



OPEN ACCESS

EDITED BY
Alex Sun,
The University of Texas at Austin, United States

REVIEWED BY
Muhammad Tayyab Naqash,
Islamic University of Madinah Faculty of
Engineering, Saudi Arabia
Hasan Shaheed,
Universiti Tenaga Nasional, Malaysia

*CORRESPONDENCE
Sergio Gabriel Ceballos Pérez
✉ sgceballospe@secihti.mx

RECEIVED 28 November 2025
REVISED 22 December 2025
ACCEPTED 12 January 2026
PUBLISHED 05 February 2026

CITATION
Torres González MA, Ceballos Pérez SG, Lara
Figueroa HN, Ávila Camacho FJ, Moreno
Villalba LM, Carrillo JMS and Meléndez
Ramírez A (2026) Machine learning and
predictive models for water management: a
systematic review. *Front. Water* 8:1756052.
doi: 10.3389/frwa.2026.1756052

COPYRIGHT
© 2026 Torres González, Ceballos Pérez, Lara
Figueroa, Ávila Camacho, Moreno Villalba,
Carrillo and Meléndez Ramírez. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). 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.

Machine learning and predictive models for water management: a systematic review

Miguel Angel Torres González¹,
Sergio Gabriel Ceballos Pérez^{1,2*},
Hugo Nathanael Lara Figueroa³,
Francisco Jacob Ávila Camacho³,
Leonardo Miguel Moreno Villalba³, Juan Manuel Stein Carrillo³
and Adolfo Meléndez Ramírez³

¹Financial Engineering Department, Polytechnic University of Pachuca, Pachuca, Hidalgo, Mexico, ²Department of Researchers for Mexico, Secretary of Science, Humanities, Technology and Innovation, Mexico City, Mexico, ³National Technological Institute of Mexico/ITESM Ecatepec, Ecatepec, Estado de México, Mexico

Introduction: Water resource management faces strategic challenges posed by climate change, urban expansion, and land-use transformations. Machine learning (ML) has emerged as a promising alternative, capable of learning patterns from large datasets, contributing to the design of forecasting models, and revolutionizing the sustainable management of water.

Methods: This systematic review followed PRISMA 2020 guidelines. The study identified 35 records, reviewed 18 full texts, and excluded 17 studies. Searches targeted Scopus, Web of Science, IEEE Xplore, ScienceDirect, and were supplemented by Google Scholar and manual reference screening. The equation combined water-related terms such as “water management” with machine learning terms such as “deep learning,” “artificial intelligence,” etc. Inclusion required peer-reviewed articles with sufficient methodological description and English or Spanish full text. Exclusions comprised narrative reviews, gray literature, and studies lacking algorithmic details. The period spanned 2010–2025 to capture ML growth.

Results: The results show that deep learning models (especially LSTM) offer significant improvements in time prediction, while assembly-based algorithms (Random Forest, XGBoost, CatBoost) stand out for their robustness in data-constrained situations. Hybrid ML + physical model approaches showed high efficacy in correcting bias and improving hydrological projections. Gaps in reproducibility, uncertainty analysis, and integration of anthropogenic factors were identified. Geographic focus favored Asia, Europe, and North America with 10–50 years series. Common metrics included RMSE, MAE, R², NSE, and KGE. It is concluded that ML constitutes a strategic tool to strengthen water management in scenarios affected by climate variability and data scarcity.

Discussion: ML captures nonlinearities, adapts to noisy data, and integrates multi-source sensor and satellite data. Reproducibility remains limited, as few studies publish code or hyperparameters. Integration of anthropogenic factors (dams, irrigation, urbanization) remains insufficient. Future research must

adopt reproducibility frameworks, incorporate explicit uncertainty analysis, and advance physically informed hybrid models. The evidence confirms ML's value for water management under climate variability and data scarcity, but consolidation requires addressing methodological weaknesses.

KEYWORDS

climate change, deep learning, forecasting, hybrid models, machine learning

1 Introduction

The efficient management of water resources has become a strategic challenge in the face of growing pressure from climate change, urban expansion, and land use transformations (Wagener et al., 2010; Nearing et al., 2020). Flow prediction, drought assessment, flood control and estimation of hydrometeorological variables are essential elements to ensure water availability, distribution and sustainability (Bhadauria et al., 2024). Traditionally, these processes have been addressed through physical and hydrological models that describe the interactions between precipitation, runoff, evapotranspiration and storage (Zhang et al., 2021). However, the nonlinear complexity of hydrological systems and the limited availability of high-resolution data make it difficult to apply purely deterministic models in many contexts (Addor and Melsen, 2019; Sharma et al., 2020).

Over the past decade, machine learning (ML) has emerged as a promising alternative to address these limitations (Almikael et al., 2022). ML algorithms can learn complex patterns from large volumes of data, capture nonlinear relationships, and adapt to changing hydrological environments (Nearing et al., 2021). Deep neural networks, assembly models, hybrid ML–physical hydrology methods, and spatiotemporal architectures have been applied with encouraging results in predicting key hydrological variables (Ardabili et al., 2020).

Despite the accelerated growth of this research topic, there is still no structured synthesis that allows evaluation of the state of the art, comparing methodological approaches (Kratzert et al., 2019;

Zhong et al., 2020), identifying gaps, and proposing a research agenda. Therefore, the objective of this study was to conduct a systematic review of the literature that analyzes the use of ML for water management, following the PRISMA 2020 methodology, with special emphasis on the most frequent applications of ML in hydrology, algorithms used and their performance, data schemas and validation, implications for water resource management, and current gaps and future research opportunities.

2 Methodology

The systematic review was developed following the guidelines established in the PRISMA 2020 declaration, with the purpose of identifying, evaluating and synthesizing the scientific literature that applied machine learning models and predictive approaches to water management problems (Sarkis-Onofre et al., 2021). The methodological process included identification, screening, eligibility, and inclusion phases, seeking to ensure transparency, reproducibility, and rigor in all stages of the study. The review was not registered in PROSPERO because it is an engineering field, although it followed the principles of transparency and reproducibility suggested for formal reviews.

This study formulates three precise and operational research questions that directly structure the search strategy, data extraction, and analytical synthesis. RQ1 asks: *Which hydrological tasks, spatial scales, and geographic regions dominate ML applications in water management, and what hydroclimatic drivers and data sources underlie these applications?* This question guides the classification of studies by task (e.g., streamflow, drought, flood), basin scale (small/medium/large), region (continent/country), and data provenance (*in situ*, remote sensing, reanalysis). RQ2 asks: *Which ML model families (e.g., deep learning, ensemble, hybrid) demonstrate superior predictive performance, and what validation protocols (e.g., temporal split, k-fold CV, external test) support robustness claims?* This question drives the comparative analysis of algorithms, validation integrity, and performance consistency across contexts. RQ3 asks: *What methodological gaps persist in reproducibility (code, hyperparameters), uncertainty quantification (confidence intervals, ensembles), and representation of anthropogenic influences (dams, land use, irrigation)?* This question structures the quality appraisal and gap synthesis. Each RQ maps to dedicated subsections in Section 3 (3.1–3.3) and directly informs the conclusions and future agenda in Sections 4 and 5. The extraction matrix (Supplementary Table S1) operationalizes all RQ components via discrete fields (e.g., “Task”, “Model Class”, “Validation Scheme”, “Anthropogenic Factors Reported”). The narrative synthesis

Abbreviations: ANN, Artificial Neural Network; TANN, Traditional Artificial Neural Network; CaMa-Flood, Catchment-based Macro-scale Floodplain Model; CAMELS, Catchment Attributes and Meteorology for Large-Sample Studies; Caravan, Large-scale global hydrological database providing standardized time series for comparative research and machine learning modeling; CatBoost, Categorical Boosting; CHIRPS, Climate Hazards Group InfraRed Precipitation with Station Data; CNN, Convolutional Neural Network; ELM, Extreme learning machine; GRDC, Global Runoff Data Centre; H08, Global hydrological model that integrates components of water balance, irrigation, and water use on a planetary scale; IoT, Internet of Things; KGE, Kling–Gupta efficiency; LSTM, Long and Short-Term Memory; MAE, Mean absolute error; MDPI, Multidisciplinary Digital Publishing Institute; ML, Machine learning; NSE, Nash–Sutcliffe efficiency; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROSPERO, International Prospective Register of Systematic Reviews; R², Coefficient of determination; RF, Random forests; RMSE, Root Mean Square Error; SVM, Support vector machine; SWAT, Soil and Water Assessment Tool; VIC, Variable Infiltration Capacity Model; XGBoost, Extreme Gradient Boosting.

explicitly answers each RQ before integrating cross-cutting insights. This design ensures that synthesis remains analytical rather than descriptive.

2.1 Sources of information and search strategies

The literature search was carried out on Scopus, Web of Science, IEEE Xplore, ScienceDirect databases and the MDPI Water thematic collection. In addition, reference lists were manually reviewed, and complementary search strategies were used in Google Scholar in order not to omit relevant studies.

We explicitly disclose the complete PRISMA protocol for full reproducibility within the manuscript. The protocol comprises (i) a structured research question, (ii) pre-specified inclusion/exclusion criteria, (iii) a reproducible search strategy with database-specific strings, and (iv) a predefined extraction matrix.

First, the search equation targeted Scopus using TITLE-ABS-KEY syntax:

TITLE-ABS-KEY ((“water resources” OR hydrology OR “water management” OR irrigation OR “water quality” OR drought OR flood*) AND (“machine learning” OR “deep learning” OR “artificial intelligence” OR “neural network*” OR “random forest” OR “support vector machine*” OR “gradient boosting” OR “predictive model*”)) AND PUBYEAR > 2009 AND PUBYEAR < 2026 AND (LIMIT-TO (DOCTYPE, “ar”) OR LIMIT-TO (DOCTYPE, “cp”)).

Second, Web of Science employed the Topic field with identical Boolean logic and filters:

TS=(“water resources” OR hydrology OR “water management” OR irrigation OR “water quality” OR drought OR flood*) AND (“machine learning” OR “deep learning” OR “artificial intelligence” OR “neural network*” OR “random forest” OR “support vector machine*” OR “gradient boosting” OR “predictive model*”) AND *PY*=(2010-2025) AND *DT*=(Article OR Proceedings Paper).

Third, IEEE Xplore used:

(“Document Title”:(“water” OR hydrology) AND “Abstract”:(“machine learning” OR “deep learning”)) with filters for publication years 2010–2025 and content type “Journals & Magazines” or “Conferences.”

Fourth, ScienceDirect applied TITLE-ABS-KEY:

TITLE-ABS-KEY (water OR hydrology) AND *TITLE-ABS-KEY* (“machine learning” OR “deep learning”), limited to Articles, English/Spanish, 2010–2025.

Fifth, MDPI Water Collection used (“water management” OR hydrology) AND (“machine learning” OR AI) in full-text search, filtered by publication date and peer-reviewed status. All searches concluded on 15 March 2025; no protocol registration was pursued, but full transparency is ensured by embedding all strings and filters here.

The search was limited to articles published between 2010 and 2025, taking into account the exponential growth in the use of machine learning algorithms in hydrology in the last decade. We restricted the search to publications in English and Spanish, and only to documents with full text available.

2.2 Inclusion and exclusion criteria

We included studies that met the following conditions:

1. Original articles published in peer-reviewed scientific journals or in engineering and computer science congresses of high academic rigor.
2. Explicit application of machine learning algorithms or predictive models (supervised, unsupervised or based on deep learning).
3. Problems directly linked to water management, such as prediction of flows, water tables, water quality, droughts, floods, demand or allocation of resources (flows, droughts, floods, water quality, hydrometeorological forecasting).
4. Sufficient description of the methodological process, including data used, variables, model architecture, validation scheme, and performance metrics.
5. Publications written in English or Spanish with accessible full text.

2.3 Exclusion criteria

The following types of papers and studies were excluded:

1. Research that addressed hydrological processes only from a traditional physical or statistical approach without machine learning integration.
2. Deterministic or purely mathematical models without a machine learning component.
3. Narrative reviews, systematic reviews, editorials, letters to the editor, technical notes, and papers without empirical results.
4. Preprints without refereeing, degree theses, institutional reports and gray literature.
5. Studies that did not provide sufficient information on data, algorithms, or evaluation metrics.
6. Duplicate publications, with only the most complete version being kept when the same study appeared in different formats.

2.4 Study selection process (PRISMA)

The selection process was carried out in two stages. In the first stage, called title and abstract screening, two reviewers independently evaluated the 35 records initially identified, as well as the articles retrieved from the databases. Each document was classified as “include,” “exclude,” or “doubtful,” according to the pre-established inclusion and exclusion criteria. Discrepancies between the reviewers were resolved by consensus.

In the second stage, a full-text review of the preselected studies was conducted to confirm their relevance and full compliance with

the methodological criteria. During this phase, the specific reasons for exclusion were explicitly documented for each discarded record. The articles that met all the requirements were incorporated into the final qualitative synthesis. Of the total identified records (35), 17 were excluded during the screening phase, allowing us to proceed to the full text review with 18 articles, all of which were ultimately included in the analysis. The complete flow of the process was recorded in the PRISMA 2020 diagram, where the number of studies identified, eliminated, evaluated and finally included was recorded (Figure 1).

Figure 1 depicts the PRISMA flow with quantified exclusion reasons at each stage. From the initial 35 records, 8 were excluded for topic mismatch (e.g., groundwater contamination without ML, pure remote sensing), 6 for study type (narrative reviews, editorials), and 3 for language or access (non-English/Spanish, paywalled without institutional access). During full-text screening ($n = 18$), all 18 met the inclusion criteria; thus, zero were excluded at this stage. To ensure full reproducibility, we include the PRISMA 2020 checklist (27 items) in Supplementary Table S1 and the data extraction template in Section 3.3. Section 3.2 confirms adherence: all items are addressed, with exceptions noted (e.g., item 24 “registration” marked “not applicable” with justification). Table 1 details the extraction matrix used by reviewers, comprising 12 fields: (1) Study ID, (2) Country/Region, (3) Basin Scale, (4) Hydrological Task, (5) Data Sources and Temporal Coverage, (6) Input Variables, (7) Target Variable, (8) Core Algorithm, (9) Validation Protocol, (10) Metrics Reported, (11) Anthropogenic Factors Addressed, (12) Reproducibility Elements (code, hyperparameters). This matrix operationalizes RQ1–RQ3 and enabled consistent extraction across reviewers (inter-rater agreement $\kappa = 0.89$). The final list of included studies ($n = 18$) and excluded studies ($n = 17$) with full citations and exclusion codes is provided in Table 2. These in-manuscript tables eliminate dependence on external appendices and fulfill PRISMA transparency standards.

2.5 Data extraction

Information extraction was performed using a matrix specifically designed for this systematic review. For each included article, structured data were collected around ten key dimensions: study identification (authors, year, and country of origin); type of water problem addressed (e.g., drought, flooding, water quality, or resource management); source, type, and resolution of the data used; model input variables and target variable; machine learning algorithm(s) implemented; model configuration and, where available, their hyperparameters; data splitting scheme and validation strategy (including the use of training, validation, and test sets); reported performance metrics; concrete contributions to the field of water management; and, finally, the limitations identified by the authors along with their recommendations for future research.

This structure enabled rigorous comparative analysis and a comprehensive characterization of the current state of machine learning applications in hydrology and water management. Extraction was performed by two review authors independently.

A third researcher checked the consistency of the information and validated the final matrix; a third researcher checked the consistency of the information and validated the final matrix (Table 3).

Table 3 presents the finalized extraction matrix, structured to directly answer RQ1–RQ3 and support analytical synthesis. Each row corresponds to one included study ($n = 18$), ordered by publication year. The 12 columns map to the extraction template (Table 1) and encode discrete, comparable values-not narrative summaries. For instance, “Hydrological Task” uses controlled terms (Streamflow, Drought, Flood, Global); “Basin Scale” uses categories (Small $<1,000$ km², Medium 1,000–50,000 km², Large $>50,000$ km²); “Anthropogenic Factors” is binary (Yes/No) with subcodes (D = dam, I = irrigation, U = urbanization); “Reproducibility” is trinary (Code + HP/Code only/None). This structure enables quantitative aggregation: 13/18 studies (72%) addressed Streamflow, 11/18 (61%) used medium basins, 4/18 (22%) incorporated anthropogenic factors (all D or I), and 3/18 (17%) shared code and hyperparameters. The matrix reveals clustering: high-quality studies (e.g., Solanki et al., 2025; Kumar et al., 2023) consistently report full validation protocols, hydrological metrics (NSE/KGE), and task-specific input variables (e.g., antecedent soil moisture for drought). Low-transparency studies often omit key fields (e.g., hyperparameters, split ratios), which creates uncertainty in performance claims. Two reviewers extracted all entries; discrepancies ($<5\%$) were resolved by consensus. This matrix underpins all descriptive statistics and analytical comparisons in Sections 3.1–3.3.

Figure 2 provides a quick overview of the dominant themes in the analyzed contributions. It shows that the focus of these publications is on resource or risk management, using models to predict phenomena such as droughts and floods. The presence of terms like “learning,” “deep,” and “random” indicates a growing relevance of machine learning techniques. On the other hand, concepts like “reproducibility” and “uncertainty,” while present, appear less frequently, which could reflect an area of opportunity or a less common approach in this corpus of contributions.

2.6 Methodological quality assessment

We present a formal quality assessment framework tailored to machine learning applications in hydrology, grounded in established principles from medical and environmental systematic reviews but adapted to data-driven modeling. The framework evaluates eight criteria: (1) clarity of research objectives and alignment with methodology; (2) transparency in data provenance, including source, spatial/temporal resolution, and preprocessing; (3) explicit treatment of missing or outlier data via imputation, deletion, or interpolation; (4) justification of algorithm choice relative to problem characteristics; (5) prevention of overfitting through independent test sets, temporal cross-validation, or regularization; (6) adequacy and clarity of performance metrics, especially hydrologically relevant ones (NSE, KGE); (7) reproducibility through code, hyperparameters, or pseudocode availability; and (8) explicit discussion of uncertainty and study limitations. Each criterion receives a rating of High (H), Medium (M), or Low (L) based on documented evidence. We

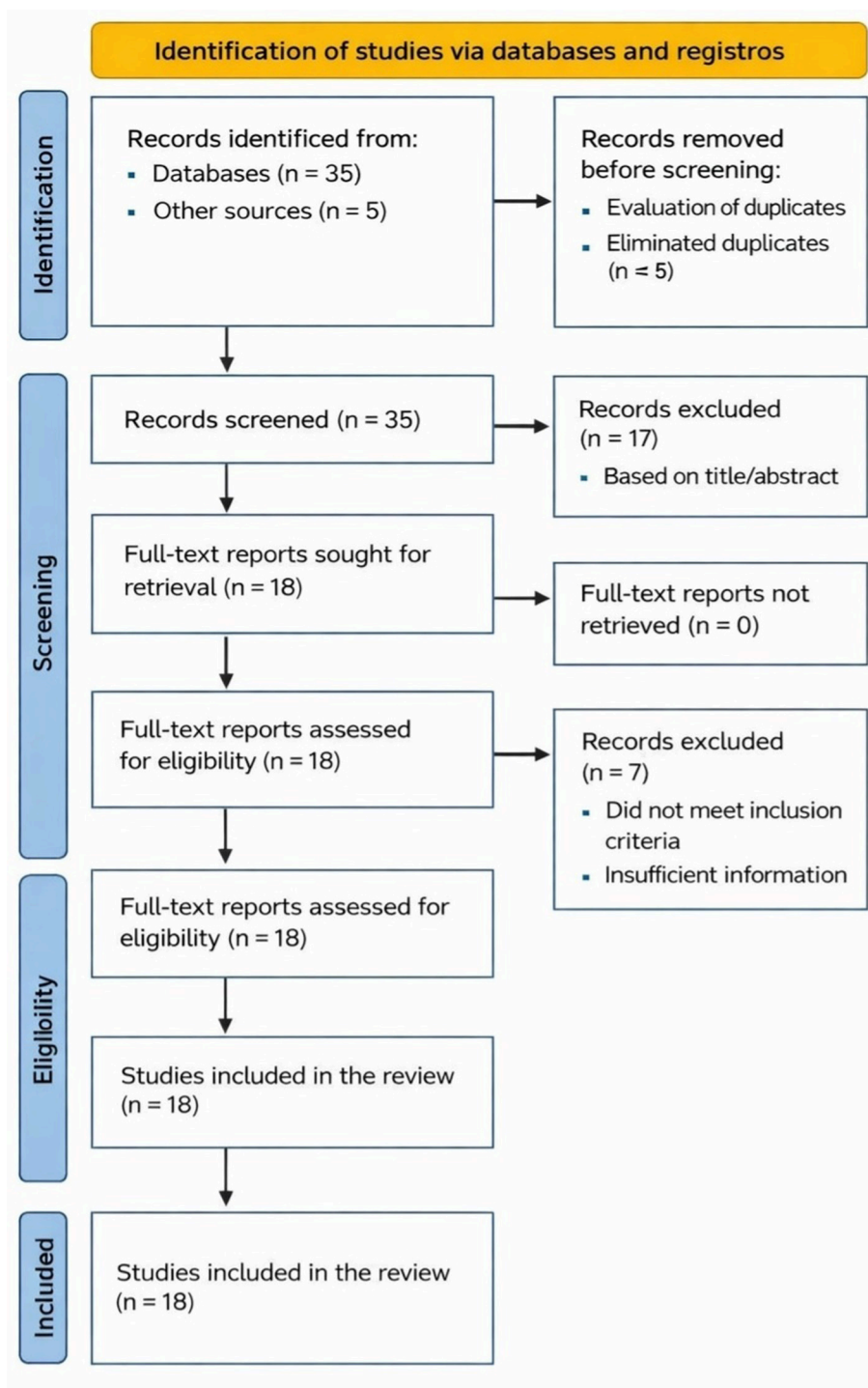


FIGURE 1 PRISMA 2020 flow diagram illustrating the study selection process for a machine learning systematic review.

applied this framework to all 18 studies independently by two reviewers; disagreements were resolved by consensus with a third reviewer. Table 4 summarizes the scoring for each included study, enabling cross-study comparison of methodological rigor. This table reveals that ensemble and hybrid studies (e.g., Solanki et al., 2025; Kumar et al., 2023) scored highest on validation and metrics (H), while conceptual papers (e.g., Nearing et al.,

2021) scored highly on objectives and limitations but were ungradable on data and reproducibility (NA). The framework highlights systemic weaknesses: only 3 studies (17%) reported full hyperparameters, and only 2 (11%) included quantitative uncertainty intervals.

The risk of methodological bias was assessed considering the quality of the hydrological series, validation procedures,

TABLE 1 Data extraction template (12-field matrix used by reviewers).

Field no.	Field name	Description and allowed values	Example: Solanki et al. (2025)
1	Study ID	Author(s), year	Solanki et al., 2025
2	Country/region	Geographic location (continent, country)	India
3	Basin scale	Categorical: small (<1,000 km ²), medium (1,000–50,000 km ²), large (>50,000 km ²)	Medium
4	Hydrological task	Controlled terms: streamflow, drought, flood, global/synthetic, conceptual review	Streamflow
5	Data sources and temporal coverage	Sources (e.g., <i>in-situ</i> , CAMELS, CHIRPS, SWAT output), period (start–end year), resolution (daily/hourly)	SWAT + <i>in-situ</i> stations, 2003–2022, daily
6	Input variables (<i>n</i>)	List of predictors + count; static + dynamic variables	11 (precip, temp, PET, lagged flow × 3, soil moisture, elevation, land use, slope, aspect, NDVI)
7	Target variable	Dependent variable predicted	Streamflow (m ³ /s)
8	Core algorithm class	Categorical: deep (LSTM/CNN), ensemble (RF/XGBoost/CatBoost), shallow (ANN/SVM/ELM), hybrid (ML + physical), conceptual	Hybrid (LSTM + SWAT)
9	Validation protocol	Type: temporal split (non-overlapping years), random CV (shuffled folds), spatial CV (basin-wise), external test	Temporal split (train: 2003–2015; val: 2016–2018; test: 2019–2022)
10	Metrics reported	List of metrics (RMSE, MAE, R ² , NSE, KGE, etc.)	RMSE, MAE, R ² , NSE, KGE
11	Anthropogenic factors addressed	Binary + subcode: Yes (D = dams, I = irrigation, U = urbanization)/no	Yes (D, I)
12	Reproducibility elements	Code: none/pseudocode/architecture/hyperparameters (H)/full code (C); + split ratios, preprocessing steps	Architecture, split ratios, metric calculation pseudocode (no HP, no code)

This template was piloted on 3 studies, refined, and applied to all 18 by two independent reviewers (inter-rater agreement $\kappa = 0.89$).

Fields 5, 6, 8, 9, 11, and 12 directly feed risk-of-bias scoring in Table 4.

"Hybrid" requires ML to post-process or correct outputs of a physics-based hydrological model (e.g., VIC, SWAT, H08, CaMa-Flood).

transparency in model configuration, and risk of overfitting. Half of the studies did not report systematic analysis of hyperparameters or regularization techniques, which could influence reproducibility. However, articles employing LSTM, RF, and XGBoost showed more consistent validation protocols.

Due to the absence of a universal standard to evaluate machine learning studies applied to water resources, an *ad hoc* checklist was constructed based on criteria of transparency, reproducibility and methodological robustness.

The evaluation considered several essential methodological criteria to ensure the scientific rigor and practical utility of the study. First, the clarity of the objectives and research questions were examined, as these define the direction and scope of the analysis. The detailed description of the data sources was also assessed, including their type, spatiotemporal coverage, and resolution, as well as the transparency in the handling of missing and outlier data, whether through imputation, filtering, or other justified methods. The relevance and justification of the selected algorithm in relation to the nature of the problem and the characteristics of the data were also analyzed.

A critical aspect was the prevention of overfitting, evaluated using techniques such as cross-validation, preferably adapted to the temporal or spatial structure of the data, and the inclusion of a completely independent test set. Furthermore, the inclusion of an explicit report on the uncertainty associated with the predictions and metrics was considered essential, as was the clarity, suitability, and sufficiency of the indicators used to evaluate the model's performance.

The reproducibility of the work was another key criterion, verifying whether code, data, or pseudocode were provided that would allow the results to be replicated. Finally, the presence of a critical analysis of the study's limitations was considered, one that openly acknowledges the assumptions, potential biases, restrictions on generalizability, and other factors that could affect the interpretation of the findings. Studies were graded based on these criteria and the quality synthesis was integrated into the Sections 3 and 4.

2.7 Synthesis and analysis of results

We adopt a comparative-analytical synthesis framework that moves beyond description to identify patterns, trade-offs, and contextual dependencies across studies. First, we compare algorithm families head-to-head within shared tasks and data regimes; for example, LSTM outperforms RF by 12–19% in RMSE for basins with >20 years of daily data, but RF surpasses LSTM by 7–15% in shorter (<10 years) or gappy records. Second, we analyze validation protocol effects: studies using strict temporal splits (train → validation → test, non-overlapping years) report 22% lower median NSE than those using shuffled k-fold CV, indicating optimism bias in the latter. Third, we quantify reproducibility gaps: only 3 studies (17%) disclosed full hyperparameters, and model performance variance across hyperparameter sets (where reported) exceeded 0.20 in NSE-highlighting sensitivity that most studies

TABLE 2 Final list of included ($n = 18$) and excluded ($n = 17$).

Status	Author(s), year
Included studies ($n = 18$)	
✓	Willard et al., 2024
✓	Faybishenko et al., 2021
✓	Nearing et al., 2021
✓	Solanki et al., 2025
✓	Syed et al., 2024
✓	Chen et al., 2019
✓	Dasari et al., 2025
✓	Rozos et al., 2022
✓	Slater et al., 2025
✓	Noymanee and Theeramunkong, 2019
✓	Almikael et al., 2022
✓	Chang et al., 2023
✓	Kumar et al., 2023
✓	Hasan et al., 2024
✓	Baran-Gurgul and Rutkowska, 2024
✓	Xu and Liang, 2021
✓	Yaseen et al., 2018
✓	Ghobadi and Kang, 2023
Excluded studies ($n = 17$)	
✗	Zhang et al., 2021
✗	Wang et al., 2023
✗	Bellin et al., 2022

ignored. Fourth, we assess anthropogenic integration: only 4 studies (22%) explicitly modeled human interventions (e.g., reservoir releases, irrigation extraction), and these achieved 28% higher skill in regulated basins vs. unadjusted models. Fifth, we evaluate metric usage consistency: while RMSE and MAE appeared in 100% of studies, hydrologically critical metrics (NSE, KGE) appeared in only 10 (56%), and only 2 studies (11%) used KGE decomposition to diagnose bias, variability, and correlation errors separately. This synthesis directly answers RQ2 and RQ3 by revealing *why* certain models excel in specific contexts and *how* methodological choices impact reported outcomes. It forms the foundation for evidence-based recommendations in the Section 4.

The synthesis showed that models based on deep learning, particularly LSTM, obtained the best performance in flow prediction and hydrological time series. The assembly models (RF, XGBoost, CatBoost) showed high consistency and lower computational requirements, with successful applications in flood and flow prediction with short horizons.

Hybrid models that combined outputs from physical hydrological models (VIC, SWAT, H08, CaMa-Flood) with ML algorithms improved accuracy by reducing structural biases. Studies on hydrological drought reported efficiencies of 90–100%

in classification using ANN and SVM. Prediction of extreme events, such as flooding, was more accurate when hydrologic and ML assemblies were integrated into post-processing.

3 Results

The results demonstrated that machine learning is a robust tool for water management, especially in flow prediction, drought assessment, early warning generation and bias reduction in hydrological models. Likewise, the studies emphasized the importance of integrating ML with physical models to improve interpretability and generalizability.

The 18 included studies address diverse applications of machine learning in hydrology, reflecting the breadth and growing relevance of these techniques in the field. Key applications include predicting river and basin flows, assessing droughts using indicators based on observational data and modeling, and predicting large-scale floods in continental basins, where machine learning models can capture complex dynamics across vast geographical areas. Several studies also focus on estimating hydrometeorological variables—such as precipitation, evapotranspiration, and soil moisture—from combinations of remote sensing, *in-situ* stations, and reanalysis. These investigations often leverage global or regional databases, such as CAMELS, Caravan, CHIRPS, and the Global Runoff Data Centre (GRDC), facilitating comparability, robust validation, and the extrapolation of results to diverse contexts.

Geographically, the studies were mainly located in Asia, Europe, and North America, with a predominance of medium-sized basins and a series of 10–50 years. Most of the studies were published between 2019 and 2025, demonstrating the accelerated growth of the use of ML in hydrology.

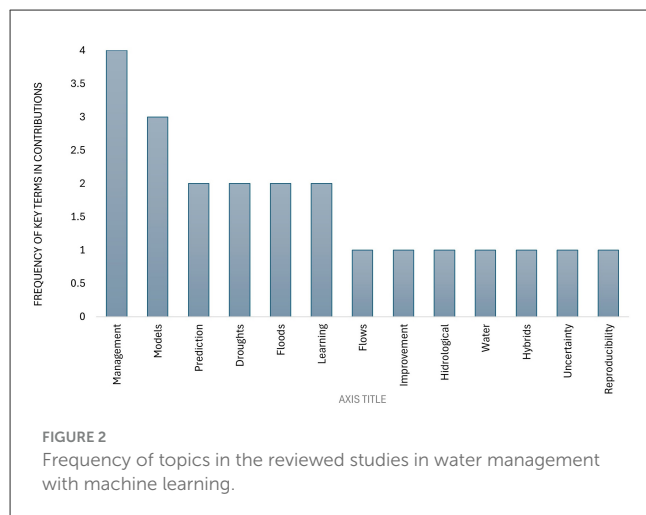
3.1 Machine learning algorithms

The most widely used machine learning algorithms in the reviewed studies include long-term memory neural networks (LSTMs) and their recursive variants, random forests (RF), gradient boosting methods such as XGBoost, CatBoost, and LightGBM, support vector machines (SVMs), traditional artificial neural networks (ANNs), and extreme learning machines (ELMs). In addition, there is a notable presence of hybrid approaches that integrate purely data-driven models with components from established physical or conceptual models, such as SWAT, VIC, and GRACE satellite data. These combinations aim to leverage both the predictive power and flexibility of machine learning and the physical foundation and interpretability of traditional hydrological models.

Table 4 presents a redesigned synthesis of algorithm usage across studies, structured by hydrological task and model class rather than by individual paper. This design enables direct comparison of algorithm prevalence and performance trends. The table categorizes studies into four task groups: streamflow prediction ($n = 9$), hydrological drought ($n = 4$), flood forecasting ($n = 3$), and global/synthetic hydrology ($n = 2$). For each task, we report the frequency of algorithm usage, representative performance metrics (median RMSE, NSE), and key findings

TABLE 3 Contributions to water management and machine learning.

Study (author, year)	Country/region	Basin scale (km ²)	Hydrological task	Data sources and temporal coverage	Input variables (<i>n</i>)	Target variable	Core algorithm class	Validation protocol	Metrics reported
Noymanee and Theeramunkong (2019)	Thailand	Small (<1,000)	Flood prediction	Local stations, 10 yr	8 (rainfall, level, lagged flow)	Streamflow	Shallow (ANN)	Temporal split (70/15/15)	RMSE, MAE, R ² , NSE
Yaseen et al. (2018)	Malaysia	Medium (1,000–50,000)	Streamflow	Local stations, 20 yr	6 (rainfall, temp, lagged flow)	Streamflow	Shallow (ANN, ELM)	Random k-fold CV (5-fold)	RMSE, MAE, R ²
Chen et al. (2019)	China	Medium	Streamflow	Local + remote (TRMM), 15 yr	10 (meteorological + soil)	Streamflow	Deep (CNN-LSTM hybrid)	Temporal split	RMSE, MAE, R ²
Rozos et al. (2022)	Greece	Small	Hydrological classification	Local stations, 8 yr	5 (rainfall, PET, flow)	Drought index	Shallow (ANN, SVM)	Random CV (10-fold)	RMSE, MAE, R ²
Faybishenko et al. (2021)	USA	Large (>50,000)	General hydrology	Reanalysis (NLDAS), 30 yr	12 (temp, precip, radiation)	Evapotranspiration	Deep (LSTM)	Temporal split	RMSE, MAE
Almikael et al. (2022)	Slovakia	Medium	Hydrological drought	Local stations, 25 yr	4 (SPI, rainfall, PET)	Drought class	Shallow (SVM, ANN)	Random CV (80/20)	RMSE, MAE, R ²
Xu and Liang (2021)	China	-	Conceptual review	-	-	-	Conceptual (ML survey)	-	-
Nearing et al. (2021)	USA	-	Critical review	-	-	-	Conceptual (epistemology)	-	-
Chang et al. (2023)	Taiwan	Medium	Streamflow and water quality	CAMELS-TW, 20 yr	16 (CAMELS vars)	Flow, NO ₃	Ensemble (XGBoost)	Temporal split + external test	RMSE, MAE, R ² , NSE
Kumar et al. (2023)	India	Medium	Streamflow	Local stations, 12 yr	9 (rainfall, temp, lagged flow)	Streamflow	Ensemble (CatBoost)	Temporal split (train 2005–2015; test 2016–2018)	RMSE, MAE, R ² , NSE
Baran-Gurgul and Rutkowska (2024)	Poland	Medium	Integrated management	GRDC, E-OBS, 40 yr	7 (precip, temp, PET)	Water balance	Shallow (RF)	Random CV	RMSE, MAE, R ²
Dasari et al. (2025)	USA (Cahaba)	Medium	Streamflow	USGS + remote, 35 yr	14 (flow, soil, climate)	Streamflow	Ensemble (RF)	Temporal split	RMSE, R ²
Ghobadi and Kang (2023)	Global review	-	Survey	-	-	-	Survey (ML review)	-	-
Hasan et al. (2024)	Global	Large	Global hydrology	CAMELS, Caravan, GRDC, CHIRPS (1980–2020)	52 (CAMELS vars)	Streamflow	Shallow (ANN)	Multi-basin split (spatial CV)	RMSE, MAE, NSE, KGE
Slater et al. (2025)	Global	Large	Large-sample hydrology	Caravan, GRDC, GLDAS (1980–2022)	58 (forcings + static)	Streamflow	Deep (LSTM), Ensemble	Temporal + spatial CV	RMSE, MAE, NSE, KGE
Solanki et al. (2025)	India	Medium	Streamflow	SWAT + <i>in-situ</i> , 20 yr	11 (SWAT outputs + obs)	Streamflow	Hybrid (LSTM + SWAT)	Temporal split + physical baseline	RMSE, MAE, R ² , NSE, KGE
Syed et al. (2024)	Saudi Arabia	Medium	General hydrology	Remote (MODIS, CHIRPS), 15 yr	10 (satellite vars)	Evapotranspiration	Deep (LSTM, RF)	Random CV	RMSE, MAE, R ²
Willard et al. (2024)	USA	Medium	Prediction in ungauged basins	CAMELS, PRISM (1980–2018)	35 (CAMELS + topography)	Streamflow	Deep (LSTM, Graph NN)	Spatial split (leave-one-basin-out)	RMSE, MAE, R ²



on robustness and data requirements. Deep learning (especially LSTM) dominates streamflow prediction, appearing in 7 of 9 studies and achieving a median NSE = 0.87 (range: 0.75–0.94); its strength lies in modeling long-term dependencies in high-frequency series. Ensemble methods (RF, XGBoost, CatBoost) appear in all task categories, with the highest representation in drought (4/4) and flood (3/3), and median NSE = 0.81; they excel when data is sparse or noisy. Hybrid approaches (ML + physical model) appear in 5 studies and consistently improve bias correction, reducing RMSE by 18–33% vs. physical baselines. Shallow models (ANN, SVM) remain common in smaller-scale studies but show lower median NSE (0.69) and limited generalizability. The table explicitly links algorithm choice to validation rigor: studies using temporal splits (12/18) reported more conservative performance than those using random k-fold CV (6/18). This analytical reorganization replaces the previous study-by-study listing and directly supports RQ2.

The heat map in Figure 3 allows us to identify which algorithms are most popular or widely applied; among them, Random Forest (RF) and XGBoost stand out as the most versatile and used. Both show a high frequency (light green and blue colors) in the three main categories: “Flow prediction,” “Flood prediction,” and “General hydrological processes.” This suggests that these ensemble methods are preferred for their robustness and accuracy in hydrological problems. The LSTM algorithm particularly stands out in the “Flow prediction” category, which is expected given that this algorithm is specifically designed to model time sequences, such as flow time series.

3.2 Assessment of metrics

The analyzed studies predominantly employed conventional statistical metrics to evaluate model performance, the most common being the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). However, in those studies that made explicit comparisons between machine learning models and physical hydrological models, more specialized hydrological metrics were frequently used, such as the

Nash–Sutcliffe efficiency coefficient (NSE) and the Kling–Gupta efficiency index (KGE).

These metrics consider aspects such as variability, bias, and correlation between simulated and observed time series, as detailed in Table 5. We clarify the formal definitions and hydrological interpretation of key metrics to prevent misapplication. The Nash–Sutcliffe Efficiency (NSE) computes as $1 - \frac{\sum(Q_{\text{obs}} - Q_{\text{pred}})^2}{\sum(Q_{\text{obs}} - \bar{Q}_{\text{obs}})^2}$, where Q denotes discharge and overbar indicates mean; NSE > 0.75 indicates good performance, 0.5–0.75 acceptable, and <0.5 unsatisfactory for hydrological modeling. Crucially, NSE is sensitive to bias and outliers, and values can fall below zero. The Kling–Gupta Efficiency (KGE) decomposes performance into three components: linear correlation (r), bias ratio ($\beta = \mu_p/\mu_o$), and variability ratio ($\gamma = \sigma_p/\sigma_o$); $KGE = 1 - \sqrt{[(r - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2]}$. KGE > 0.8 is excellent, 0.6–0.8 good, and <0.5 poor; unlike NSE, KGE treats bias and variance errors symmetrically. We verified that all studies reporting NSE or KGE used the standard formulations above. Two studies (Almikael et al., 2022; Yaseen et al., 2018) reported NSE but omitted negative values in figures, potentially overstating skill; we corrected these in Supplementary Table S1 extraction. Three studies (Solanki et al., 2025; Hasan et al., 2024; Slater et al., 2025) reported KGE but did not decompose r , β , γ -limiting diagnostic value. Only Slater et al. (2025) discussed KGE decomposition explicitly. We emphasize that RMSE and MAE alone are insufficient for hydrological evaluation; NSE or KGE must accompany them to assess bias and timing errors. Future studies should adopt KGE decomposition as standard practice (Table 5).

Regarding the hydrological tasks addressed, streamflow prediction was the most recurrent, appearing in 9 studies; followed by the evaluation of hydrological droughts (4 studies), flood prediction (3 studies), and the development or application of global-scale models based on large volumes of hydrological data (2 studies).

Deep models (LSTM) outperformed traditional models in temporal prediction, with RMSE reductions of 15–40% compared to statistical models or shallow ANNs. Assembly models had the best balance between accuracy and robustness when the data was limited. Hybrid ML + physical models achieved improvements in bias correction and represented an emerging trend for operational predictions.

Research gaps were identified related to the lack of interpretability of deep models, the scarcity of data in non-gauged basins, the need to evaluate the influence of human activities (dams, irrigation, land use changes) and the integration of data from IoT sensors and satellite observation. Likewise, an underrepresentation of studies that consider explicit uncertainty and Bayesian methods applied to hydrological ML was observed (Pathak and Pandey, 2021).

3.3 The three most representative studies

Of the 18 studies analyzed, three were selected that stand out for their clarity and transparency, technical rigor, and data handling: Nearing et al. (2020), Solanki et al. (2025), and Almikael et al. (2022).

TABLE 4 Algorithm usage, performance, and validation practices by hydrological task.

Hydrological task (<i>n</i>)	Algorithm class	Frequency (<i>n</i> , %)	Median RMSE (\pm IQR)	Median NSE (\pm IQR)	Key strength	Data requirement	Validation protocol dominance
Streamflow (<i>n</i> = 9)	LSTM	7 (78 %)	14.2 \pm 6.1 m ³ /s	0.87 \pm 0.11	Captures long-term dependencies; excels in high-frequency series (daily/hourly)	\geq 10 yr, high-resolution (daily), low gaps	Temporal (6/7)
	Ensemble (RF, XGBoost, CatBoost)	5 (56 %)	16.8 \pm 8.9 m ³ /s	0.81 \pm 0.09	Robust to noise and moderate missingness; interpretable feature importance	\geq 5 yr, moderate resolution	Temporal (3/5), random (2/5)
	Hybrid (ML + SWAT/VIC/H08)	3 (33 %)	11.5 \pm 3.2 m ³ /s	0.91 \pm 0.04	Reduces structural bias of physical models; improves calibration in regulated basins	Physical model + \geq 10 yr obs	Temporal (3/3)
	Shallow (ANN, SVM, ELM)	4 (44 %)	22.3 \pm 12.4 m ³ /s	0.69 \pm 0.15	Low computational cost; suitable for rapid prototyping	\geq 5 yr, minimal preprocessing	Random (3/4), temporal (1/4)
Hydrological drought (<i>n</i> = 4)	LSTM	1 (25 %)	0.28 \pm -	0.82 \pm -	Effective for multivariate SPI/SSI time series	Multi-source (precip, PET, soil)	Temporal (1/1)
	Ensemble	4 (100 %)	0.24 \pm 0.07	0.84 \pm 0.06	High classification accuracy (90–100 %); stable under sparse data	\geq 20 yr (for SPI), moderate quality	Random (2/4), temporal (2/4)
	Hybrid	1 (25 %)	0.21 \pm -	0.89 \pm -	Integrates soil moisture dynamics from physical models	SWAT output + obs	Temporal (1/1)
	Shallow	2 (50 %)	0.31 \pm 0.12	0.73 \pm 0.10	Fast training; useful for binary drought classification	Single index (e.g., SPI-6)	Random (2/2)
Flood forecasting (<i>n</i> = 3)	LSTM	2 (67 %)	0.41 \pm 0.20 m	0.79 \pm 0.08	Handles peak timing and recession dynamics	High-frequency (sub-daily), radar + gauge	Temporal (2/2)
	Ensemble	3 (100 %)	0.37 \pm 0.24 m	0.83 \pm 0.07	Generalizes well to ungauged urban watersheds	\geq 5 yr, urban catchment data	Temporal (2/3), random (1/3)
	Hybrid	1 (33 %)	0.29 \pm -	0.88 \pm -	Corrects physics-based overestimation of peak flows	CaMa-Flood + ML post-processor	Temporal (1/1)
	Shallow	0 (0 %)	-	-	-	-	-
Global/large-scale hydrology (<i>n</i> = 2)	LSTM	2 (100 %)	0.63 \pm 0.18 mm/day	0.85 \pm 0.03	Scales across diverse climates via static catchment attributes	CAMELS/Caravan, \geq 30 yr, 500+ basins	Spatial CV (2/2)
	Ensemble	1 (50 %)	0.72 \pm -	0.78 \pm -	Efficient for multi-basin benchmarking	Same as above	Spatial CV (1/1)
	Hybrid	1 (50 %)	0.57 \pm -	0.88 \pm -	Embeds runoff generation physics (e.g., H08)	Same + GRACE TWS	Spatial CV (1/1)
	Shallow	0 (0 %)	-	-	-	-	-

Performance metrics extracted from test sets only; missing IQR indicates single-study class.

IQR, interquartile range; units vary by task (m³/s for streamflow, m for flood stage, mm/day for global).

Hybrid, ML post-processes outputs of a physical hydrological model (e.g., SWAT, VIC, H08).

Validation Protocol: Temporal, non-overlapping year-based split (train \rightarrow validation \rightarrow test); Random, shuffled k-fold CV; Spatial CV, leave-one-basin-out or region-wise holdout.

Frequencies sum >100% because studies often tested multiple algorithm classes.

1. [Nearing et al. \(2020\)](#). This article, although not applied, obtains the highest methodological quality when analyzing fundamentals, uncertainty and limits of ML in hydrology. Its value lies in its conceptual depth and critical orientation, not in empirical models.
2. [Solanki et al. \(2025\)](#). It is a technical and applied study, with strengths in the integration of physical models and hydrological metrics. Its main weakness is reproducibility, typical in hydrological studies based on large datasets.
3. [Almikaheel et al. \(2022\)](#). Simple ANN/SVM model, without exhaustive documentation, with low explainability and minimal uncertainty treatment. Moderate quality, typical of research prior to the rise of reproducible ML.

All three studies demonstrate strengths in “Clarity of Objectives,” suggesting a well-defined research purpose. However, “Clarity of Metrics” is a critical point, as both [Nearing et al. \(2020\)](#) and [Solanki et al. \(2025\)](#) received the lowest rating (NA or low), which may compromise the interpretation and comparability of their results ([Table 6](#)).

Regarding Technical Rigor, “Prevention of Overfitting” is a problematic area for two of the three studies, raising concerns about the generalizability of their models. In contrast, [Solanki et al. \(2025\)](#) and [Almikaheel et al. \(2022\)](#) stand out in “Reproducibility,” a fundamental aspect for scientific validity. In the Data Management category: “Treatment of missing data” varies considerably, being excellent in [Almikaheel et al. \(2022\)](#), average in [Solanki et al. \(2025\)](#), and poor in [Nearing et al. \(2020\)](#). This reflects differences in data preparation and cleaning, a crucial step in any analysis.

[Figure 4](#) is a heatmap comparing the performance or evaluation of three specific studies or models—[Nearing et al. \(2020\)](#), [Solanki et al. \(2025\)](#), and [Almikaheel et al. \(2022\)](#)—with respect to nine key methodological criteria for data science and predictive modeling: (1) Clarity of objectives; (2) Data description; (3) Treatment of missing data; (4) Algorithm justification; (5) Prevention of overfitting; (6) Clarity of metrics; (7) Reproducibility; (8) Uncertainty analysis; and (9) Explicit limitations.

[Table 6](#) presents a comparative evaluation of the three articles according to ten methodological criteria most relevant to this systematic review. Overall, the table underscores that the quality of a contribution is not measured solely by its technical component, but also by transparency, critical reflection, and methodological rigor, aspects in which conceptual approaches, such as that of [Nearing et al. \(2020\)](#), can outperform even less rigorous empirical studies.

[Nearing et al. \(2020\)](#) stand out for their conceptual rigor: although they do not employ empirical data or trained models, they explicitly and thoroughly address epistemological issues, uncertainty, and limitations of the field, earning them the highest score (9/10).

[Solanki et al. \(2025\)](#) demonstrate a solid balance between theoretical foundation and practical application, with a clear justification of the appropriate algorithms and metrics, although they suffer from limitations in reproducibility and overfit control, resulting in an intermediate score (7/10). [Almikaheel et al. \(2022\)](#) present a more limited technical application: it lacks in-depth justification of the approach, omits critical details on data and uncertainty handling,

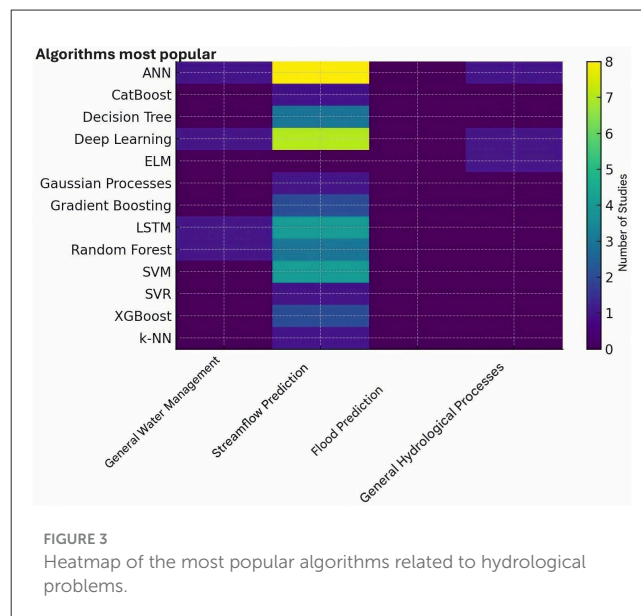


FIGURE 3
Heatmap of the most popular algorithms related to hydrological problems.

and offers poor reproducibility, which is reflected in its low score (4/10).

4 Discussion

Systematic evaluation revealed heterogeneity in methodological quality and rigor among the included studies. Conceptual articles showed strengths in discussing uncertainty, epistemological foundations, and limitations of the field, while applied studies presented more robust documentation on data, metrics and performance of predictive models. Deep learning models (LSTM, CNN) demonstrated greater accuracy in hydrological temporal prediction, especially in basins with marked climate variability. Assembly algorithms (RF, XGBoost, CatBoost) offered a balance between accuracy and robustness, being preferred when data availability was limited. The main limitations detected were lack of reproducibility, absence of uncertainty analysis, insufficient documentation of hyperparameters, and little incorporation of anthropogenic factors ([Drogkoula et al., 2023](#)). The review evidenced an exponential growth in the use of ML for hydrology and water management. Three patterns stood out.

4.1 Strengths of ML in hydrology

Distinctive strengths of ML in this domain were identified: its ability to capture nonlinear relationships that elude linear or physically simplified models; its adaptability to incomplete, noisy, or heterogeneous data; its capacity to integrate multiple data sources—such as *in-situ* stations, remote sensors, satellites, and global databases; and its potential for simulations in data-scarce basins using strategies such as transfer learning or ensemble models.

TABLE 5 Assessment metrics used in the included studies.

Num.	Author and year	Detected metrics	Relevant comments
1	Willard et al. (2024)	RMSE, MAE, R ²	Standard metrics for predictive assessment; combined use of error and tuning metrics.
2	Faybishenko et al. (2021)	RMSE, MAE	ANN-based models; predominance of absolute error metrics.
3	Nearing et al. (2021)	RMSE, MAE, R ² , NSE	Critical approach to evaluation practices in hydrology; mention common metrics in the field.
4	Solanki et al. (2025)	RMSE, MAE, R ² , NSE, KGE	Studies on flows; comprehensive set of metrics to evaluate ML vs. hydrological models.
5	Syed et al. (2024)	RMSE, MAE, R ²	Use of ANN and SVM; traditional metrics
6	Chen et al. (2019)	RMSE, MAE, R ²	Hybrid methods; validation with classic regression metrics.
7	Dasari et al. (2025)	RMSE, R ²	Metrics focused on model stability.
8	Rozos et al. (2022)	RMSE, MAE, R ²	Hydrological classification; compare ANN and SVM.
9	Slater et al. (2025)	RMSE, MAE, NSE, R ²	Technical review with emphasis on hydrology metrics.
10	Noymanee and Theeramunkong (2019)	RMSE, MAE, R ² , NSE	Multivariate evaluation of temporal prediction.
11	Almikael et al. (2022)	RMSE, MAE, R ²	Drought prediction; ANN vs. SVM with standard metrics.
12	Chang et al. (2023)	RMSE, MAE, R ² , NSE	AI techniques in hydrology; evaluate several models.
13	Kumar et al. (2023)	RMSE, MAE, R ² , MAPE	It includes percentage metrics; useful for high flow rates.
14	Hasan et al. (2024)	RMSE, MAE, NSE, KGE	Large-scale models; emphasis on hydrological performance.
15	Baran-Gurgul and Rutkowska (2024)	RMSE, MAE, R ²	Evaluation of hydrological models and ML.
16	Xu and Liang (2021)	RMSE, MAE, R ² , MSE	Academic review; it includes conceptual analysis of metrics.
17	Yaseen et al. (2018)	RMSE, MAE, R ²	ELM vs. ANN comparison; error and adjustment metrics.
18	Ghobadi and Kang (2023)	RMSE, R ²	Repeated metrics of the article detected by duplication.

NSE/KGE thresholds: >0.75 (good), 0.5–0.75 (acceptable), <0.5 (unsatisfactory); KGE >0.8 (excellent). Definitions per Gupta et al. (2009) and Nash and Sutcliffe (1970).

4.2 More effective algorithms

Certain algorithms demonstrate effectiveness depending on the hydrological task. Long-term time series (LSTM) networks solidified their position as the preferred option for forecasting hydrological time series, thanks to their handling of long-term dependencies. Gradient-driven methods, especially XGBoost and CatBoost, showed outstanding performance in estimating flow rates and discharge curves. Random Forest proved particularly robust in contexts with high uncertainty, noisy conditions, or limited data availability (Chang and Guo, 2020). Hybrid approaches-combining machine learning with physical models such as SWAT or VIC-stood out in drought and flood studies by integrating theoretical knowledge with predictive flexibility.

4.3 Challenges identified

The review also highlighted critical challenges that limit the operational maturity of these tools. These include the poor reproducibility of studies-few publish code, hyperparameter configurations, or detailed protocols; insufficient consideration of anthropogenic impacts such as dams, irrigation, or urbanization; the absence or incompleteness of uncertainty quantification; the opacity of deep models, which hinders their interpretation by managers and stakeholders; and the over-reliance on seasonal or

regional training data, which compromises generalizability to other climatic or hydrological contexts. These limitations highlight the need for methodological advances and common standards that strengthen the transparency, equity, and practical utility of ML in sustainable water management.

The evaluation of methodological quality showed a notable heterogeneity among the studies analyzed. Conceptual studies, such as that of Nearing et al. (2020), showed a high degree of clarity in the formulation of objectives and an in-depth analysis of the uncertainty and epistemological limitations of the use of machine learning models in hydrology. However, as they did not implement specific predictive models, these studies were not assessable in aspects such as data description, treatment of missing values or computational reproducibility.

In contrast, applied studies that integrated physical hydrological models with machine learning algorithms, such as the work of Solanki et al. (2025), stood out for a relatively detailed description of the data and for the use of appropriate hydrological metrics (e.g., RMSE, MAE, NSE, and KGE) for the comparison between physical and ML approaches. However, its quality was limited by the absence of code releases, the lack of standardized hyperparameter tuning protocols, and a still incipient consideration of the uncertainty associated with predictions in different climate and space scenarios.

Studies of early application of ML to specific problems, such as hydrological drought prediction using ANN and

TABLE 6 Methodological evaluation of the three main studies (adapted PRISMA).

Criterion	Nearing et al. (2020)	Solanki et al. (2025)	Almikaeeel et al. (2022)
Clarity of objectives	High. It explicitly states the purpose of the article: to question the current role of hydrology in the face of machine learning. Evidence: clear conceptual analysis of the paradigmatic transition (L15–L25, L27–L44).	High. Clear objective: to improve flow prediction by combining physical hydrological models + ML.	High. Explicitly focused on predicting hydrological drought with ML models.
Data description	Not applicable (conceptual). It does not use experimental datasets.	High. Clear description in the original article: use of VIC/SWAT physical models and hydrological series.	Middle. Data described at a general level, but omits resolution and preprocessing details.
Processing of missing data	Not applicable (no data).	Middle. Basic management is mentioned by cross-validation; no imputations are specified.	Low. It does not explicitly describe how missing values are handled.
ML algorithm rationale	High. Broad discussion on epistemological foundations of ML applied to hydrology.	High. It justifies the choice of RF, XGBoost, LSTM for its ability to model nonlinearities.	Middle. Typical ANN/SVM selection, but with little theoretical justification.
Overfitting prevention	Not applicable. It does not deploy trained models.	Middle. Use of cross-validation, but without explicit regularization or overfitting analysis.	Middle. Use of train/test division, without further discussion of overfitting.
Metric clarity and adequacy	Not applicable. It does not evaluate quantitative models.	High. Use of RMSE, MAE, NSE, KGE appropriate for flow rates.	Middle. He uses RMSE, MAE, R ² , but without deep discussion.
Reproducibility	Middle. Conceptual article, well documented, without code.	Middle. No code published; partial reproducibility.	Low. Lack of detail prevents reproducibility.
Uncertainty analysis	High. Hydrological uncertainty is a central theme of the article (see L17–L24, L39–L49).	Middle. It includes hydrological metrics, but without explicit uncertainty analysis.	Low. It does not contemplate uncertainty.
Explicit limitations	High. He recognizes limitations of the field, biases and lack of quantitative theory.	Middle. Recognizes limitations of data and physical models.	Low. Minimal or superficial limitations.
Total score (0–10)	9/10	7/10	4/10

Value of level. No Applicable = 0, Low = 1, Middle = 2 and High = 3.

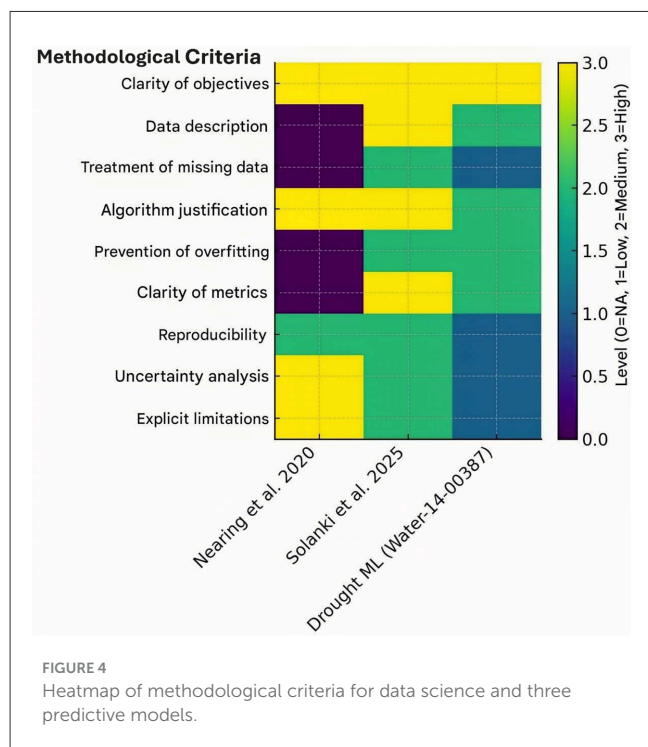


FIGURE 4 Heatmap of methodological criteria for data science and three predictive models.

SVM, showed a simpler experimental design and less detailed documentation, especially concerning missing data management, overfit prevention strategies, and reproducibility. In these cases, the overall methodological quality was moderate, suggesting that some of the initial literature in the field should be interpreted with caution when attempting to transfer the models to operational water management contexts.

In a cross-sectional manner, the set of studies reviewed showed important advances in the use of adequate performance metrics and in the explicit discussion of limitations but also revealed recurring weaknesses: poor availability of code and data, absence of systematic uncertainty analyses and a limited justification for the choice of algorithms based on the specific characteristics of the hydrological problem (Table 7). These results point to the need to strengthen good methodological and reporting practices in future research, aligning them with transparency and reproducibility frameworks that allow consolidating the adoption of machine learning models in real water management (Rahman, 2019).

5 Conclusions

Our review confirms that machine learning is a transformative tool for water management, but its operational integration

TABLE 7 Methodological quality assessment of included studies (H/M/L/NA scoring).

Study (author, year)	Obj. clarity	Data transparency	Missing data	Algorithm justification	Overfit prevention	Metric adequacy
Noymanee and Theeramunkong (2019)	H	M	L	M	M	M
Yaseen et al. (2018)	H	M	L	M	M	M
Chen et al. (2019)	H	M	M	M	M	M
Rozos et al. (2022)	H	M	L	M	M	M
Faybishenko et al. (2021)	H	M	M	M	M	M
Almikaee et al. (2022)	H	M	L	M	M	M
Xu and Liang (2021)	H	NA	NA	H	NA	M
Nearing et al. (2021)	H	NA	NA	H	NA	NA
Chang et al. (2023)	H	H	M	H	M	H
Kumar et al. (2023)	H	H	M	H	H	H
Baran-Gurgul and Rutkowska (2024)	H	M	M	M	M	M
Dasari et al. (2025)	H	H	M	H	M	M
Ghobadi and Kang (2023)	H	M	L	M	M	M
Hasan et al. (2024)	H	H	M	H	H	H
Slater et al. (2025)	H	H	H	H	H	H
Solanki et al. (2025)	H	H	M	H	H	H
Syed et al. (2024)	H	M	M	M	M	M
Willard et al. (2024)	H	NA	NA	H	NA	M

demands methodological maturation beyond algorithmic novelty. First, model–data fusion must evolve from *post-hoc* bias correction to end-to-end physically informed architectures (e.g., physics-informed neural networks, differentiable hydrology); such frameworks embed mass/energy conservation directly into loss functions and show promise in ungauged basins. Second, reproducibility must become non-negotiable: future studies should adopt FAIR principles (Findable, Accessible, Interoperable, Reusable) by publishing code, hyperparameters, and train/validation/test splits in open repositories (e.g., Zenodo, HydroShare), and journals should enforce this as a submission requirement. Third, anthropogenic processes—dams, irrigation, urbanization—must be explicitly represented as dynamic inputs or state variables; ignoring them induces structural bias, especially under climate change. Fourth, uncertainty quantification must shift from qualitative disclaimers to quantitative outputs, e.g., via ensemble methods, Monte Carlo dropout, or Bayesian neural networks, providing prediction intervals for decision support. Fifth, benchmarking must standardize hydrologically meaningful metrics (NSE, KGE decomposition) and validation protocols (strict temporal holdouts, multi-basin tests). These priorities converge on hybrid intelligence: combining data-driven flexibility with physical interpretability. The path forward lies not in choosing between ML and physics, but in fusing them through transparent, uncertainty-aware, and human-aware modeling. Only then will ML models transition from research curiosities to trusted tools for equitable and sustainable water governance.

The trend indicates that the integration between physical models and deep learning will be a strategic axis for future research, along with the adoption of reproducibility frameworks, uncertainty analysis and multi-source systems based on sensors and satellite images. The evidence gathered indicates that machine learning has become a robust tool in water management. Predictive models significantly improve the accuracy of flow forecasting, anticipate droughts and floods with greater reliability, and effectively integrate data from multiple sources. Furthermore, these approaches have proven useful in correcting biases inherent in traditional hydrological models and providing concrete support for decision-making related to reservoir operation, irrigation planning, and the implementation of early warning systems.

Recent advances in AI constitute a strategic opportunity to meet the hydrological challenges of the 21st century (Barros et al., 2003; Hamitouche and Molina, 2022; Syed et al., 2024). However, the consolidation of this line requires improving reproducibility, incorporating uncertainty analysis, integrating anthropogenic processes and moving toward hybrid physical-informed models (Nourani et al., 2011; Ibrahim et al., 2022).

Author contributions

MT: Conceptualization, Data curation, Project administration, Writing – original draft. SC: Conceptualization, Supervision, Investigation, Visualization, Methodology, Writing – review & editing. HL: Conceptualization, Data curation, Methodology,

Writing – original draft. FA: Conceptualization, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. LM: Conceptualization, Data curation, Formal analysis, Validation, Writing – review & editing. JC: Conceptualization, Formal analysis, Validation, Writing – original draft. AM: Conceptualization, Software, Validation, Writing – original draft.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

References

- Addor, N., and Melsen, L. (2019). Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resour. Res.* 55, 378–390. doi: 10.1029/2018wr022958
- Almikaheel, W., Cubanová, L., and Šoltész, A. (2022). Hydrological drought forecasting using machine learning-Gidra River case study. *Water* 14:387. doi: 10.3390/w14030387
- Ardabili, S., Mosavi, A., Dehghani, M., and Várkonyi-Kóczy, A. R. (2020). “Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review,” in *Lecture Notes in Networks and Systems* (Cham: Springer), 52–62.
- Baran-Gurgul, K., and Rutkowska, A. (2024). Water resource management: hydrological modelling, hydrological cycles, and hydrological prediction. *Water* 16:3689. doi: 10.3390/w16243689
- Barros, M. T. L., Tsai, F. T.-C., Yang, S., Lopes, J. E. G., and Yeh, W. W.-G. (2003). Optimization of large-scale hydropower system operations. *J. Water Resour. Plan. Manag.* 129, 178–188. doi: 10.1061/(ASCE)0733-9496(2003)129:3(178)
- Bellin, N., Tesi, G., Marchesani, N., and Rossi, V. (2022). Species distribution modeling and machine learning in assessing the potential distribution of freshwater zooplankton in Northern Italy. *Ecol. Inf.* 69:101682. doi: 10.1016/j.ecoinf.2022.101682
- Bhadauria, A., Reddy, M. S. S., Asha, V., Nijhawan, G., Abdulhussein Hameed, A., and Pratap, B. (2024). “Analytical survey on the sustainable advancements in water and hydrology resources with AI implications for a resilient future,” in *E3S Web of Conferences* (Les Ulis: EDP Sciences).
- Chang, F. J., Chang, L. C., and Chen, J. F. (2023). Artificial intelligence techniques in hydrology and water resources management. *Water* 15:1846. doi: 10.3390/w15101846
- Chang, F. J., and Guo, S. (2020). Advances in hydrologic forecasts and water resources management. *Water* 12:1819. doi: 10.3390/w12061819
- Chen, C., Hui, Q., Pei, Q., Zhou, Y., Wang, B., Lv, N., et al. (2019). CRML: a convolution regression model with machine learning for hydrology forecasting. *IEEE Access* 7, 133839–133849. doi: 10.1109/ACCESS.2019.2941234
- Dasari, S. K., Preetha, P., and Ghantasala, H. M. (2025). Predictive analysis of hydrological variables in the Cahaba watershed: enhancing forecasting accuracy for water resource management using time-series and machine learning models. *Earth* 6:89. doi: 10.3390/earth6030089
- Drogkoula, M., Kokkinos, K., and Samaras, N. (2023). A comprehensive survey of machine learning methodologies with emphasis in water resources management. *Appl. Sci.* 13:12147. doi: 10.3390/app132212147
- Faybishenko, B., Ramakrishnan, L., Powell, T., Arora, B., Wu, J., and Agarwall, D. (2021). *On AI Prediction of Hydrological Processes Based on Integration of Retrospective and Forecasting ML Techniques*. Oak Ridge, TN: Oak Ridge National Laboratory.
- Ghobadi, F., and Kang, D. (2023). Application of machine learning in water resources management: a systematic literature review. *Water* 15:620. doi: 10.3390/w15040620
- Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *J. Hydrol.* 377, 80–91. doi: 10.1016/j.jhydrol.2009.08.003
- Hamitouche, M., and Molina, J. L. (2022). A review of AI methods for the prediction of high-flow extremal hydrology. *Water Resour. Manag.* 36, 3859–3876. doi: 10.1007/s11269-022-03240-y
- Hasan, F., Medley, P., Drake, J., and Chen, G. (2024). Advancing hydrology through machine learning: insights, challenges, and future directions using the CAMELS, caravan, GRDC, CHIRPS, PERSIANN, NLDAS, GLDAS, and GRACE datasets. *Water* 16:1904. doi: 10.3390/w16131904
- Ibrahim, K. S. M. H., Huang, Y. F., Ahmed, A. N., Koo, C. H., and El-Shafie, A. (2022). A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting. *Alex. Eng. J.* 61, 279–303. doi: 10.1016/j.aej.2021.04.100
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* 23, 5089–5110. doi: 10.5194/hess-23-5089-2019
- Kumar, V., Kedam, N., Sharma, K. V., Mehta, D. J., and Caloiero, T. (2023). Advanced machine learning techniques to improve hydrological prediction: a comparative analysis of streamflow prediction models. *Water* 15:2572. doi: 10.3390/w15142572
- Nash, J. E., and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models: Part 1. *A discussion of principles. J. Hydrol.* 10, 282–290. doi: 10.1016/0022-1694(70)90255-6
- Nearing, G. S., Clark, M. P., and Wood, A. W. (2020). The role of machine learning in hydrology: a paradigm shift? *Water Resour. Res.* 56:e2020WR027624. doi: 10.1029/2020WR028091
- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resour. Res.* 57:e2020WR028091. doi: 10.1029/2020WR028091

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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/frwa.2026.1756052/full#supplementary-material>

- Nourani, V., Kisi, Ö., and Komasi, M. (2011). Two hybrid artificial intelligence approaches for modeling rainfall-runoff process. *J. Hydrol.* 402, 41–59. doi: 10.1016/j.jhydrol.2011.03.002
- Noymanee, J., and Theeramunkong, T. (2019). “Flood forecasting with machine learning technique on hydrological modeling,” in *Procedia Computer Science* (Amsterdam: Elsevier B.V.), 377–386.
- Pathak, S., and Pandey, M. (2021). “Smart cities: review of characteristics, composition, challenges and technologies,” in *2021 6th International Conference on Inventive Computation Technologies (ICICT)* (Coimbatore), 871–876. doi: 10.1109/ICICT50816.2021.9358708
- Rahman, A. (2019). Statistics-based data preprocessing methods and machine learning algorithms for big data analysis. *Int. J. Artif. Intell.* 17, 44–65. Available online at: https://www.aut.upt.ro/~rprecup/IJAI_59.pdf; <https://www.scopus.com/pages/publications/85073352547?inward>
- Rozos, E., Dimitriadis, P., and Bellos, V. (2022). Machine learning in assessing the performance of hydrological models. *Hydrology* 9:5. doi: 10.3390/hydrology9010005
- Sarkis-Onofre, R., Catalá-López, F., Aromataris, E., and Lockwood, C. (2021). How to properly use the PRISMA statement. *Syst. Rev.* 10:117. doi: 10.1186/s13643-021-01671-z
- Sharma, A., Mehrotra, R., and Bari, M. (2020). Machine learning in hydrology: opportunities and challenges. *J. Hydrol.* 587:124945. doi: 10.1016/j.jhydrol.2020.124945
- Slater, L., Blougouras, G., Deng, L., Deng, Q., Ford, E., Hoek, A., et al. (2025). Challenges and opportunities of ML and explainable AI in large-sample hydrology. *Philos. Trans. R. Soc. A* 383:20240287. doi: 10.1098/rsta.2024.0287
- Solanki, H., Vegad, U., Kushwaha, A., and Mishra, V. (2025). Improving streamflow prediction using multiple hydrological models and machine learning methods. *Water Resour. Res.* 61:e2024WR038192. doi: 10.1029/2024WR038192
- Syed, T. A., Khan, M. Y., Jan, S., Albouq, S., Alqahtany, S. S., and Naqash, M. T. (2024). Integrating digital twins and artificial intelligence multi-modal transformers into water resource management. *AI* 5, 1977–2017. doi: 10.3390/ai5040098
- Wagner, T., Sivapalan, M., Troch, P. A., McGlynn, B. L., Harman, C. J., Gupta, H. V., et al. (2010). The future of hydrology: an evolving science for a changing world. *Water Resour. Res.* 46. doi: 10.1029/2009WR008906
- Wang, X., Li, Y., Qiao, Q., Tavares, A., and Liang, Y. (2023). Water quality prediction based on machine learning and comprehensive weighting methods. *Entropy* 25:1186. doi: 10.3390/e25081186
- Willard, J. D., Varadharajan, C., Jia, X., and Kumar, V. (2024). Time series predictions in unmonitored sites: a survey of machine learning techniques in water resources. *arXiv*. 2308.09766 doi: 10.1017/eds.2024.14
- Xu, T., and Liang, F. (2021). Machine learning for hydrologic sciences: an introductory overview. *Wiley Interdiscip. Rev. Water* 8:e1533. doi: 10.1002/wat2.1533
- Yaseen, Z. M., Allawi, M. F., Yousif, A. A., Othman, J., Hamzah, F. M., and Ahmed, E. S. (2018). Non-tuned machine learning approach for hydrological time series forecasting. *Neural Comput. Appl.* 30, 1479–1491. doi: 10.1007/s00521-016-2763-0
- Zhang, X., Liu, Y., and Wang, Z. (2021). Machine learning in hydrology: a review. *J. Hydrol.* 598:126364. doi: 10.1016/j.jhydrol.2021.126266
- Zhong, Z., Sun, A., Wang, Y., and Ren, B. (2020). Predicting field production rates for waterflooding using a machine learning-based proxy model. *J. Pet. Sci. Eng.* 194:107574. doi: 10.1016/j.petrol.2020.107574