However, the scatter in tilt also increases with latitude as the average sizes of active regions decrease with latitude. Considering the many factors that affect the evolution of active regions, it is not clear why Joy's law persists. Joy's tendency is only a few degrees but it has been considered highly significant. It has been the topic of many statistical studies following Babcock's (1961) and Leighton's (1969) suggestions that the tilt is the main factor in why the trailing polarity is the one that first reaches the pole during every 11-year sunspot cycle.

Attributing significance to the tilt of the axis of bipolar active regions is puzzling. The lessening of tilt over time is an observed trend in major active regions (Weart, 1970). Additionally, the largest active regions show the least tilt. Moreover, Howard (1992) and Harvey (1994) have each shown that essentially all of the magnetic flux that reaches the solar polar regions is due to the largest active regions and complexes of active regions. Considering these observations, necessarily calls into question the assumption that tilt is the main factor in why the trailing polarity is the one that dominantly reaches the pole during every 11-year sunspot cycle.

An important result about filaments by Tang (1987) helps offer an alternative reason why a major portion of the trailing polarities of active regions reach the pole before the leading polarity. While

it was already known that filaments can form along any boundary between opposite polarity magnetic fields (Babcock and Babcock, 1955; Smith et al., 1965; Tandberg-Hanssen, 1974), Tang (1987) study had revealed that twice as many filaments form at polarity boundaries between active regions than within the center of active regions between their two main areas of opposite polarity.

In continuing this explanation of selective magnetic field migration, we will call the filaments that form within active regions "W filaments" and those that form between active regions "B filaments." This division of filaments into two main categories was already noted by Babcock and Babcock (1955), Smith et al., 1965) and Tandberg-Hanssen (1974). In the active region belts in each hemisphere and during periods of the formation of many new large active regions, there are often series of filaments in the middle of the active region belt alternating like: W B W B W B W B…. To be more specific, if we add the magnetic polarity designation of (+) for the western polarity, (−) for the eastern magnetic polarity in the northern hemisphere, the series of W and B filaments would look like −W + B −W + B − W + B − W + B ….

To develop, the B filaments can only form and become long-lived structures, when magnetic reconnection in the corona first connects some of the negative (−) trailing polarity magnetic fields of the active regions to the west with the (+) polarity of the active regions to the east. Then, the B fields can be represented more completely as +B− and the whole series can be represented as −W+ +B− −W+ +B− −W+ +B− −W+ +B− …. Relative to the W filaments, the (+) polarity is the leading or western polarity and (−) is the following or eastern polarity. But, relative to the B filaments, the (−) negative polarity is the leading or western polarity and the (+) polarity is the trailing or eastern polarity.

Differential rotation, in the northern hemisphere, acts on all of these magnetic fields such that W filaments and B filaments become diagonal. If north is up and west is to the right, and diagonal lines represent the filaments, the series can be depicted as −\+ +\− −\+ +\− −\+ +\− −\+ +\− …. This depiction should make it obvious that the B filaments have acquired their magnetic fields from some of the preceding and following active regions as they spread. It is commonly observed that the B filaments only form when the spreading of opposite polarity magnetic fields come in contact and begin canceling.

Relative to the W filaments in active regions in the northern hemisphere, differential rotation gradually adds a north–south component as indicated by the symbols (−\+). This tends to put the leading (+) polarity more to the north and the (−) polarity more toward the south in active regions in the northern hemisphere. However, we know from observations that it is the (−) polarity [not the (+) polarity] that first reaches the north pole for the sample solar cycle being represented here.

Next, we note the same effect of differential rotation acting on the B filaments that is represented by (+\−). In this case, differential rotation has pushed the (−) polarity more northward toward the pole and the (+) polarity more southward. Therefore, as differential rotation continues, the set of B filaments (+\−) are favored strongly to next contribute to the polar fields and the W filaments (−\+) might be the ones more likely to migrate toward the equator as shown in Figure 11. This preferential motion of a set of filaments toward the poles and two sets toward the equator are shown in the graph of long-term filament evolution in Figure 11 (Figure 6 in Cliver, 2014, which he adapted from Bocchino, 1933).

For the southern hemisphere, the entire pattern is reversed. Therefore, in the southern hemisphere, the next polarity reversal will result from differential rotation acting on the B filament channels and their surrounding fields in the same way as is for the northern hemisphere. This favors the (+) polarity fields to migrate toward the south pole as expected.

The choice of which polarity of magnetic fields arrives at the poles during every solar maximum is not as simple as described above. This is because some active regions develop roughly north or south of other active regions. In these cases, B filaments can form quickly with nearly east-to-west or west-to-east orientations. Differential rotation has little effect in changing the direction of these filament channels. However, these channels can slowly move toward areas of lower magnetic flux density. That direction is usually toward the poles. This shows that these filament channels can also be a determining factor in deciding which polarity reaches the poles first. The origin of subpolar crown filaments and their channels have to be studied to see how their channels quantitatively contribute to the role of herding specific polarity magnetic fields toward the poles. In the same hemisphere, some filament orientations would favor (+) magnetic flux reaching the poles in some solar circumstances and the (−) polarity magnetic fields in other circumstances. These filaments are consistent with our knowledge that one or more multiple polar reversals are sometimes observed before a main polarity is established in the polar regions.

An important mechanism of the formation of B filaments was first put forth by Hansen and Hansen (1975). They describe the formation of a nearly horizontal quiescent B filament, and its filament channel, forming between opposite polarity magnetic fields. They give the specific example of a new filament channel, and its filament, and its overlying coronal loop system forming between the polar magnetic field and the opposite polarity magnetic fields of one pole of an active region in the active region belt in the northern hemisphere. This is the same mechanism of formation of filament channels and filaments which applies to all B filaments. Physically, all filaments form by the same mechanisms (Martin et al., 2012). However, the solar circumstances, that precede and accompany their formation, are different for W filaments in active regions and B filaments that form between active regions.

The conclusion from the observations cited are

(1) It is unlikely that the tilt of active regions plays a significant role in determining what magnetic field is transported to the poles during the rise to solar maximum; differential rotation would tend to guide the leading polarity westward and poleward and the following polarity westward and southward. This would tend to effectively remove the tilt of active regions most commonly oriented east-to-west or west-to-east.
(2) Filament channels, that form under the coronal loop systems between large active regions (or decaying active regions), are the prime candidates for allowing differential rotation to guide them, and the newly associated part of the trailing polarity of the preceding active region, westward and poleward.
(3) In specific solar circumstances, poleward of the active region belts, other factors might determine, or contribute toward

determining, which polarity will reach the poles, among these factors are: the action of canceling and merging of magnetic fields, filament formation, differential rotation, random walk, and meridional circulation. These circumstances could account for some cases in which the leading polarity migrates poleward instead of only the trailing polarity.

The qualitative picture, derived here from combining the results in published studies, requires verification. What is required in the future are more detailed quantitative studies of the evolution of solar magnetic fields than have yet been made or modeled. This especially includes the role of canceling magnetic fields, filament channel formation, and differential rotation. Such studies should be made over long intervals of time during the sunspot phase and the non-sunspot phase of the 22-year solar cycles.

The identity of the origin magnetic fields is lost rather quickly due to canceling and merging. It is highly desirable for studies to be done by visually identifying and tracking changes in addition to by attempting to follow the identity of magnetic features by computer programs alone. Models can be adjusted to appear to correctly fit a solar situation involving multiple parameters, even without including all of the essential observed parameters, such as canceling features. These can contribute to an observed pattern in some solar circumstances, and not in others.

8 Initiation of the polarity reversal of active regions at the beginning of 22-year Hale Solar Cycles

In this section, we explore the literature relevant to the question: why do active regions change polarity at the beginning of each 22-year solar cycle? One might first also ask, "are there any other phenomena in addition to active regions which reverse their west-to-east or east-to-west orientation every 11 years?"

In this author's awareness, there is only one group of closely related phenomena that yields a positive answer to this question. Well-documented in the literature, these phenomena are polar crown filaments, subpolar crown filaments, and their filament channels (Rust, 1967; Leroy, 1978; Leroy et al., 1983; Bommier et al., 1985; Leroy, 1988; Martin, 1990; Bommier et al., 2021). Understanding how and why the polar crown filaments and subpolar crown filaments develop opposite east-to-west and west-to-east directions every 11 years requires at least a minimal understanding of both their evolution and the creation of the direction of the magnetic fields along filament channels and filaments.

In Figure 12, provided by V. Bommier, the several polar crown filaments in the northern hemisphere are along the polarity reversal boundary designated by the thin green line. This boundary was formerly often called "the neutral line" but the true meaning is that the magnetic field is horizontal along the polarity boundary. The red arrows represent the magnitudes and directions of the magnetic fields at a point on the filament at the base of the arrow. The arrows are in opposite directions in the polar crown filament and subpolar crown filaments. Specifically, in the northern hemisphere, the direction of the magnetic fields in the polar crown filaments are west-to-east and the subpolar crown filaments are east-to-west.

The opposing east-to-west or west-to-east directions of the polar crown and subpolar crown filaments were initially recognized in articles by Rust (1967), Leroy et al. (1983), Bommier et al. (1985), and Martin (1990). Reproduced in Figure 12 is a synoptic chart from the Meudon Observatory with the directions of the magnetic field superposed by Bommier et al. (2021). These synoptic charts refine and summarize the original work in Bommier et al. (1985) and references therein. Figure 12 shows the magnetic field directions for the polar crown and subpolar crown filaments for solar rotation no. 1697 (August 1980) during solar cycle 21.

The synoptic chart in Figure 12 is during the interval of increased rate of migration of the polar crown filaments toward the poles. The disappearance of the polar crown filament in the northern hemisphere occurred in the first half of 1981, shortly before the disappearance of the polar magnetic fields. The disappearance of the polar crown filaments is signified in Figure 13 by the change in polarity of the polar fields and in Figure 14 (Figure 5 from Harvey, 1994) by the curved lines representing the polar crown filaments and filament channels when they nearly reach the solar poles at 90° in latitude. They do not quite reach the pole before there is insufficient magnetic field to sustain them. Upon the complete disappearance of the polar crown filament, by definition, the subpolar crown filament is the new polar crown filament.

The solar cycle reversal of the direction of the polar crown filaments around the time of the sunspot maximum has become especially well known since the early measurements of the horizontal component of the magnetic fields in polar and sub-polar crown filaments using the Hanle effect (Bommier and Leroy (1998).

This reversal in the direction of magnetic fields of the polar crown filaments is linked to the reversal of the polarity of the polar crown magnetic fields shown by Harvey (1994) in Figure 14 from synoptic charts of magnetograms for cycles 21 and 22. The reversals of the direction of the magnetic fields along filaments were also confirmed by Martin (1998) for cycles 21, 22, and 23. However, Martin (1990) used the chiral method of determining the direction of the magnetic field along filaments. That method required using the observed directions of barbs on filaments from H-alpha images, supported by the knowledge of the polarity of the photospheric magnetic fields on each side of the filament from magnetograms available at the Big Bear Solar Observatory. The chiral method is explained using many different examples in Martin et al. (2008). This method of determining the field direction along filaments was found to be consistent with the measurements of the magnetic field in filaments using the Hanle effect (Bommier and Leroy, 1998). That meant one can employ either method to extrapolate the alternating field directions along polar crown filaments for many solar cycles forward and backward in time.

Babcock and Babcock (1955) first demonstrated that filaments are markers of polarity boundaries. This led to recognizing that the alternating direction of the polar and subpolar crown filaments is due to the alternating bands of the decayed remnants of active regions that have evolved into north-to-south tiers in latitude (Leroy, 1978; Leroy et al., 1984; Bommier and Leroy, 1998). The solar cycle reversals of the direction of the magnetic field along filaments in both hemispheres are simply due to the poleward migration of alternating tiers of positive and negative magnetic fields at high latitudes that are well explained in Bommier et al. (2021).

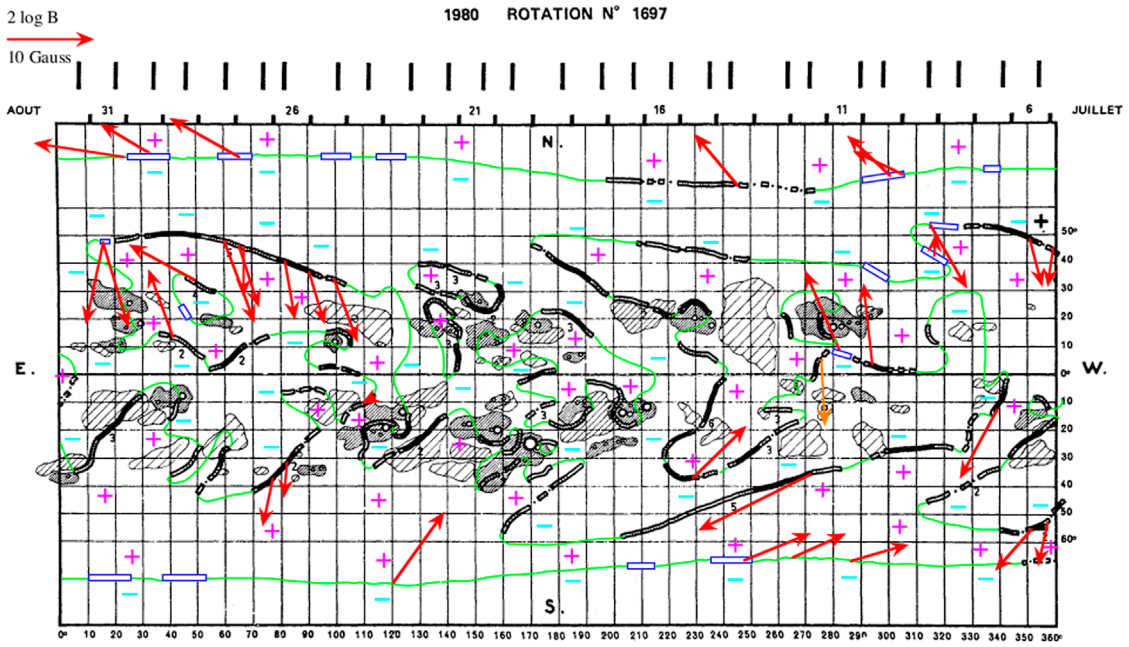


FIGURE 12
This synoptic chart of active regions and filaments for the whole of solar rotation no. 1697 is from the Meudon Solar Observatory. From right to left, the vertical lines above the map show the fraction of the central section of each full disk image which was incorporated into this chart. Photospheric magnetic fields are represented by red, + symbols for positive magnetic fields, and green – symbols for negative fields. These background polarities and the polarity boundaries (in green) have been added from McIntosh maps (NOAA/SEC) by Bommier et al. (2021) along with the red arrows which are proportional to the magnetic fields in the filaments. Polar crown filaments along the polar crown magnetic field boundaries in the northern and southern hemispheres are seen to be opposite in their west-to-east and east-to-west magnetic field directions. All of the arrows show the magnetic fields in the filaments are inverse. (They are in the – to + direction rather than + to – that is relative to the polarity of the major background fields on the two sides of each filament.) Along the next equatorward polarity boundary are the subpolar crown filaments which are opposite in their east-to-west or west-to-east direction from the polar crown filaments. The arrows on these charts were originally published in black and white by Bommier et al. (1985). (This updated color version was provided by the courtesy of V. Bommier and J. L. Leroy).

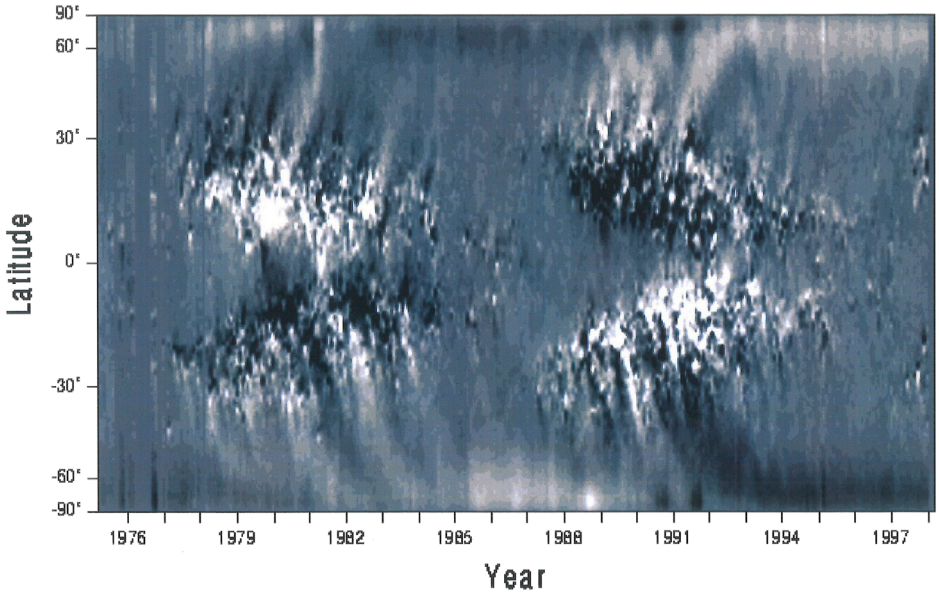


FIGURE 13
This butterfly diagram of magnetograms for solar cycle 21 shows the change of the polar field from positive to negative in the northern hemisphere and from negative to positive in the southern hemisphere in 1981. The northern hemisphere polar field remains negative until the next reversal to positive in 1991 during solar cycle 22. The southern hemisphere polar field reversed from positive to negative in 1990 (credit: Harvey, 1994, Figure 1, with permission from Springer-Verlag).

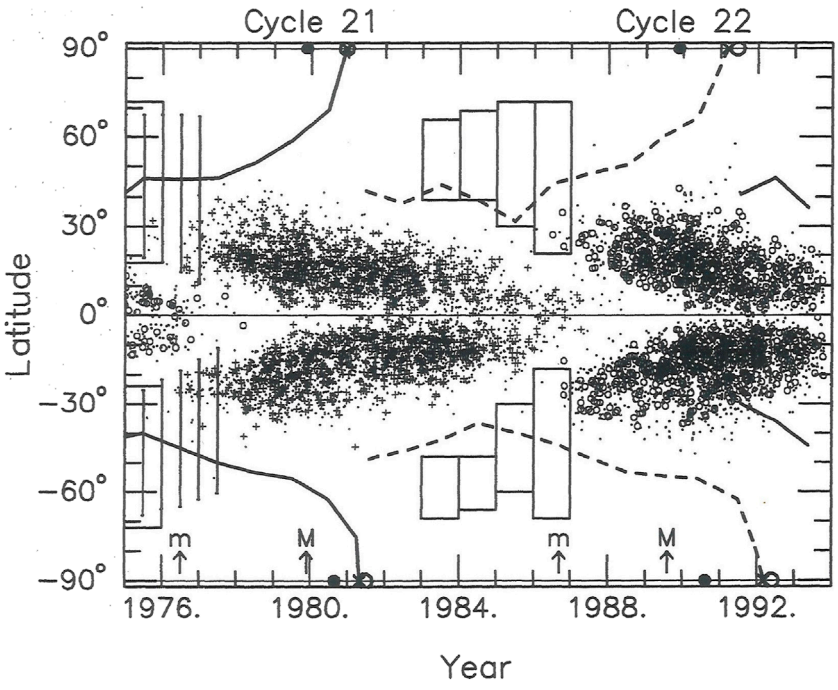


FIGURE 14
Butterfly diagram of sunspot regions of cycles 20, 21, and 22. The solid and dashed lines are the annual averages of the polarity inversion which represent the tracks of the polar crown filament for cycles 21 and 22. The latitude ranges of the high latitude belt of ephemeral active regions are shown as rectangles and small Ca II plage regions by vertical bars. Along the horizontal axis, 'M' is the time of solar maximum and 'm' is solar minimum (credit K. L. Harvey (1994) Figure 5, permission from Springer-Verlag).

The next questions are 1) is the west-to-east or east-to-west magnetic field direction of the ephemeral regions at the beginning of a 22-year solar cycle the same as that of the polar crown filament or the subpolar crown filament? 2) Does the band of new active regions at the beginning of a 22-year solar cycle spatially coincide with the polar crown or subpolar crown filament? The answers to these questions are also in published articles.

For the filaments:

A. Figure 12 (Figure 13 in Bommier et al. (1985)) shows the direction of the magnetic field of the polar crown filaments during August 1980, several months before they disappeared near the solar rotational poles. Their horizontal components are west-to-east for that time in solar cycle 21 and for all odd numbered cycles.
B. In Figure 5 by Harvey (1994), reproduced here in Figure 14, the west-to-east orientation applies to the polar crown filament in solar cycle 21 before its disappearance noted at the top of Harvey's Figure 5. Therefore, the direction along the subpolar crown filaments in that cycle were east-to-west. That same subpolar crown filament is, by definition, the next polar crown filament (Figure 14 (Harvey's Figure 5)). Therefore, the direction along the next polar crown filament in the northern hemisphere will be east-to-west for solar cycle 22 and for all even numbered cycles.

For the ephemeral regions:

As consistently observed, the polarity of the polar fields is always the same as the western polarity of the active regions in the same hemisphere and cycle. The polarity of the polar fields in the northern hemisphere, prior to the solar maximum in solar cycle 21, is positive as seen in Figure 13 (Figure 1 in Harvey, 1994). Therefore, active regions in the northern hemisphere in solar cycle 21 are oriented west-to-east according to the alternating pattern per solar cycle. Thus, the preferred direction of the ephemeral regions in the northern hemisphere in cycle 22 is the opposite; it is east to west.

Comparing these independently determined results for the orientations of polar crown filaments and ephemeral regions, the answer for the question above is: the magnetic fields of ephemeral active regions of a new solar cycle have the same preferred orientation as the polar crown filaments. In addition, in Figure 14, the latitude of the polar crown filament is approximately 45° in each hemisphere. This is slightly lower in latitude than the average latitude of 55° where the new 22-year solar cycles begin. However, the span of coronal arches above the polar crown filament can be 20° or more in width. It is the whole volume beneath the coronal loops that is relevant in this comparison. It is also important to notice that the polar crown filament is already established around the time of the emergence of the new active regions in the 22-year solar cycles.

The second question was "do the polar crown filaments and their channels coincide with the band of ephemeral regions at the beginning of a 22-year solar cycle?" The answer is in Figure 14 (Figure 5 in Harvey, 1994). The blocks represent ephemeral regions early in solar cycles 21 and 22. The curved line represents the track of the polar crown filaments and their filament channels. For approximately 3 or more years, the polar crown filament coincides

with some part of the latitude bands of the new cycle ephemeral active regions.

With these correlations, established in the literature, between the polar crown filaments and ephemeral active regions in both space and time, it is reasonable to make a general hypothesis about how the polarity reversal of solar cycles are initiated:

Hypothesis: The apparent reversal of the orientation of ephemeral active regions at the start of every 22-year solar cycle is due to their aligning with the dominant east-to-west or west-to-east orientation of the filament channels and channel environment into which the ephemeral regions emerge. The alignment process is suggested to be magnetic reconnection.

The direction of the magnetic fields of the environment of filament channels and filaments varies significantly throughout the volume they occupy in the chromosphere and corona. This volume can include the varying directions of the magnetic fields of the coronal loop system over a filament and filament channel, the filament spine, the filament barbs, and the underlying photosphere. Therefore, substantial scatter in the direction of the ephemeral region magnetic fields should be expected to exist and depend on what part of the environment is the closest to the individual ephemeral regions at the time when the magnetic reconnections are taking place between it and the environmental magnetic fields.

This hypothesis is based on the observation that magnetic reconnection between magnetic fields, that are partially aligned, causes them to become more aligned.

Direct evidence favorable or unfavorable to this hypothesis is expected to be difficult to achieve due to the high latitude at which ephemeral active regions form early in 22-year solar cycles. Very high-resolution magnetograms and/or filtergrams are essential to determine the magnetic polarity information required about the ephemeral active region, such as their changes over time.

Searching for relationships between ephemeral regions and filaments and their filament channels at lower latitudes might lead to learning whether there is more evidence for or against the hypothesis. Such new studies could show a requirement to refine the hypothesis. New studies might also lead to asking new questions about what could cause a bias toward west-to-east or east-to-west orientations of active regions of any size.

This hypothesis should apply to not only ephemeral active regions during the start of a solar cycle but also to all active regions whose magnetic flux is initially less than that of the environmental magnetic fields into which the active region emerges.

Evidence in favor of the above hypothesis is shown in Figure 7 in the article by Tlatov et al. (2010). As stated above, large scatter in the orientation of both environmental magnetic fields of filaments and filament channels, and the orientations of the active regions, which emerge into the environmental magnetic fields, is expected. Therefore, statistical associations are not expected to be strong for ephemeral active regions. However, the degree of alignment should increase as the magnitude of the interacting magnetic fields increases.

A single statement summary of this section is the reversal of polarity of ephemeral regions at the beginning of each 22-year solar cycle is due to the reversed direction of the magnetic fields of the filament channel and environment of the channel into which they emerged. Reconnection between the ephemeral region magnetic fields and its environmental magnetic fields results in increased alignment of the ephemeral region structure with the west-to-east or east-to-west direction of the magnetic field into which it emerges.

9 Discussion: the transition to a new paradigm for solar cycles

This review has been written during a transition to a new paradigm in how solar cycles are viewed. The former paradigm was centered on sunspots and continued with new observations of the active region magnetic fields that encompass sunspots. In the previous paradigm, solar cycles were linear successions of sunspot/active region cycles of approximately 11 years. Two 11-year cycles were required for a complete magnetic cycle.

The slow transition to the new paradigm changed the focus from the sunspot belt to the global Sun. The emphasis switched from the 11-year sunspot cycle to the acknowledgment of the extended solar cycle. This recognition first depended on the new full-disk daily magnetograms at the Kitt Peak National Solar Observatory. They had sufficiently high spatial resolution to reveal vast numbers of small-scale active regions that did not reach a magnitude great enough to produce sunspots. Singly, these tiny active regions seemed insignificant; collectively, they have a colossal effect on how solar cycles are viewed. The magnetic flux they bring to the solar surface is greater than that in the active regions with sunspots. As if ephemeral regions are in control, their numbers increase early in each solar cycle. While continuing to increase their numbers, the ephemeral active regions begin to usher-in a lot of small active regions that produce small sunspots, followed by a lesser number of active regions with medium-area sunspots, and then a few active regions that produce large sunspots. After the solar maximum, for all sizes of active regions, the ephemeral active regions decrease their population but first they usher-out those active regions with sunspots approximately in reverse order to their ushering-in.

The culmination of the extended cycle is the recognition of two 22-year solar cycles present in separate latitude bands continuously. Two cycles each of 22 years duration are essential because each 22-year band of regions in latitude contains a dominance of west-to-east or east-to-west oriented active regions. The two cycles are offset by 11 years with the beginning of each 22-year cycle at the highest latitudes where the ephemeral active regions can be identified. The beginning of each 22-year cycle coincides with the disappearance of the polar magnetic fields of single polarity. These polar magnetic caps and the equator serve as markers of the hidden mechanism of the solar cycle clock. They mark the timing of the two 22-year cycles which are out of phase by 11 years with more precision than could be imagined earlier. The accuracy of this solar cycle clock is a harbinger of more accurate future predictions and offers new views of historical records.

Over the next decade, it is expected that the new paradigm will translate into new ideas, observations, and models for solar cycles. These will encompass observations yet to be fully understood about active regions. This will lead to greater understanding of their orientation patterns, peculiarities of their sunspots, tilts with respect to latitude and longitude, clustering in longitude, global polarity reversals, and depths of origin.

Notably, these key items for solar cycle research are almost the same as originally listed by Babcock (1961). In light of the new paradigm, they are echoes of Babcock's research call, "Of particular value would be more and better measurements of the distribution and quantity of magnetic flux in BMR (bipolar magnetic regions) as a function of age."

Now the little "bipolar magnetic regions" that bring on solar cycles are discussed as "ephemeral active regions" and "elementary bipoles." As the progenitors of both larger active regions and their sunspots, what is their origin?

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgments

The author expresses appreciation to Prof. Oddbjorn Engvold for reading and commenting on the scientific content of the manuscript, to Prof. Piet Martens for discussing the hypothesis for the initiation of polarity reversals at the beginning of 22-year solar cycles, to novelist Gil Roscoe and song-writer Alan Higgins for assistance in detailed editing of the manuscript, and to student Gabriel Higgins for resolving complications of fussy software.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Altrock, R. C. (1988). "Variation of solar coronal Fe XIV 5303 Å emission during solar cycle 21, in solar and stellar coronal structure and dynamics," in *Proceedings of the ninth Sacramento peak summer symposium, sunspot, NM*. Editor R. C. Altrock (Sunspot, NM: National Solar Observatory), 414.

Altrock, R. C. (1997). An 'extended solar cycle' as observed in Fe XIV. *Sol. Phys.* 170, 411–423. doi:10.1023/a:1004958900477

Anderson, M., and Martin, S. F. (2005). "Formation of an extraordinarily long filament channel," in Large-scale Structures and their Role in Solar Activity, Proceedings of the Conference held, Sunspot, New Mexico, USA, 18-22 October, 2004. Editors K. Sankarasubramanian, M. Penn, and A. Pevtsov (IEEE), 201.

Babcock, H. D. (1959). The Sun's polar magnetic field. *Astrophys. J.* 130, 364. doi:10.1086/146726

Babcock, H. D., and Babcock, H. W. (1955). The Sun's magnetic field 1952-1954. *Astrophys. J.* 121, 349. doi:10.1086/145994

Babcock, H. D., and Babcock, H. W. (1958). "Photospheric magnetic fields," in *Electromagnet phenomena in cosmical physics. Editor bo lehnert, IAU symposium no. 6* (Cambridge: Cambridge University Press), 239–247.

Babcock, H. D., and Livingston, W. C. (1958). National academy of Sciences: abstracts of papers presented at the annual meeting, 28-30 april 1958, Washington, D.C. *Science* 127, 1058–1063. doi:10.1126/science.127.3305.1058

Babcock, H. W. (1947). Zeeman Effect in stellar spectra. *Astrophys. J.* 105, 105B. doi:10.1086/144887

Babcock, H. W. (1961). The topology of the Sun's magnetic field and the 22-year cycle. *Astrophys. J.* 133, 572. doi:10.1086/147060

Babcock, H. W., and Babcock, H. D. (1952). Mapping the magnetic fields of the Sun. *PASP* 64, 282. doi:10.1086/126495

Balasubramaniam, K. S., Pevtsov, A., and Rogers, J. (2004). Statistical properties of superpenumbral whorls around sunspots. *Astrophys. J.* 608, 1148–1155. doi:10.1086/420759

Beer, J., Tobias, S., and Weiss, N. (1998). An active sun throughout the maunder minimum. *Sol. Phys.* 181, 237–249. doi:10.1023/a:1005026001784

Bernasconi, P. N., Rust, D. M., and Geogoulis, M. K. (2002). Moving dipolar features in an emerging flux region. *Solar Phys.* 209, 119–139.

Bocchino, G. (1933). Migrazione delle protuberanze durante il ciclo undecennale dell'attivata solare. *Oss. Mem. Oss. Astrofus. Arcetri* 51, 5–47.

Bommier, V., Degl'Innocenti, E. L., Leroy, J.-L., et al. (1994). Complete determination of the magnetic field vector and of the electron density in 14 prominences from linear polarization measurements in HeI D3 and H alpha lines. *Solar Phys.* 154, 231–326.

Bommier, V., and Leroy, J. L. (1998). Global pattern of the magnetic field vector above neutral lines from 1974 to 1982: Pic-du-Midi observations of prominences, New Perspectives in Solar Prominences, IAU Colloquium 167. *Asp. Conf. Ser.* 150, 434–438.

Bommier, V., Leroy, J. L., and Sahal- Brechot, S. (1985). "MSFC workshop: measurements of solar fields, huntsville,", Alabama, USA, 15-18 May, 1984. Editor M. J. Hagyard (NASA Conference Publication), 375.2374.

Bommier, V., Leroy, J.-L., and Sahal-Brechot, S. (2021). 24 synoptic maps of average magnetic field in 296 prominences measured by the Hanle effect during the ascending phase of Solar Cycle 21. *Astron. & Astrophys* 647, A60. Id. A60. doi:10.1051/0004-6361/202038868

Bondar, N. I., and Katsova, M. M. (2020). *Long-term activity in photospheric low-mass stars with strong magnetic fields*. arXiv:201.

Born, R. (1974). First phase of active regions and their relation to the chromospheric network. *Sol. Phys.* 38, 127–131. doi:10.1007/bf00161830

Brandenburg, A. (2006). "Location of the solar dynamo and near-surface shear," in *MHD theory and observations: a high resolution perspective*. Editors H. Uitenbroek, J. Leibacher, and R. F. Stein (ASP Conference) 354, 121–126.

Bray, R. J., and Loughhead, R. E. (1964). *Sunspots*. London: Chapman & Hall.

Bruzek, A. (1969). Motions in arch filament systems. *Sol. Phys.* 8, 29–36. doi:10.1007/bf00150655

Carrington, R. C. (1858). On the distribution of the solar spots in latitude since the beginning of the year 1854; with a map. *MNRAS* 19, 1–3. doi:10.1093/mnras/19.1.1a

Chae, J., Martin, S. F., Yun, H. S., Kim, J., Lee, S., Goode, P. R., et al. (2001). Small magnetic bipoles emerging in a filament channel. *Astrophys. J.* 548, 497–507. doi:10.1086/318661

Chatterjee, S., Hegde, M., Banerjee, D., Ravindra, B., and McIntosh, S. W. (2020).Time-latitude distribution of for 10 solar cycles: a study using Kodaikanal, Meudon, and Kanzelhohe data, *Earth Space Sci.*, e00666, 7.

Chian, A. C.-L., Rempel, E. L., Silva, S. S. A., Rubio, L. B., and Gosic, M. (2023). Intensification of magnetic field in merging magnetic flux tubes driven by supergranular vortical flows. *MNRAS* 518, 4930–4942. doi:10.1093/mnras/stac3352

Cliver, E. W. (2014). The extended cycle of solar activity and the Sun's 22-year magnetic cycle. *Space Sci. Rev.* 186, 169–189. doi:10.1007/s11214-014-0093-z

Eddy, J. A. (1976). The maunder minimum. *Science* 192, 1189–1202. doi:10.1126/science.192.4245.1189

Eddy, J. A. (1983). The maunder minimum: a reappraisal. *Sol. Phys.* 89, 195–207. doi:10.1007/bf00211962

Engvold, O., and Zirker, J. B. (2016). The parallel worlds of Christoph scheiner and Galileo galilei. *Hist. Astron* 7 (3), 332–345. doi:10.1177/0021828616662406

Frazier, E. N. (1972a). The magnetic structure of arch filament systems. *Sol. Phys.* 26, 130–141. doi:10.1007/bf00155113

Frazier, E. N. (1972b). The relations between chromospheric features and photospheric magnetic fields. *Sol. Phys.* 24, 98–112. doi:10.1007/bf00231086

Hale, G. E. (1908). On the probable existence of a magnetic field in sunspots. *Astrophys. J.* 28, 315. doi:10.1086/141602

Hale, G. E. (1913). Preliminary results of an attempt to detect the general magnetic field of the Sun. *Astrophys. J.* 38, 27–30. doi:10.1086/142013

Hale, G. E., and Ellerman, F. (1918). The polarity of sunspots. *PAAS* 3, 81.

Hale, G. E., Ellerman, F., Nicholson, S. B., and Joy, A. H. (1919). The magnetic polarity of sunspots. *Astron. J.* 49, 153–178. doi:10.1086/142452

Hale, G. E., and Nicholson, S. B. (1925). The law of sunspot polarity. *Astrophys. J.* 62, 270–300. doi:10.1086/142933

Hale, G. E., and Nicholson, S. B. (1938). *Magnetic observations of sunspots*, No. 498. Washington, D. C: Carnegie Inst. of Washington Publ.

Hansen, S. F., and Hansen, R. T. (1975). Differential rotation and reconnection as basic causes of some coronal reorientations. *Solar Phys.* 44, 503–508.

Harvey, K. (2000). *Solar active regions: ephemeral, encyclopedia of Astronomy and astrophysics*. Bristol: Institute of Physics Publishing. Available at http://eaa.iop.org/abstract/0333750888/2275.

Harvey, K., and Martin, J. (1973). Observations of moving magnetic features near sunspots. *Sol. Phys.* 28, 61–71. doi:10.1007/bf00152912

Harvey, K. L. (1992). The cyclic behavior of solar activity. *Sol. Cycle ASP Conf. Ser.* 27, 335–367.

Harvey, K. L. (1993a). *PhD thesis*. University of Utrecht.

Harvey, K. L. (1994). "The solar magnetic cycle," in *Solar surface magnetism* (Kluwer: Kluwer Academic Publishers), 347–363.

Harvey, K. L., Harvey, J. W., and Martin, S. F. (1975). Ephemeral active regions in 1970 and 1973. *Sol. Phys.* 40, 87–102. doi:10.1007/bf00183154

Harvey, K. L., and Zwaan, C. (1993b). Properties and emergence patterns of bipolar active regions: I.Size distribution and emergence frequency. *Sol. Phys.* 148, 85–118. doi:10.1007/bf00675537

Hathaway, D. H. (2015). The solar cycle. *Sol. Phys.* 7, 1. doi:10.12942/lrsp-2010-1

Howard, R., and Labonte, B. J. (1980). The sun is observed to be a torsional oscillator with a period of 11 years. *Astrophys. J. Part 2 - Lett. Ed.* 239, L33–L36. doi:10.1086/183286

Howard, R. F. (1992). On the large-scale distribution of solar magnetic fields. *Sol. Cycle ASP Conf.* 27, 44.

Jarboe, T. R., Benedett, T. E., Everson, C. J., Hansen, A., Hossack, A. C., Morgan, K. D., et al. (2017). *Evidence for a shallow thin magnetic structure and solar dynamo: the driver of torsional oscillations*.

Jiang, J., Cameron, R. H., and Schussler, M. (2015). *The cause of the weak solar cycle*. arXiv:1507.01764v1[Astro-ph.SR].

Jiao, Q., Jiang, J., and Wang, Z.-F. (2021). Sunspot tilt angles revisited: dependence on the solar cycle strength. *Astron. & Astrophys*, 11615v1. arXiv:2106.

Jin, C., Zhou, G., Zhang, Y., and Wang, J. (2020). The bipolar magnetic emergence in the solar polar region. *Astrophys. J. Lett.* 889, L26. doi:10.3847/2041-8213/ab65bf

Kutsenko, A. S. (2020). The rotation rate of solar active and ephemeral regions – I. Dependence on morphology and peak magnetic flux. *MNRAS* 000, 5159–5166. doi:10.1093/mnras/staa3616

Leamon, R. J., McIntosh, S. W., and Chapman, S. C. (2000a). *The Hale cycle clock*. Bristol, AGU Fall Meeting SH053.

Leamon, R. J., McIntosh, S. W., and Marsh, D. R. (2018). Termination of solar cycles and correlated tropospheric variability. *ADS*. arXiv-prints.

Leamon, R. J., McIntosh, S. W., and Title, A. M. (2020b). Deciphering solar magnetic activity: the solar cycle clock. *Front. Astron. Space Sci.* 9, 886670. doi:10.3389/fspas.2022.886670

Leamon, R. J., McIntoshChapman, S. W. S. C., and Watkins, N. W. (2019). *Timing terminators: forecasting sunspot cycle 25*. arXiv.

Leighton, R. B. (1964). Transport of magnetic fields on the Sun. *Astrophys. J.* 140, 1547–1562L. doi:10.1086/148058

Leighton, R. B. (1969). A magneto-kinematic model of the solar cycle. *Astrophys. J.* 156, 1–26. doi:10.1086/149943

Leighton, R. B., Noyes, R. W., and Simon, G. W. (1962). Velocity fields in the solar atmosphere. I. Preliminary Report. *Astrophys. J. Let.* 135, 474–499. doi:10.1086/147285

Leroy, J.-L. (1978). On the orientation of magnetic fields in quiescent prominences. *Astron. Astrophys.* 64, 647–252.

Leroy, J.-L. (1988). "Solar and stellar coronal structure and dynamics," in Proceedings of the 9th Sacramento Peak Summer Symposium, Sunspot, New Mexico, 17–21 August 1987. Editor R. C. Altrock (IEEE), 422.

Leroy, J.-L. (1989). "Dynamics and structure of quiescent solar prominences," in *Dynamics and structure of solar prominences*. Editor E. R. Priest (Dordrecht: Kluwer Academic Publishers), 77.

Leroy, J.-L., Bommier, V., and Sahal-Bréchot, S. (1983). The magnetic field in the prominences of thepolar crown. *Sol. Phys.* 83, 135.

Leroy, J.-L., Bommier, V., and Sahal-Bréchot, S. (1984). New data on the structure of quiescent Prominences. *Astron. Astrophys.* 131-33.

Leroy, J.-L., and Noens, J.-C. (1983). Does the solar activity cycle extend over more than an 11-year period? *Astron. Astrophys.* 120, L1–L2.

Litvinenko, Y., Chae, J., and Park, S.-Y. (2007). Flux pile-up magnetic reconnection in the solar photosphere. *Astrophys. J.* 662, 1302–1308. doi:10.1086/518115

Litvinenko, Y. E. (2010). Evolution of the axial magnetic field in solar filament channels. *Astrophys. J.* 720, 948–952. doi:10.1088/0004-637x/720/1/948

Litvinenko, Y. E., and Martin, S. F. (1999). Photospheric magnetic reconnection and canceling magnetic features on the sun. *Astrophys. J.* 515, 435–440. doi:10.1086/307001

Livi, S. H. B., Wang, J., and Martin, S. F. (1985). The cancellation of magnetic flux I - on the quiet Sun. *Aust. J. Phys.* 38, 855–873. doi:10.1071/ph850855

Loughhead, R. E., and Bray, R. J. (1979). *Sunspots*. USA, Dover Publishers, Inc.

Makarov, V. I., and Sivaraman, K. R. (1989a). Evolution of latitude zonal structure of the large-scale magnetic field in solar cycles. *Sol. Phys.* 119, 35–44. doi:10.1007/bf00146210

Makarov, V. I., and Sivaraman, K. R. (1989b). New results concerning the global solar cycle. *Sol. Phys.* 123, 367–380. doi:10.1007/bf00149112

Martin, S. F. (1988). The identification and interaction of network, intranetwork and ephemeral region magnetic fields. *Solar Phys.* 117, 243.

Martin, S. F. (1990). Elementary bipoles of active regions and ephemeral active regions. *Soc. Astron. Ital. Mem.* 61, 293–315.

Martin, S. F. (1998). Filament chirality: a link between fine-scale and global patterns new perspectives in solar prominences, iau colloquium 167. *Asp. Conf. Ser.* 150, 419–429.

Martin, S. F. (2015). "The magnetic field structure in prominences from direct and indirect observations, Chapter 9," in *Solar prominences*. Editors J.- C Vial, and O. Engvold (Germany: Springer Publishers), 205–235.

Martin, S. F. (2018). Observational evidence of shallow origins for the magnetic fields of solar cycles. *Front. Astron. Space Sci.Sec. Stellar Sol. Phys.* 5, 315. doi:10.3389/fspas.2018.00017

Martin, S. F., Bilimoria, R., and Tracadas, P. W. (1994). "Magnetic field configurations basic to filament channels and filaments," in *Solar surface magnetism*. Editors R. J. Rutten, and C. J. Schrijver (Dordrecht: Kluwer Academic Publishers), 303.

Martin, S. F., Engvold, O., and Lin, Y. (2008). A method of resolving the 180-degree ambiguity by employing the chirality of solar features. *Sol. Phys.* 250, 31–51. doi:10.1007/s11207-008-9194-8

Martin, S. F., and Harvey, K. H. (1979). Ephemeral active regions during solar minimum. *Sol. Phys.* 64, 93–108. doi:10.1007/bf00151118

Martin, S. F., Livi, S. H. B., and Wang, J. (1985). The cancellation of magnetic flux II - in a decaying active region. *Aust. J. Phys.* 38, 929–959. doi:10.1071/ph850929

Martin, S. F., Panasenco, O., Berger, M. A., Engvold, O., Lin, Y., Pevtsov, A. A., et al. (2012). "The build-up to eruptive events viewed as the development of chiral systems, the Second ATST East Meeting: magnetic Fields from the Photosphere to the Corona," in ASP Conference Series, San Francisco, 15 Dec 2012. Editors T. Rimmele, A. Tritschler, F. Woger, V. Collados, H. Socos Navarro, R. Schlichenmaier, et al. (Astronomical Society of the Pacific), 157.463.

Martinez-Pillet, V., Lites, B. W., and Skumanich, A. (1998). Vector magnetic fields of emerging solar flux, *ESSP* 417, 259–261.

Mazumder, R. (2019). *Properties of solar filaments in solar cycle 20-23 from McIntosh Archive*. article id. 080.

Mazumder, R., Chatterjee, S., Nandy, D., and Banerjee, D. (2021). Solar cycle evolution of filaments over a century: investigations with the Meudon and McIntosh hand-drawn archives. *Astrophys. J.* 919, 125. doi:10.3847/1538-4357/ac09f6

McIntosh, S. W., and Leamon, R. J. (2015). Deciphering solar magnetic activity: on grand minima in solar activity, *Front. Sol. Stellar Phys. article, Front. Astronomy Space Sci.*, Vol.2, 2, doi:10.3389/fspas.2015.00002

McIntosh, S. W., and Leamon, R. J. (2017). Deciphering solar magnetic activity: spotting solar cycle 25. *Front. Astron. Space Sci.* 4, 4. doi:10.3389/fspas.2017.00004

McIntosh, S. W., Leamon, R. J., Egeland, R., Dikpati, M., Altrock, R. C., Banerjee, D., et al. (2021). Deciphering solar magnetic activity: 140 years of the 'extended solar cycle' – mapping the Hale cycle. *Solar Phys.* 296, 89.

McIntosh, S. W., Leamon, R. J., Egeland, R., Dikpati, M., Fan, Y., and Rempel, M. (2019). What the sudden death of solar cycles can tell us about the nature of the solar interior. *Sol. Phys.* 294, 88. doi:10.1007/s11207-019-1474-y

McIntosh, S. W., Scherrer, P. H., Svalgaard, L., and Leamon, R. J. (2022). Uniting the Sun's Hale magnetic cycle and "extended solar cycle" paradigms. *Front. Astronomy Space Sci.* 9, 923049. doi:10.3389/fspas.2022.923049

McIntosh, S. W., Wang, X., Leamon, R. J., Davey, A. R., Howe, R., Krista, L. D., et al. (2014). Deciphering solar magnetic activity. I. On the relationship between the sunspot cycle and the evolution of small magnetic features. *Astrophys. J.* 792, 12. doi:10.1088/0004-637X/792/1/12

Metcalf, T. S., Buccino, A. P., Brown, B. P., Mathur, S., Soderblom, D. R., Henry, T. J., et al. (2012). *Magnetic activity cycles in the exoplanet host star Eta Eridani*. arXiv:1212.4425v1 [astro-ph.SR].

Munoz-Jaramillo, A., Benjamin, N., and Campusano, L. E. (2022). *Solar anti-Hale bipolar magnetic regions: a distinct population with systematic properties*. arXiv:2203.11898v1 [astro-ph.SR] 22 Mar 2022.

Parker, E. N. (1955). The Formation of sunspots from the solar toroidal field. *Astrophys. J.* 121, 491–507. doi:10.1086/146010

Parnell, C. E., DeForest, C. R., Hagenaar, H. J., Johnston, B. A., Lamb, D. A., and Welsch, B. T. (2009). A power-law distribution of solar magnetic fields over more than five decades in flux. *Astrophys. J.* 698, 75–82. doi:10.1088/0004-637x/698/1/75

Pesnell, W. D. (2012). Solar cycle predictions (review). *Sol. Phys.* 281, 507–532.

Petrie, G. (2022). *Polar global magnetic field evolution and global flux transport*. Germnay: Springer. doi:10.21203/rs.3.rs-2298525/v1

Petrovay, K. (2010). Solar cycle prediction, living review. *Sol. Phys.* 17 (1). article id.2.

Priyadarchi, A., Hegde, M., Jha, B. K., Chatterjee, S. M., Mandal, S., Chowdhury, M., et al. (2022). *Detection of solar filaments using suncharts from Kodaikanal solar Observatory archive employing a clustering approach*. arXiv:2212.12176v1 [astro-ph.SR].

Richardson, R. S. (1941). The nature of solar hydrogen vortices. *Astrophys. J.* 41, 24. doi:10.1086/144238

Robbrecht, E., Wang, Y.-M., Sheeley, Jr., and Rich, N. B. (2010). On the 'extended' solar cycle in coronal emission. *Astrophys. J.* 716, 693–700. doi:10.1088/0004-637x/716/1/693

Rust, D. (1967). Magnetic fields in quiescent solar prominences. I. Observations. *Astrophys. J.* 150, 313–326. doi:10.1086/149333

Rust, D. M. (1968). in *Structure and development of solar active regions, IAU symp. No. 35*. Editor K. O. Kiepenheuer, 77.

Rust, D. M., and Martin, S. F. (1994). "A Correlation between sunspot whirls and filament type," in ASP Conf. Series Astron. Soc. Pacific, San Francisco, 3 September 1993. Editors K. S. Balasubramaniam, and G. W. Simon (IEEE), 337.

Schwabe, M. (1844). Sonnenbeobachtungen im Jahre 1843. Von Herrn Hofrath Schwabe in Dessau. *Astron. Nachrichten* 21, 233.

Simon, G. W., and Leighton, R. B. (1964). Velocity fields in the solar atmosphere. III. Large-scale motions, the chromospheric network, and magnetic fields. *Astrophys. J.* 140, 1120. doi:10.1086/148010

Smith, S. F., Ramsey, H., and Howard, R. F. (1965). Some characteristic properties of solar magnetic fields. *Astron. J.* 70, 330. doi:10.1086/109605

Snodgrass, H. B. (1987). Torsional oscillations and the solar cycle. *Sol. Phys.* 110, 35–49. doi:10.1007/bf00148200

Spörer, G. (1894). Beobachtungen von sonnenflecken in den jahren 1885 bis 1893, Pub. *Potsdam Obs.* 10 (1), 3–147.

Srivastava, A. K., McIntosh, S. W., Arge, N., Banerjee, D., Dipankar, G., Dikpati, M., et al. (2018). The extended solar cycle: muddying the waters of solar/stellar dynamo modeling or providing crucial observational constraints? *Front. Astronomy Space Sci. Mini-Review* 5. doi:10.3389/fspas.2018.00038

Strous, L. H., and Zwaan, C. (1999). Phenomena in an emerging active region, II. Properties of the dynamic small-scale structure. *Astrophys. J.* 527, 435–444. doi:10.1086/308071

Tandberg-Hanssen, E. (1974). *Solar prominences*. Dordrecht, Holland: D. Reidel Publishing C.

Tandberg-Hanssen, E. (1995). *The nature of solar prominences*. Dordrecht: Kluwer Academic Publishers.

Tang, F. (1987). Quiescent prominences – where are they formed?. *Solar Phys.* 107, 233.

Tlatov, A. G. (1996). A model of the solar magnetic cycle with a meridian stream closed inside the convective zone. *Quantum Electron* 39, 794–798. doi:10.1007/bf02120957

Tlatov, A. G., Makarov, M. I., and Sivaraman, K. R. (1989). New results concerning the global solar cycle, *Sol. Phys*. 123, 367, 380. doi:10.1007/bf00149112

Tlatov, A. G., Vasil'eva, V. V., and Pevtsov, A. A. (2010). Distribution of magnetic bipoles on the Sun over three solar cycles. *Astrophys. J.* 717, 357–362. doi:10.1088/0004-637x/717/1/357

Vial, J.-C., and Engvold, O. (2015). in *Solar prominences*. Editors J.-C. Vial, and O. Engvold (Switzerland: Springer International Publishing).

Vrabec, D. (1974). "Streaming magnetic features near sunspots, chromospheric fine structure. Editor R. Grant Athay, IAU Symp. 56, Surfer's Paradise, Qld., Australia," in *3-7 september 1973* (Dordrecht; Boston: Reidel), 201–231.

Wang, H., and Li, H. (2022). Rediscovery of 23 historical records of naked-eye sunspot observations in AD 1618. *Sol. Phys.* 297, 127. Article 127. doi:10.1007/s11207-022-02046-w

Wang, Y.-M., Nash, A. G., and Sheeley, N. R., Jr. (1989). Magnetic flux transport on the Sun. *Science* 249, 712–718. doi:10.1126/science.245.4919.712

Wang, Y. M., and Sheeley, N. R., Jr. (1994). The rotation of photospheric magnetic fields: a random walk transport model. *Astrophys. J.* 430, 399–412. doi:10.1086/174415

Wang, Y.-M., Sheeley, N. R., Jr., and Nash, A. G. (1991). A new solar cycle model including meridional circulation. *Astrophys. J.* 383, 431–442. doi:10.1086/170800

Weart, S. (1970). The birth and growth of sunspots. *Astrophys. J.* 162, 97–922.

Webb, D. F., Davis, J. M., and McIntosh, P. S. (1984). Observations of the reappearance of polar coronal holes and the reversal of the polar magnetic field. *Sol. Phys.* 92, 109–132. doi:10.1007/bf00157239

Weber, M., Schunker, H., Jouve, L., and Isik, E. (2023). Understanding active region origins and emergence on the sun and other cool stars. *Space Sci. Rev.* 219, 63. arXiv.2306.06536v1 [astro-ph.SR] 10 June 2023. doi:10.1007/s11214-023-01006-5

Wilson, P. R., Altrock, R. C., Harvey, K. L., Martin, S. F., and Snodgrass, H. B. (1988). The extended solar activity cycle. *Nature* 333, 748–750. doi:10.1038/333748a0

Wood, P., and Martens, P. (2003). Measurements of flux cancellation during filament formation. *Sol. Phys.* 218, 123–135. doi:10.1023/b:sola.0000013027.52104.c6

Zwaan, C. (1978). On the appearance of magnetic flux in the solar photosphere. *Sol. Phys.* 60, 213–240. doi:10.1007/bf00156523

Frontiers in
Astronomy and Space Sciences

Explores planetary science and extragalactic astronomy in all wavelengths

Advances the understanding of our universe - from planetary science to extragalactic astronomy, to high-energy and astroparticle physics.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact