ARTIFICIAL INTELLIGENCE-BASED FORECASTING AND ANALYTIC TECHNIQUES FOR ENVIRONMENT AND ECONOMICS MANAGEMENT

EDITED BY: Wendong Yang, Shaolong Sun, Yiliao Song, Pei Du, Yan Hao and Jianzhou Wang

PUBLISHED IN: Frontiers in Environmental Science, Frontiers in Earth Science, Frontiers in Ecology and Evolution and Frontiers in Marine Science



frontiers Research Topics



frontiers

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 978-2-83250-472-7
DOI 10.3389/978-2-83250-472-7

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

ARTIFICIAL INTELLIGENCE-BASED FORECASTING AND ANALYTIC TECHNIQUES FOR ENVIRONMENT AND ECONOMICS MANAGEMENT

Topic Editors:
Wendong Yang, Shandong University of Finance and Economics, China
Shaolong Sun, Xi'an Jiaotong University, China
Yiliao Song, University of Technology Sydney, Australia
Pei Du, Xi'an Jiaotong University, China
Yan Hao, Shandong Normal University, China
Jianzhou Wang, Dongbei University of Finance and Economics, China

Citation: Yang, W., Sun, S., Song, Y., Du, P., Hao, Y., Wang, J., eds. (2022). Artificial Intelligence-Based Forecasting and Analytic Techniques for Environment and Economics Management. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83250-472-7

Table of Contents

05 ***An Ensemble Prediction System Based on Artificial Neural Networks and Deep Learning Methods for Deterministic and Probabilistic Carbon Price Forecasting***
Yi Yang, Honggang Guo, Yu Jin and Aiyi Song

27 ***An Outlier-Robust Point and Interval Forecasting System for Daily $PM_{2.5}$ Concentration***
Ziqi Yin and Xin Fang

42 ***Comprehensive Evaluation of Water Carrying Capacity in Hebei Province, China on Principal Component Analysis***
Huadong Wang, Yunhong Xu, Rini Suryati Sulong, HuiLing Ma and Lifeng Wu

51 ***Research on Influence Factors of Air Quality in the Host Area of the 24th Winter Olympic Games Based on Grey Relational Analysis***
Wei Wang, Yan Chen, Yuhan Xie and Lifeng Wu

61 ***Hybrid Data Mining Forecasting System Based on Multi-Objective Optimization and Selection Model for Air pollutants***
Yanwen Huang, Yuanchang Deng, Chen Wang and Tonglin Fu

84 ***Modeling of Direct Economic Losses of Storm Surge Disasters Based on a Novel Hybrid Forecasting System***
Hongbo Guo, Kedong Yin and Chong Huang

100 ***Design of a Combined System Based on Multi-Objective Optimization for Fine Particulate Matter ($PM_{2.5}$) Prediction***
Lu Bai, Hongmin Li, Bo Zeng and Xiaojia Huang

118 ***A Quantitative Assessment on Ecological Compensation Based on Water Resources Value Accounting: A Case Study of Water Source Area of the Middle Route of South-To-North Water Transfer Project in China***
Junfei Chen, Qian Wang and Qian Li

132 ***Multivariate Adaptive Step Fruit Fly Optimization Algorithm Optimized Generalized Regression Neural Network for Short-Term Power Load Forecasting***
Feng Jiang, Wenya Zhang and Zijun Peng

145 ***Forecasting Energy Consumption Based on SVR and Markov Model: A Case Study of China***
Zhaosu Meng, Huike Sun and Xi Wang

160 ***Prediction of Air Pollution Interval Based on Data Preprocessing and Multi-Objective Dragonfly Optimization Algorithm***
Jiyang Wang, Jingrui Li and Zhiwu Li

178 ***Forecasting New COVID-19 Cases and Deaths Based on an Intelligent Point and Interval System Coupled With Environmental Variables***
Zongxi Qu, Yongzhong Sha, Qian Xu and Yutong Li

191 ***Tourism Demand Forecasting Considering Environmental Factors: A Case Study for Chengdu Research Base of Giant Panda Breeding***
Jianhong He, Dong Liu, Yulei Guo and Daohua Zhou

202 ***COVID-19 Pandemic, Air Quality, and $PM_{2.5}$ Reduction-Induced Health Benefits: A Comparative Study for Three Significant Periods in Beijing***
Fangfang Cai, Kedong Yin and Mengying Hao

229 ***Analysis of Ambient Air PM10-Bound Pollutants Surrounding an Industrial Site and Their Prediction Using Artificial Neural Network***
M.G. Bonelli, M. Cerasa, E. Guerriero, A. Manni, S. Mosca, M. Perilli and G. Rossetti

236 ***Research and Application of the Mode Decomposition-Recombination Technique Based on Sample-Fuzzy Entropy and K-Means for Air Pollution Forecasting***
Chunhua Niu, Ziqi Niu, Zongxi Qu, Lili Wei and Yutong Li

247 ***Utilizing a Two-Dimensional Data-Driven Convolutional Neural Network for Long-Term Prediction of Dissolved Oxygen Content***
Dashe Li and Xuan Zhang

267 ***CGAN Based Improved ASI Retrieval Algorithm for Antarctic Sea Ice Concentration***
Xingdong Wang, Yue Zhao, Shuhui Yang, Yuhua Wang and Donghui Shangguan

276 ***Study on Spatial Evolution and Mechanism of County Economic Differences in Hubei Province From 2005 to 2020***
Liang Jiang, Jie Chen, Chunyan Zhang, Ye Tian, Guoxi Wu and Jing Luo

296 ***A Novel Hybridized Neuro-Fuzzy Model With an Optimal Input Combination for Dissolved Oxygen Estimation***
Saman Maroufpoor, Saad Sh. Sammen, Nadhir Alansari, S.I. Abba, Anurag Malik, Shamsuddin Shahid, Ali Mokhtar and Eisa Maroufpoor

313 ***Evidence of the Middle-Income Trap in Latin American Countries: Factor Analysis Approach Using Regression and the ARDL Model***
Lisette Paola Cervantes Galvan, Carlos Carmona Campo, Savo Stanojevic and David Villada Alzate

frontiers in Environmental Science

ORIGINAL RESEARCH
published: 17 September 2021
doi: 10.3389/fenvs.2021.740093

An Ensemble Prediction System Based on Artificial Neural Networks and Deep Learning Methods for Deterministic and Probabilistic Carbon Price Forecasting

Yi Yang[1], *Honggang Guo*[2]*, *Yu Jin*[2] *and Aiyi Song*[2]

[1]School of Information Science and Engineering, Lanzhou University, Lanzhou, China, [2]School of Statistics, Dongbei University of Finance and Economics, Dalian, China

Carbon price prediction is important for decreasing greenhouse gas emissions and coping with climate change. At present, a variety of models are widely used to predict irregular, nonlinear, and nonstationary carbon price series. However, these models ignore the importance of feature extraction and the inherent defects of using a single model; thus, accurate and stable prediction of carbon prices by relevant industry practitioners and the government is still a huge challenge. This research proposes an ensemble prediction system (EPS) that includes improved data feature extraction technology, three prediction submodels (GBiLSTM, CNN, and ELM), and a multiobjective optimization algorithm weighting strategy. At the same time, based on the best fitting distribution of the prediction error of the EPS, the carbon price prediction interval is constructed as a way to explore its uncertainty. More specifically, EPS integrates the advantages of various submodels and provides more accurate point prediction results; the distribution function based on point prediction error is used to establish the prediction interval of carbon prices and to mine and analyze the volatility characteristics of carbon prices. Numerical simulation of the historical data available for three carbon price markets is also conducted. The experimental results show that the ensemble prediction system can provide more effective and stable carbon price forecasting information and that it can provide valuable suggestions that enterprise managers and governments can use to improve the carbon price market.

Keywords: carbon price forecasting, ensemble prediction system, deep learning methods, error distribution function, multiobjective optimization algorithm

OPEN ACCESS

Edited by:
Wendong Yang,
Shandong University of Finance and Economics, China

Reviewed by:
Yao Dong,
Jiangxi University of Finance and Economics, China
Jamshid Piri,
Zabol University, Iran

***Correspondence:**
Honggang Guo
ghg970612@163.com

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: *12 July 2021*
Accepted: *23 August 2021*
Published: *17 September 2021*

Citation:
Yang Y, Guo H, Jin Y and Song A (2021) An Ensemble Prediction System Based on Artificial Neural Networks and Deep Learning Methods for Deterministic and Probabilistic Carbon Price Forecasting. Front. Environ. Sci. 9:740093. doi: 10.3389/fenvs.2021.740093

INTRODUCTION

This section describes the research background, provides a literature review, and states the purpose and innovation of this study.

Research Background

With the rapid development of the economy, the environment and the climate will inevitably change. Climate change is clearly a common problem facing all countries. On February 16, 2005, the Kyoto

Protocol went into effect. According to the situation in each country, specific emission reduction plans and schedules were formulated. In January 2005, the EU emissions trading scheme (EU ETS), which was designed to achieve the emission reduction targets stipulated in the Kyoto protocol, was introduced (Arouri et al., 2012). The EU ETS allocates carbon trading quotas to different emission entities according to its regulations, and entities that exceed the quota must purchase emission rights from entities that are lower than the quota through the carbon trading market. This measure of using a market trading mechanism provides valuable experience for solving the problem of global climate change.

As the world's largest carbon dioxide emitter (in 2018, its total carbon dioxide emissions reached 10 billion tons, accounting for approximately 30% of global carbon dioxide emissions), China has successively established eight carbon emission trading markets since 2013. However, this system is still in the construction stage, and the market mechanism is not perfect and needs further improvement. By studying the regular price fluctuation pattern of the EU ETS and China's carbon trading market, analyzing the influencing factors, and forecasting the carbon market price accordingly, we can better understand the fluctuation law of the carbon market and obtain a reference for formulating carbon market policies and mechanisms to improve the ability to regulate this market.

Carbon prices have important implications for governments, companies, and long-term investors. For governments, carbon pricing is one of the mechanisms used to reduce carbon emissions, and it can also be a source of revenue. Companies can use internal carbon pricing to assess the impact of mandatory carbon pricing on their businesses and to identify potential climate risks and revenue opportunities. Long-term investors are using carbon pricing to reevaluate their investment strategies. Therefore, regardless of the point of view, it is necessary to establish an accurate and stable carbon price forecasting system.

Literature Review

Most of the research methods used in carbon price prediction rely on historical data to build models to predict carbon prices. Carbon prices display high volatility and nonlinear structure, and many studies of carbon price prediction based on historical data have been conducted in recent years. The prediction methods can be divided into three categories: 1) statistical measurement methods; 2) artificial intelligence methods; and 3) decomposition integration hybrid forecasting methods.

Statistical Measurement Method

As a classical time series forecasting method, statistical measurement methods, including linear regression models, autoregressive integrated moving averages (ARIMAs), generalized autoregressive conditional heteroscedasticity (GARCH) models, and gray model GM (1, 1) (Chevallier, 2009; Byun and Cho, 2013; Zhu and Wei, 2013) are widely used in carbon trading price prediction and volatility analysis. For example, Benz and Trück (2008) proposed the Markov transition and AR-GARCH model for stochastic modeling and analyzed the short-term price of the carbon dioxide emission quota of the EU ETS. Through the empirical results obtained, it was demonstrated that the prediction performance of the Markov state transition model is better than that of the GARCH model. Zhu and Wei (2013) combined least squares SVM with the ARIMA model, and the results showed that the developed model was more robust than the single-prediction model. Zhu B et al. (2018) used grey correlation analysis to analyze the carbon price market. The traditional statistical model has high prediction accuracy and wide adaptability in linear and stable time series. However, because carbon prices show strong volatility, nonlinearity, and instability, traditional statistical measurement methods cannot capture internal structural characteristic data (Lu et al., 2019). Therefore, accurate forecasting of carbon prices requires the use of a method with a strong nonlinear feature extraction ability that enables it to take into account potential nonlinear characteristics. In addition, the traditional statistical measurement method is more suitable for the long-term prediction of time series, and its short-term carbon price prediction performance is poor (Cheng and Wang, 2020).

Owing to the shortcomings of statistical models, ***artificial intelligence methods*** (**AI**) have gradually become widely used in time series prediction; these methods are suitable for nonlinear prediction without any assumption of data distribution (Wang et al., 2020). Increasing evidence shows that the performance of AI in nonlinear time series is better than that of other models (Zhang et al., 2017). AI, including back-propagation neural networks (BPs), multilayer perceptual neural networks (MLPs), least squares support vector regression (LSSVR), and hybrid prediction methods combined with optimization algorithms, have also been widely used in carbon price forecasting. Atsalakis (2016) combined a hybrid fuzzy controller called PATSOS with an adaptive neuro fuzzy inference system (ANFIS). The research shows that this method can produce accurate and timely prediction results. Fan et al. (2015) studied the chaotic characteristics of the EU ETS, used the neural network model of MLP to predict carbon prices, and found that the forecasting accuracy of the model was significantly improved. Tian and Hao (2020) used phase space reconstruction technology and the ELM under the multiobjective grasshopper optimization algorithm (MOGOA-ELM) to predict the trend of the EU ETS and China's carbon prices. The empirical results show that this method can be used effectively to predict carbon prices.

In recent years, with the development of deep learning theory (DL) in image detection, audio detection, and other fields, DL has become the focus of many scholars (Liu et al., 2021). The unique storage unit structure of deep learning allows it to retain past historical data and has significant advantages for processing time series data that feature long processing intervals and delays (Zhang B et al., 2018). Niu et al. (2020) combined LSTM and GRU to establish a deep learning recursive forecasting unit for forecasting multiple financial data. Liu et al. (2020) proposed a new wind speed prediction model based on an error correction strategy and the LSTM algorithm to predict short-term wind speed. The experimental results demonstrated that its performance is better than that of other comparable models.

However, application of deep learning frameworks to carbon price prediction is still very limited.

In addition to the selection and optimization of prediction methods, data preprocessing technology also plays an indispensable role in the prediction accuracy of the prediction model (Wang et al., 2021). ***Decomposition and integration methods***, including empirical mode decomposition (EMD), singular spectrum denoising (SSA), and variational mode decomposition (VMD) are widely used in time series data preprocessing. These methods aim to decompose and reconstruct the original time series data and extract the effective features of the time series. Decomposing the original time series into a series of simple patterns that exhibit strong regularity can significantly improve the prediction accuracy of time series. Wei et al. (2018) used wavelet transform and kernel ELM to predict carbon prices. Zhu J et al. (2018) explored an efficient prediction model based on VMD mode reconstruction and optimal combination and thereby greatly improved the prediction accuracy of carbon prices. However, the above decomposition methods still have some shortcomings. For example, in wavelet decomposition and VMD, it is necessary to determine the wavelet basis function and the decomposition level. Although in EMD it is not necessary to determine the number of decomposition levels, mode aliasing and insufficient noise separation cannot be solved (Jin et al., 2020). Therefore, it is very important to extract the nonlinear peculiarities of carbon prices by using appropriate data preprocessing methods.

A single prediction model cannot achieve good performance on every dataset. Therefore, researchers began to focus on combination forecasting models. In essence, combination forecasting models combine different hybrid forecasting methods or single forecasting methods using weighting. In many experimental studies, it is found that the use of a combination of prediction methods produces better predictions than the use of a method that is based on a single-prediction model. The advantage of using combination models is that different time series may have different information sets, information features, and modeling structures, and the use of a combination of prediction methods can result in good performance in the case of such structure mutations. Although the use of a combination forecasting method to forecast time series is very common, use of a combination forecasting model to forecast carbon prices is still in its infancy.

The above analysis indicates that most research on carbon prices is driven by single or hybrid forecasting models and that it tends to emphasize prediction strategies that are based on certainty and to largely ignore the importance of uncertainty analysis of carbon prices. Regardless of the type of prediction model used, there are inherent and irreducible uncertainties in each prediction that will greatly increase the possibility of miscalculation (Du et al., 2020). Therefore, quantification of the uncertainty of carbon price prediction plays an indispensable role in exploring the complexity of the carbon price market and strengthening the ability to conduct effective market anti-risk management.

Objectives and Contributions

To supplement the existing research on carbon price prediction, an ensemble prediction system (EPS) based on the ICEEMDAN data preprocessing method, the deep learning algorithm (DL), the extreme learning machine (ELM), and the multiobjective dragonfly optimization algorithm (MODA) is developed and used to analyze the certainty and uncertainty of carbon prices. Specifically, ICEEMDAN is employed to decompose and reconstruct the original carbon price data and extract the effective features of the data, and the results are transferred into the submodels of EPS as training data (the submodels are ICEEMDAN-GBiLSTM, ICEEMDAN-CNN, and ICEEMDSAN-ELM). Using the MODA, the final carbon price point forecast results are then obtained through a weighted combination of the submodel prediction results. For interval prediction, the upper and lower bounds of the prediction interval are constructed based on the prediction value of ESP and the best fit distribution function of error, namely, the T location-scale (TLS) distribution. The main innovations presented in this study are as follows:

1) ***An effective ensemble prediction system of carbon prices is developed.*** Two hybrid prediction models based on a deep learning algorithm (ICEEMDAN-GBiLSTM and ICEEMDAN-CNN) and a feedforward neural network (ICEEMDAN-ELM) are combined to ***overcome the inherent defects of a single hybrid prediction model.***
2) A deep learning recurrent neural network, GBiLSTM, is first proposed as a prediction submodel of the EPS. ***GBiLSTM combines two recursive deep learning algorithms; it can effectively deal with time series with long memory and increase the accuracy of carbon price forecasting.***
3) The MODA is employed as an effective method of weighting the ensemble prediction system. It optimizes the weight coefficient of ***the ensemble model from the perspective of prediction accuracy and prediction stability,*** thereby overcoming the obvious defect that single objective optimization can only select one objective function.
4) ***To overcome the nonlinearity and strong volatility of the original carbon price data, an effective time series preprocessing technique is developed.*** ICEEMDAN sequence decomposition technology is employed to decompose and reconstruct the original carbon price data, extract the salient features of the data, and improve the prediction accuracy of the EPS.
5) ***By fitting the best error distribution, the uncertainty of the carbon price is mined.*** In the past, the error distribution of a prediction was usually assumed to be a Gaussian distribution. In this study, ***five types of parameter distribution functions are used to fit the prediction error***, the best error distribution function is found, and the ranges of carbon price interval prediction are constructed.

The remainder of the study is organized as follows. In Section *Model Theory and Related Work* and Section *Ensemble Prediction System and its Interval Forecasting Framework*, we introduce the theoretical method and the framework used in the proposed EPS.

Section *Experiment and Analysis* describes the experimental data and the prediction performance evaluation index. The point prediction and interval prediction of the carbon price are then simulated. Section *Discussion* is a further discussion of EPS, and a summary of the study is presented in Section *Conclusion*.

MODEL THEORY AND RELATED WORK

This section introduces the corresponding theories and describes the functions of the data preprocessing module, the combination prediction module, and the uncertainty mining module of the EPS prediction system.

Data Preprocessing

The data processing module includes the data feature extraction method, which is based on improved complex ensemble empirical mode decomposition with adaptive noise (ICEEMDAN), and the data feature selection method, which is based on the partial autocorrelation function (PACF).

Data Feature Extraction

To improve the problem of mode aliasing in the traditional noise reduction method EMD and the slight residual noise in CEEMDAN, the ICEEMDAN technology is improved. The CEEMDAN method adds Gaussian white noise during the decomposition process, while the ICEEMDAN method adds a special type of white noise, $E_k(\omega^{(i)})$, that is, the k-th IMF component of the Gaussian white noise (M.E. Torres et al., 2011; M.A. Colominas et al., 2014). The local mean value of the added noise is calculated for each modal component, and the IMF is defined as the difference between the residual signal and the local mean.

1) The definition operator $E_k(\cdot)$ represents the k-th IMF after EMD decomposition, and $M(\cdot)$ represents the local mean value of the signal. There is $\boldsymbol{E}_1(x) = x - \boldsymbol{M}(x)$. Operator means taking the mean value, and x represents the original data of the study, and then the local average value is calculated by EMD:

$$\boldsymbol{x}^{(i)} = \boldsymbol{x} + \alpha_0 \boldsymbol{E}_1\left(\omega^{(i)}\right) \tag{1}$$

where $\omega^{(i)}$ is the ith white noise added and α_0 is the standard deviation of the noise. The first residual component $\mathrm{r}_1 = \langle M(x^i)\rangle$ is obtained by taking the local mean value. The first intrinsic mode function IMF_1 value $\widetilde{\boldsymbol{d}}_1 = \mathrm{x} - r_1$ is calculated.

2) The value $\widetilde{d}_2$ of the second mode component IMF_2 is calculated:

$$\widetilde{\boldsymbol{d}}_2 = r_1 - r_2 = r_1 - \langle \boldsymbol{M}(r_1 + \alpha_1 \boldsymbol{E}_2(\omega^{(i)}))\rangle \tag{2}$$

3) The k-th residual is calculated:

$$\boldsymbol{r_k} = \langle \boldsymbol{M}\left(r_{k-1} + \alpha_{k-1}\boldsymbol{E_k}\left(\omega^{(i)}\right)\right)\rangle, k = 2, 3, \cdots, N \tag{3}$$

4) The value of the k-th mode component $IMF_k.\widetilde{\boldsymbol{d}}_{\mathbf{k}} = r_{k-1} - r_k$, is calculated, and **Eq. 3** is repeated until the residual satisfies the iteration termination condition, which is ***Cauchy*** convergence. The standard deviation between two adjacent ***IMF*** components $\chi = \left\|\widetilde{\mathrm{d}}_k - \widetilde{\mathrm{d}}_{k-1}\right\|_2 / \left\|\widetilde{\mathrm{d}}_{k-1}\right\|_2$ is less than a specified value.

In this study, ICEEMDAN is used to decompose the original carbon price data into several intrinsic mode functions (IMFs). The IMF with the highest frequency is removed, and the remaining IMFs are included. Through this method of deconstruction and reorganization, the problem of strong volatility and randomness of the original data is solved. The data features are effectively extracted, and the prediction veracity of the model is increased.

Data Feature Selection

The partial autocorrelation function (PACF) is an effective method for distinguishing the structural features of sequences (Jiang et al., 2020). It can be used to calculate the partial correlation between the time series and its lag term. If Φ_{kj} is employed to represent the j-th regression coefficient in the k-order autoregressive equation, the model can be expressed as follows:

$$\boldsymbol{x}_{\mathrm{t}} = \boldsymbol{\Phi}_{k1}x_{t-1} + \boldsymbol{\Phi}_{k2}x_{t-2} + \cdots + \boldsymbol{\Phi}_{kk}x_{t-k} + \mu_t \tag{4}$$

where x_t is the time series and Φ_{kk} is the last coefficient. If Φ_{kk} is defined as a function of lag time k, then Φ_{kk}, $k = 1,2...$ is named partial autocorrelation function.

In this study, PACF is used to find the lag terms that have the strongest correlation with the time series; these are then used as the input characteristics of the forecast model.

Ensemble Prediction Module

The prediction value calculated by the ensemble prediction system is obtained using an ensemble of the prediction results of different single-prediction components through the weighting strategy. In this section, the three submodes of the proposed EPS and the MODA weighting optimization strategy are introduced.

Convolutional Neural Network

A CNN is an incompletely connected DL network structure that is composed of two special neural networks: a convolution layer and a down sampling layer (Wang, 2020). The neurons in each layer of the CNN are locally connected, enabling them to realize hierarchical feature extraction and transformation of the input. Neurons with the same connection weight are connected to different regions of the upper neural network; in this way, a translation-invariant neural network is obtained (Wang, 2018).

1) ***The training of the Convolution Layer.*** The CNN is connected to the local region of the feature surface by a convolution kernel. The output characteristic surface size of each convolution layer must meet the following requirement:

$$\boldsymbol{oMapN} = \left(\frac{(\boldsymbol{iMap}n - C\boldsymbol{Window})}{C\boldsymbol{Interval}} + 1\right) \tag{5}$$

In **Eq. 5**, ***oMapN*** is the number of output feature surfaces of each convolution layer, ***iMapN*** is the number of input feature surfaces, ***CWindow*** is the size of the convolution kernel, and ***CInterval*** is the sliding step size of the convolution kernel.

In general, to ensure integral division in the above formula, it is necessary to train the number of parameters for each convolution layer of the CNN so as to satisfy the following condition:

$$C\boldsymbol{Params} = \left(i\boldsymbol{Map} \times C\boldsymbol{Window} + 1\right) \times o\boldsymbol{Map} \tag{6}$$

where ***CParams*** is the number of parameters, ***iMap*** is the input feature surface, and ***oMap*** is the output feature surface.

The output value x_{nk}^{out} can be obtained by the convolution layer; the formula is as follows:

$$\boldsymbol{x}_{\boldsymbol{nk}}^{\boldsymbol{out}} = \boldsymbol{f}_{\boldsymbol{cov}}\left(\boldsymbol{x}_{1\boldsymbol{h}}^{\boldsymbol{in}} \times \boldsymbol{w}_{1(h)n(k)} + \boldsymbol{x}_{1(\boldsymbol{h}+1)}^{\boldsymbol{in}} \times \boldsymbol{w}_{1(h+1)n(k)} + \boldsymbol{x}_{1(\mathrm{h}+2)}^{\boldsymbol{in}} \times \boldsymbol{w}_{1(h+2)n(k)} + \cdots + \boldsymbol{b}_{\boldsymbol{n}}\right) \tag{7}$$

where x_{mk}^{in} is the input value, b_n is the offset value of the output characteristic surface n, and $f_{\mathrm{cov}}(\cdot)$ is the excitation function. The excitation function is usually the ReLU function, and the formula for its calculation is as follows:

$$\boldsymbol{f}_{\boldsymbol{cov}}(x) = \boldsymbol{MAX}(0, x) \tag{8}$$

2) ***The output of the Pooling Layer.*** The pooling layer is also composed of several feature surfaces, and the number of feature surfaces does not change. The output value of the pooling layer is

$$\boldsymbol{t}_{\boldsymbol{nl}}^{\boldsymbol{out}} = \boldsymbol{f}_{\boldsymbol{sub}}\left(t_{nq}^{in}, t_{n(q+1)}^{in}\right) \tag{9}$$

where t_{nq}^{in} is the output value of the q-th neuron on the n-th input characteristic surface of the pooling layer and $f_{sub}(\cdot)$ is a function that takes either the maximum value or the mean value. The size ***DoMapN*** of each output feature surface of the pooling layer is

$$\boldsymbol{DoPapN} = \left(\frac{o\boldsymbol{Map}N}{D\boldsymbol{Window}}\right) \tag{10}$$

3) Full connection layer output. In the CNN structure, one or more fully connected layers are connected after the multiple convolution layers and the pooling layers are obtained. The ReLU function is also used in the excitation function of the whole connected layer.

In this study, CNN, as a component of the combined forecasting system, forecasts the carbon price.

Deep Learning Recursive Network Structure (GBiLSTM)

In this study, we developed a deep learning recurrent network structure, which is a hybrid of BiLSTM and GRU. The structure diagram is shown in **Figure 1**.

Bidirectional Long Short Term Memory Neural Network

BiLSTM is an improved network of LSTM. LSTM cannot capture information from back to front; however, BiLSTM can solve this problem. When bidirectional sequence information is captured, the time series can be predicted more accurately (Hochreiter and Schmidhuber, 1997).

1) The LSTM mechanism consists of three memory gates: an input gate (i_t), a forgetting gate (f_t), and an output gate (O_t). The specific expression is as follows:

$$\begin{cases} \boldsymbol{i}_t = \sigma\left(\omega_i \boldsymbol{X}_t + R_i \boldsymbol{H}_{t-1} + \boldsymbol{B}_i\right) \\ \boldsymbol{f}_t = \sigma\left(\omega_f \boldsymbol{X}_t + R_f \boldsymbol{H}_{t-1} + \boldsymbol{B}_f\right) \\ \boldsymbol{o}_t = \sigma\left(\omega_o \boldsymbol{X}_t + R_o \boldsymbol{H}_{t-1} + \boldsymbol{B}_{\mathrm{o}}\right) \\ \widetilde{\boldsymbol{c}} = \boldsymbol{f}_t \otimes c_{t-1} \\ \boldsymbol{H}_t = o_t \otimes \boldsymbol{tanh}(c_t) \end{cases} \tag{11}$$

where x_t, σ, and c_t represent the input sample, the ***sigmoid*** activation function, and the storage unit of time t, respectively, and (B_f, B_i, B_o) and (ω_f, ω_i, ω_o) represent the deviation and the weight matrix, respectively, of each gate. The symbol ⊗ represents the corresponding multiplication of elements. First, $\boldsymbol{H}_{\mathrm{t}-1}$, c_{t-1}, and $\boldsymbol{X}_{\mathrm{t}}$ transmit input information to the LSTM unit. The LSTM gate then interacts with$\boldsymbol{X}_{\mathrm{t}}$. After a new cell state c_{t-1} is established. In this stage, $\boldsymbol{f}_t$ determines which information needs to be stored or deleted and then updates the cell status.

2) Because BiLSTM transmits time series data to LSTM from both the forward and backward directions, it has two output layers: a forward layer $H_t^f = o_t^f \otimes \boldsymbol{tanh}(c_t^f)$ and a backward layer $H_t^b = o_t^b \otimes \boldsymbol{tanh}(c_t^b)$.
3) The final predicted output value $\widehat{y}_t$ is obtained by integrating the forward layer and the backward layer; in **Eq. 12**, $\alpha$and $\beta$are numerical factors that satisfy the equation $\alpha + \beta = 1$ (Shi et al., 2015).

$$\begin{cases} \boldsymbol{H}_{\mathrm{t}} = \alpha \boldsymbol{H}_{\mathrm{t}}^{\mathrm{f}} + \beta \boldsymbol{H}_{\mathrm{t}}^{\mathrm{b}} \\ \widehat{y}_t = \sigma\left(\boldsymbol{H}_t\right) \end{cases} \tag{12}$$

Gated Recurrent Unit

GRU is an effective variant of LSTM; its structure is simpler than that of LSTM, and it can well capture the nonlinear relationship between sequence data, thereby effectively alleviating the problems of traditional RNN gradient disappearance (Chung et al., 2014).

1) The GRU model has two gating units: an update gate $\mathrm{z_t}$ and a reset gate $\mathrm{r_t}$. An update gate is employed to equilibrate the historical information. The smaller the value of the update gate is, the more concentrated the output of the model on the information of the previous hidden layer $\boldsymbol{h}_{\mathrm{t}}$ is

$$\boldsymbol{Z}_t = \sigma\left(\boldsymbol{W}_z.\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]\right) \tag{13}$$

$$\boldsymbol{r}_t = \sigma\left(\boldsymbol{W}_r.\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]\right) \tag{14}$$

In **Eqs 13**, **14**, Wis the model weight, and σ is the activation function.

2) By resetting the gate r_t, the candidate vector $\widetilde{\boldsymbol{h}}_t = tanh\left(\boldsymbol{W}.\left[\boldsymbol{r}_t * \boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]\right)$ can be calculated. Taking the

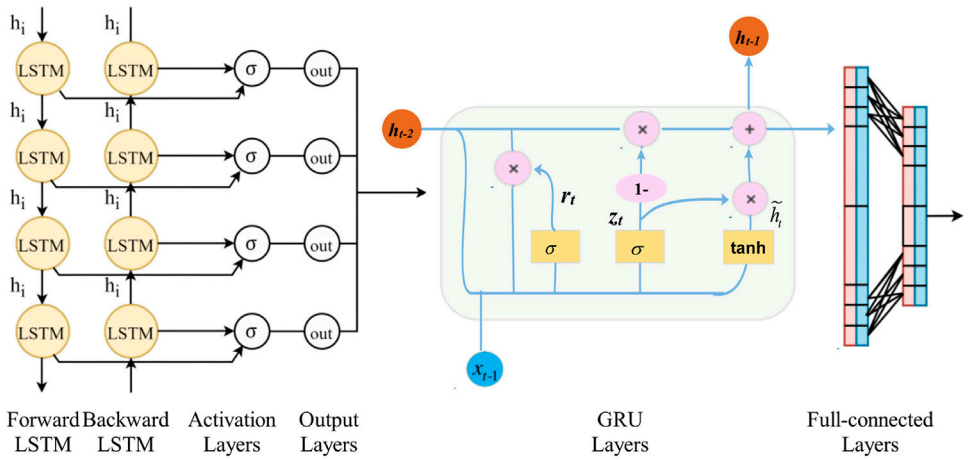


FIGURE 1 | Flowchart of the proposed bidirectional long short-term memory-gated recurrent unit (GBiLSTM) model.

value of the update gate as the weight, $\tilde{h}_t$ to be added and the state of the hidden layer at the previous time step are recorded as the output of the GRU network at time step t, as follows:

$$\boldsymbol{h}_t = (1 - \boldsymbol{Z}_t) * \boldsymbol{h}_{t-1} + \boldsymbol{Z}_t * \tilde{\boldsymbol{h}}_t \tag{15}$$

4) A set of training samples is input into the GRU; finally, the final output o is obtained by adding the fully connected layer after the GRU layer.

$$\begin{cases} \boldsymbol{h}_{end} = \boldsymbol{GRU}(x_{t-1}, x_{x-2}, \ldots, x_{t-w}) \\ o = \boldsymbol{V}.\boldsymbol{h}_{end} \end{cases} \tag{17}$$

In this study, a deep learning recursive network structure (GBiLSTM) based on BiLSTM and GRU is constructed. To reduce fitting error, the time series are trained by the BiLSTM layer and then transferred to the GRU network. Through this double deep learning layer network structure, we can better fit the carbon price data and reduce the prediction error.

Extreme Learning Machine

ELM is a type of feedforward neural network. On the premise of randomly selecting the input layer weight and the hidden layer neuron threshold, the output weight of the ELM can be obtained through a one-step calculation. ELM has the advantages of higher network generalization ability and strong nonlinear fitting ability (G B Huang et al., 2006; Jiang et al., 2021a; Jiang et al., 2021b).

1) For N different inputs (x_i, t_i) and $x_i \in R^p, t_i \in R^p, i = 1, 2, \ldots, N$, the ELM with L nodes and the excitation function $f(x)$ can be expressed as

$$\sum_{i=1}^{N} \beta_i f\left(\boldsymbol{w}_i.\boldsymbol{x}_j + \boldsymbol{b}_i\right) = y_j \qquad i = 1, \ldots, N \tag{19}$$

where $\boldsymbol{w}_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$ is the weight connecting the i-th hidden layer node and the input node, β_i is the connection weight vector, and y_j is the output value of the j-th node. The training of the network is equivalent to approximating N training samples with zero error; that is, $\hat{w}_i, \hat{b}, \hat{\beta}$make

$$\sum_{i=1}^{L} \hat{\beta}_i f\left(\hat{\boldsymbol{w}}_i.x_j + \hat{b}_i\right) = t_j, j = 1, \ldots, N \tag{20}$$

$$\boldsymbol{H}\beta = \boldsymbol{T}; \tag{21}$$

$$\boldsymbol{H} = \begin{bmatrix} f(\boldsymbol{w}_1 x_1 + \boldsymbol{b}_1) & \ldots & f(\boldsymbol{w}_L x_1 + \boldsymbol{b}_L) \\ \vdots & & \vdots \\ f(\boldsymbol{w}_1 x_N + \boldsymbol{b}_1) & \ldots & f(\boldsymbol{w}_L x_N + \boldsymbol{b}_L) \end{bmatrix}_{N \times L} \tag{22}$$

In **Eq. 20** through **Eq. 22**, $\beta = \left[\beta_1^T, \ldots, \beta_L^T\right]_{L \times m}; T = \left[t_1^T, \ldots, t_N^T\right]_{N \times m}$, and the i-th column of H represents the output vector of the i-th hidden layer node corresponding to the i-th hidden layer neuron of the input $x_1, x_2, \ldots, x_N$.

2) The input connection weight W and the hidden layer node bias b can be randomly selected at the beginning of training, and the output connection weight $\hat{\beta}$ can be solved by solving the linear **Eq. 23**.

$$\min_\beta \left\| \boldsymbol{H}\beta - \boldsymbol{T} \right\| \tag{23}$$

3) The solution is $\hat{\boldsymbol{\beta}} = \boldsymbol{H}^{\dagger}\boldsymbol{T}$; $H^{\dagger}$ is the Moore-Penrose generalized inverse of the hidden layer output matrix H.

In this study, ELM is used as an excellent traditional neural network prediction component of the combined forecasting system to predict carbon prices.

Combination Strategy

It is generally believed that no single prediction model can achieve the best prediction performance for all datasets. Combining the values predicted by different prediction models usually reduces the overall risk of incorrect model selection. It is hoped that the diversity of models can help improve the final prediction results. However, the previously developed average weighting and weighted weighting methods cannot guarantee the

global optimality of the results (Wang, Y et al., 2018), and it is necessary to find an adaptive variable weight combination strategy.

In this study, the MODA algorithm is used to weigh the three prediction components. For the weighting strategy, we formulated the MODA algorithm as a linear programming (LP) problem to minimize the loss function. These theories are introduced in detail below:

Ensemble Method

Owing to the different weights given to each individual component of the ensemble prediction system, the formula used in the ensemble forecasting method is as follows:

$$\boldsymbol{f}\left(\hat{x}_{c,t}\right) = \sum_{j=1}^{m} \boldsymbol{\omega}_j \boldsymbol{f}\left(\hat{x}_{j,t}\right), \quad t = 1, 2, \ldots \tag{24}$$

where $f(\hat{x}_{c,t})$ is the final output, $\boldsymbol{f}(\hat{x}_{j,t})$ is the prediction component of the EPS, m is the number of submodels, and ω_j is the weight of the component models. The experimental results demonstrate that the ensemble model can obtain ideal results when these weights are in the range of [−2, 2].

Multiobjective Dragonfly Optimization algorithm

The dragonfly algorithm is a population-based heuristic intelligent algorithm that is easy to understand and implement (Mirjalili, 2016). The dragonfly algorithm is inspired by the static and dynamic group behaviors of dragonflies. In the static group behavior, the group preys; in the dynamic group behavior, the group migrates. These two behaviors are very similar to the two important stages in heuristic optimization algorithms: exploration and development. In this research, the MODA is applied to increase the accuracy and stability of the prediction system (Song and Li, 2017).

The mathematical expression methods are as follows:

1) The degree of separation refers to avoiding collisions between dragonflies and adjacent individuals.

$$\boldsymbol{S_i} = \sum_{j=1}^{N} \boldsymbol{X} - \boldsymbol{X_j} \tag{25}$$

2) Alignment means that the trends in movement speed are the same in adjacent individuals.

$$\boldsymbol{A_i} = \frac{\sum_{j=1}^{N} \boldsymbol{V_i}}{\boldsymbol{N}} \tag{26}$$

3) Cohesion refers to the tendency of dragonflies to gather near the center of adjacent individuals.

$$\boldsymbol{C}_i = \frac{\sum_{j=1}^{N} \boldsymbol{X}_j}{\boldsymbol{N}} - \boldsymbol{X} \tag{27}$$

4) Food attraction is the degree of attraction of dragonflies to food.

$$\boldsymbol{F_i} = \boldsymbol{X}^{+} - \boldsymbol{X} \tag{28}$$

5) The repulsive force of natural enemies refers to the repellence of the group to natural enemies when dragonflies encounter natural enemies.

$$\boldsymbol{E}_i = \boldsymbol{X}^{-} + \boldsymbol{X} \tag{29}$$

In **Eq. 25** through **Eq. 29**, X is the position of the current dragonfly individual, $\boldsymbol{X_j}$ represents the position of the j-th adjacent dragonfly, $\boldsymbol{V_j}$ represents the speed of the j-th adjacent dragonfly, N represents the number of individuals adjacent to the i-th dragonfly individual, $\boldsymbol{X}^{+}$ indicates the location of the food source, and $\boldsymbol{X}^{-}$ indicates the position of the natural enemy.

Based on the above five behaviors, the step length and the position of the next generation of dragonflies are calculated as follows:

$$\Delta \boldsymbol{X}_{t+1} = \left(s\boldsymbol{S_i} + a\boldsymbol{A_i} + c\boldsymbol{C_i} + f\boldsymbol{F_i} + e\boldsymbol{E_i}\right) + \omega \Delta \boldsymbol{X}_t \tag{30}$$

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t + \Delta \boldsymbol{X}_{t+1} \tag{31}$$

Whether dragonflies are adjacent to each other can be judged by the Euclidean distance, which is similar to a circle with a radius of r around each dragonfly, and all individuals in the circle are adjacent. To speed up the convergence, the radius r should gradually increase during the iterative process and should finally include the entire search space (Sun et al., 2018). At the beginning of the iteration, the radius r is very small, and some individuals may have no adjacent individuals. To enhance the search power of the algorithm, the random walk is adopted to replace the step update formula, as shown below.

$$\boldsymbol{Levy}(\boldsymbol{x}) = 0.01 \times \frac{\boldsymbol{r}_1 \times \sigma}{|\boldsymbol{r}_2|^{\frac{1}{\beta}}} \tag{32}$$

In **Eq. 32**, r_1 and r_2 represent random numbers between [0,1], β is a constant (here, 3/2), and $\sigma$is calculated as follows:

$$\sigma = \left[\frac{\boldsymbol{\Gamma}\left(1+\beta\right)\boldsymbol{sin}\left(\frac{\pi\beta}{2}\right)}{\boldsymbol{\Gamma}\left(\frac{1+\beta}{2}\right)\beta 2^{\left[\frac{\beta-1}{2}\right]}}\right]^{\frac{1}{\beta}} \tag{33}$$

The corresponding position update formula can be derived as shown in the following formula:

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t + Levy\left(\boldsymbol{d}\right) \times \boldsymbol{X}_t \tag{34}$$

In **Eq. 34**, d represents the dimension of the position vector. In MODA, the nondominated Pareto optimal solution that is obtained in the optimization process is stored and retrieved through the storage unit of the external archive. More importantly, to improve the distribution of solutions in the document and maintain the diversity of Pareto solution sets, the algorithm uses a roulette method with probability $\boldsymbol{P_i} = \boldsymbol{c}/\boldsymbol{N_i}$ to keep the nondominated solution sets well distributed. N_i represents the number of solutions near the i-th solution, and c is a constant.

Objective Function of MODA

Generally, the multiobjective optimization problem can be regarded as the solution of the constraint problem. The

constraint problem with J inequalities and K equations can be written as follows:

$$\boldsymbol{Min}\, F(x) = \left(obf_1(x), obf_2(x), \cdots, obf_s(x)\right)^T \tag{35}$$

$$\begin{aligned} &s.t.\ \boldsymbol{g}_j(\boldsymbol{x}) \geq 0,\ j = 1, 2, ..., J, \\ &\boldsymbol{h}_k(\boldsymbol{x}) = 0, k = 1, 2, ..., K, \end{aligned} \quad x \in \Omega \tag{36}$$

where $\left(obf_1(x), obf_2(x), \cdots, obf_s(x)\right)^T$ is the decision vector.

In this study, the objective of the optimization algorithm is to determine the weight of each single-prediction componentω_j to minimize the error between the final combined forecast value $f(\hat{x}_{c,t})$ and the real value of the carbon price Y. The optimization algorithm can be expressed as follows:

$$\boldsymbol{Min}\begin{cases} \boldsymbol{obf}_1 = \boldsymbol{std}\left(f(\hat{x}_{c,t}), Y\right) = \boldsymbol{std}\left(\sum_{j=1}^{m}\omega_j f\left(\hat{x}_{j,t}\right) - Y_j\right) \\ \boldsymbol{obf}_2 = \boldsymbol{MAPE}\left(f(\hat{x}_{c,t}), Y\right) = \boldsymbol{MAPE}\left(\sum_{j=1}^{m}\omega_j f\left(\hat{x}_{j,t}\right) - Y_j\right) \end{cases} \tag{37}$$

$$s.t.\ -2 \leq \sum_{j=1}^{J}\omega_j \leq 2$$

Therefore, we can solve the component weight ω_j:

$$\boldsymbol{\omega}_j = \boldsymbol{arg}Min\begin{cases} \boldsymbol{Std}\left(f(\hat{x}_{c,t}), Y\right) \\ \boldsymbol{MAPE}\left(f(\hat{x}_{c,t}), Y\right) \end{cases} \tag{38}$$

$$s.t.\ -2 \leq \sum_{j=1}^{J}\omega_j \leq 2$$

Through continuous iteration of the MODA optimization algorithm, the weight vector $\omega = (\omega_1, \omega_2, ..., \omega_j)$ that minimizes the error between the combination forecast value $f(\hat{x}_{c,t})$ and the real value of carbon price Y is obtained. In this study, m = 3.

Uncertainty Mining Module

The uncertainty information of point prediction results can be used to more deeply analyze the characteristics of carbon prices. In this article, an innovative interval prediction scheme based on prediction error distribution modeling in the training stage is proposed. Unlike previous research in which it is assumed that the prediction error follows a Gaussian distribution, this article uses maximum likelihood estimation (MLE) to conduct statistical research on carbon price error data and to explore its distribution. Among the five distribution functions developed, the function that best fits the distribution of carbon price prediction error is found. Based on its probability distribution function (PDF), the upper and lower bounds of the carbon price prediction interval are constructed. The details of the five distribution functions and interval prediction methods are given below.

Distribution Function

The probability distribution function plays a very important role in resource evaluation and interval prediction. This study attempts to use different DFSs to fit the distribution function of prediction error, hoping to analyze the time series in a new way and to mine its uncertainty characteristics. In this section, five types of model prediction error distribution functions (stable, extreme value, normal, logistic, and t location-scale (TLS) functions) are introduced. The relevant probability density functions are shown in **Table 1**.

Interval Prediction Theory

Under the significance level α, for the limit of the model prediction error interval (I_{min} and I_{max}), the probability formula of the prediction model error value $\hat{y}_{err}$ and the prediction error true value Y_{err} can be expressed as follows (Song et al., 2015):

$$\boldsymbol{P}(\boldsymbol{I}_{\min} \leq \boldsymbol{Y}_{err} \leq \boldsymbol{I}_{\max}) = 1 - 2\alpha \tag{39}$$

Since the error value of the prediction model is a random variable, Eq. 50 can also be written as follows:

$$\boldsymbol{P}\{\boldsymbol{I}_{\min} \leq \boldsymbol{Y}_{err} \leq \boldsymbol{I}_{\max} | \boldsymbol{E}(Y_{err}) = \hat{y}_{err}\} = 1 - 2\alpha \tag{40}$$

In addition, we assume that the prediction error of the future prediction model has the same distribution function as that of the historical prediction model. Therefore, the probability distribution function (PDF) based on the historical error data of the prediction system can be regarded as a distribution function of future prediction error (Chen and Liu, 2021). Thus, the upper and lower bounds of the function at a certain confidence level can be calculated.

$$\left\{ (\boldsymbol{I}_{\min}, \boldsymbol{I}_{\max}) \middle| \boldsymbol{I}_{\min} \leq \boldsymbol{Y}_{err} \leq \boldsymbol{I}_{\max}, \int_{I_{\min}}^{I_{\max}} f\left(x|\widehat{\Theta}\right)dx = 1 - 2\alpha \right\} \tag{41}$$

The above equation can also be written as

$$\left[\widehat{\boldsymbol{I}}_{\min}, \widehat{\boldsymbol{I}}_{\max}\right] = \left[\widehat{\boldsymbol{I}}_{\min}, \widehat{\boldsymbol{y}}_{err}\right]\ \left[\widehat{\boldsymbol{y}}_{err}, \widehat{\boldsymbol{I}}_{\max}\right] \tag{42}$$

$$\int_{I_{\min}}^{\widehat{y}_{err}} f\left(x|\widehat{\Theta}\right)dx = \mathrm{F}\left(\widehat{y}_{err}\right) - \alpha \tag{43}$$

$$\int_{\widehat{y}_{err}}^{\widehat{I}_{\max}} f\left(x|\widehat{\Theta}\right)dx = 1 - \mathrm{F}\left(\widehat{y}_{err}\right) - \alpha \tag{44}$$

After the optimal statistical distribution of the prediction error is determined, the upper $\widehat{U}$and lower $\widehat{L}$ bounds of the carbon price prediction interval can be constructed.

$$\left[\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{U}}\right] = \left[\widehat{y}_{forecast} + \widehat{I}_{\min}, \widehat{y}_{forecast} + \widehat{I}_{\max}\right] \tag{45}$$

In **Eq. 45**, $\widehat{y}_{forecast}$ is the carbon price predicted by the carbon price prediction model.

ENSEMBLE PREDICTION SYSTEM AND ITS INTERVAL FORECASTING FRAMEWORK

This section introduces in detail the specific process used in this study. A brief overview of EPS and its uncertainty exploration is shown in **Figure 2**.

TABLE 1 | Probability distribution function (PDF) of the five distribution functions used in the study.

Distribution functions	PDF	Parameters
Extreme value	$f(x;\mu;\sigma)=\sigma^{-1}\exp-\left(\frac{x-\mu}{\sigma}\right)\exp\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right),x>0$	$\mu>0$location parameter $\sigma>0$scale parameter
Logistic	$f(x;\mu;\sigma)=\frac{\exp\left\{\frac{x-\mu}{\sigma}\right\}}{\sigma\left(1+\exp\left[-\frac{(x-\mu)}{\sigma}\right]\right)^2}$	$\mu>0$location parameter $\sigma>0$scale parameter
Normal	$f(x;\mu;\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	$\mu>0$location parameter $\sigma>0$scale parameter
Stable	$f(x;\gamma;\alpha;\beta;\delta)=\exp(-\gamma^{\alpha\vert x\vert}\left[{}^{\alpha}+\delta x\right)$	$0<\alpha<2;-1\le\beta\le1$ shape parameter $0<\gamma<\infty;-\infty<\delta<\infty$ scale parameter
T Location-Scale	$f(x;\mu,\sigma,\upsilon)=\Gamma\frac{\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi\Gamma\left(\frac{\upsilon}{2}\right)}\left(1+\frac{1}{\upsilon}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{\frac{\upsilon+1}{2}}}$	$\upsilon>0$ shape parameter $\mu>0$location parameter $\sigma>0$scale parameter

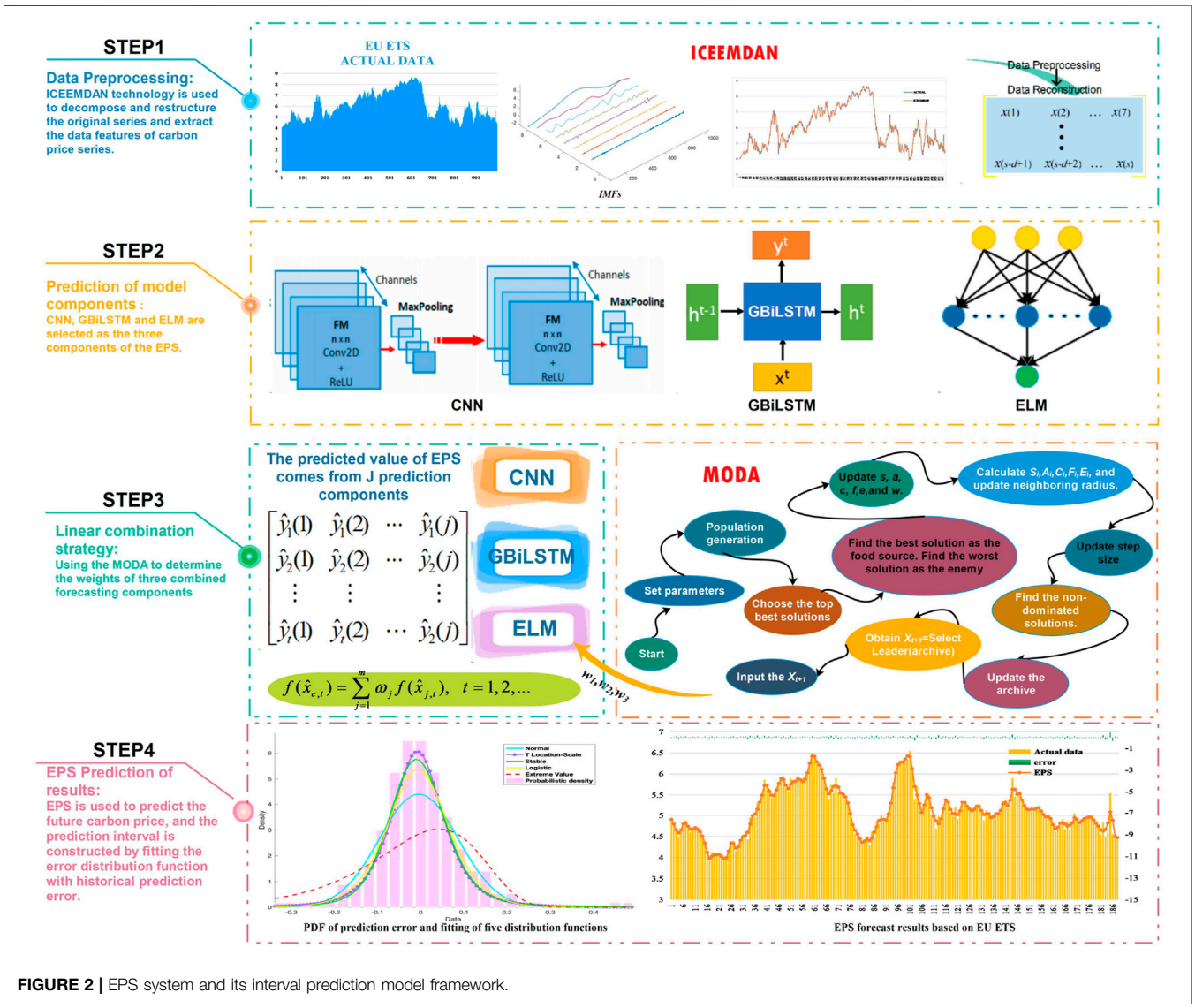


FIGURE 2 | EPS system and its interval prediction model framework.

Step 1: Data Preprocessing and Feature Selection Module

In this article, ICEEMDAN technology is employed to decompose and reconstruct the original carbon price data. ICEEMDAN decomposes the original carbon price into several IMFs and residual terms. The IMF with the highest frequency is then eliminated, and the remaining IMFs are reorganized to extract the effective features of the data. For multivariate time series, effective feature selection is also very important. In this study, partial autocorrelation analysis (PACF) is employed to determine the input feature length of carbon price prediction to achieve feature selection.

Step 2: EPS of Model Components

Owing to the high randomness, volatility, and instability of carbon price data, it is not easy to find its rules of motion, and the single hybrid prediction model has inherent defects. Therefore, the use of a combination of hybrid forecasting models is an effective means of obtaining satisfactory prediction performance and improving prediction accuracy. In this study, two deep learning hybrid models (ICEEMDAN-GBiLSTM and ICEEMDAN-CNN) and a feedforward neural network (ICEEMDAN-ELM) are used as the prediction components of the EPS. They have high prediction accuracy and good learning ability for time series.

Step 3: Component Ensemble Strategy

Given the advantages and disadvantages of different hybrid models, it is very important to select a weight combination strategy with strong adaptability and good fusion effect to compensate for the defects of the individual hybrid models and improve the performance and accuracy of carbon price prediction. Therefore, the MODA is selected to determine the fusion weight among the three prediction model components.

Step 4: Exploring Uncertainty

Quantifying the uncertainty associated with carbon price prediction is a considerable challenge. In this study, a new interval prediction scheme based on forecasting error distribution modeling in the model training stage is proposed. Unlike previous research based on the assumption that the prediction error follows a Gaussian distribution, this article uses MLE to conduct statistical research on carbon price error data and to explore its distribution. Among the five DFs developed, the function that best fits the distribution of carbon price prediction error is found. After confirming that the optimal fit to the distribution of forecast error is provided by the t location-scale, the upper and lower bounds of the carbon price prediction interval are constructed based on its PDF.

EXPERIMENT AND ANALYSIS

This section will introduce the experimental setup and analysis in detail, including the simulation experiment dataset and three different groups of empirical experiments that are used to verify the prediction performance of EPS.

Data Selection and Analysis

In this article, three datasets based on the carbon price market (the EU Emission Trading System (EU ETS), the Shenzhen (SZ), and the Beijing (BJ) datasets) are used as experimental data. The dataset can be downloaded from the wind database (http://www.wind.com.cn/). The first 80% of each dataset is used as the training set, and the last 20% is used as the test set. Specifically, for the EU emission trading system dataset, a total of 1,000 daily quota settlement prices from July 10, 2013 to May 3, 2017 are selected. For the Shenzhen and Beijing datasets, this study used daily spot carbon price data collected from January 14, 2014 to February 7, 2017, including 800 data points. Detailed statistical descriptions of the three datasets are given in **Table 2**. In addition, in constructing the model input vector, we adopted a rolling acquisition mechanism.

Model Parameter Setting

The model parameters determine the performance of the prediction system to a large extent. The different parameters of the EPS proposed in this study are obtained by referring to the literature and to the results of the experiments conducted in this study. The parameter settings for each component of the ensemble system are listed in **Table 3**; this information is valuable and useful because it provides a reference for future research.

Evaluation Index System

To quantify the performance of the developed system, this study constructs an evaluation system using a variety of error evaluation criteria. The system is evaluated and analyzed based on the deterministic point estimation evaluation index and the probabilistic interval estimation evaluation index (Wang R. et al., 2018; Jiang et al., 2021). In the deterministic prediction part, four evaluation indicators, MAE, RMSE, MAPE, and IA, are used. MAE can better express the prediction error under actual conditions. RMSE reflects the deviation between the prediction value and the true value. MAPE expresses the accuracy of prediction using the ratio of error to true value. IA is applied to measure the concordance between the predicted value and the actual value. During interval prediction and evaluation, the three general indicators FICP, FINAW, and AWD are employed to evaluate the quality of the prediction interval. FICP reflects the possibility that the original value falls within the forecast period. FINAW measures the width of the prediction interval. AWD represents the degree of deviation between the observed value and the prediction interval. For FICP, unlike other indicators, a larger value indicates better performance of the model. **Table 4** lists the details of the above evaluation indicators.

Experiment 1: Comparison of Different Data Processing Methods

In this experiment, the original carbon price series and the data based on ICEEMDAN, EEMD, and singular spectrum

TABLE 2 | : Statistical description of the carbon prices reported at three sites.

Statistical Indicators		Number	Max	Min	Mean	Std
Equation		—	—	—	$\text{Mean} = \sum_{i=1}^{N} x_i / N$	$S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
EU ETS	Total	1,000	8.67	3.93	6.01	1.24
	Training	800	8.67	4.02	6.24	1.26
	Testing	200	6.54	3.93	5.08	0.57
BJ	Total	800	77	30.63	49.56	7.08
	Training	640	77	30.63	48.96	7.65
	Testing	160	69	39.45	51.98	3.02
SZ	Total	800	88.45	17.83	43.13	16.05
	Training	640	88.45	18.98	46.92	15.58
	Testing	160	42.81	17.83	28.01	5.39

TABLE 3 | Model parameters.

Model	Parameters	Default value
MODA	Maximum number of iterations	50
	Maximum number of archives	500
	Dragonfly number	30
	Upper and lower limits of the weight coefficient	[–2,2]
	Objective functions	$Min\begin{cases} obf_1 = Std(f(\tilde{x}_{c,t}), Y) \\ obf_2 = MAPE(f(\tilde{x}_{c,t}), Y) \end{cases}$
ICEEMDAN	Noise standard deviation	0.05
	Number of realizations	50
	Maximum number of sifting iterations	500
ELM	Input nodes number	Based on PAC
	Output nodes number	1
	Hidden nodes number	5
	Learning rate	0.001
	Iterations number	200
GBiLSTM	Number of inputs	Based on PAC
	Number of hidden units	200
	Number of outputs	1
	Maximum iteration	250
	Initial learning rate	0.01
	Training algorithm	Adam
CNN	Number of inputs	Based on PAC
	Number of outputs	1
	The kernel sizes	2×2
	The activation function	Relu
	The max pooling layer size	2
	Training algorithm	Adam

decomposition (SSA) are used as the training input for different prediction models. The purpose of this experiment is to explore the effect of using different signal decomposition techniques on the prediction accuracy of prediction models. **Table 6** compares the results obtained using the corresponding models.

Feature Selection Analysis

The prediction performance of both machine learning and deep learning methods is closely related to the input variables. The PACF method is used to select appropriate features as the best input of the prediction model. The best input characteristics obtained from the PACF results of each subsequence are shown in **Table 5**. (In the follow-up experiments, the input units of each prediction model were obtained according to the PACF results.)

Prediction Results Obtained Using the Different Data Preprocessing Methods

To verify the effectiveness of the ICEEMDAN data preprocessing method in data feature extraction, in this experiment the performance of ICEEMDAN is compared with that of the classical feature extraction methods EEMD and SSA. The detailed results are described below.

TABLE 4 | Basic evaluation metrics.

Metric	Definition	Equation
MAE	The mean absolute error	$\boldsymbol{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left\|y_{PRE(i)} - y_{ACT(i)}\right\|$
RMSE	Root mean squared error	$\boldsymbol{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{PRE(i)} - y_{ACT(i)})^2}$
MAPE	The mean absolute percentage error	$\boldsymbol{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left\|\frac{y_{PRE(i)} - y_{ACT(i)}}{y_{ACT(i)}}\right\| \times 100\%$
IA	Concordance index	$\boldsymbol{IA} = 1 - \frac{\sum_{i=1}^{N}(y_{PRE(i)} - y_{ACT(i)})^2}{\sum_{i=1}^{N}(\left\|y_{PRE(i)} - \bar{y}_{ACT}\right\| + \left\|y_{ACT(i)} - \bar{y}_{ACT}\right\|)^2}$
FICP	Forecast interval coverage probability	$\boldsymbol{FICP} = \sum_{i=1}^{N} c_i \times 100\%/N$
FINAW	Forecast interval normalized average width	$\boldsymbol{FINAW} = \sum_{i=1}^{N}(U_i - L_i) \times 100\%$
AWD	Accumulated width deviation of testing dataset	$\boldsymbol{AWD} = \sum_{i=1}^{N} AWD_i/N$ $\boldsymbol{AWD}_i = \begin{cases} L_i - y_{ACT(i)}/U_i - L_i, & y_{ACT(i)} < L_i \\ 0, L_i \le y_{ACT(i)} \le U_i \\ y_{ACT(i)} - U_i/U_i - L_i, & y_{ACT(i)} > U_i \end{cases}$

Note: *This table lists the full names and calculation methods of the evaluation indices included in the evaluation system.* N *is the size of the test sample,* $\bar{y}$ *is the average value ofy,* $y_{ACT(i)}$ *is the* i*-th actual value, and* $y_{PRE(i)}$ *is the* i*-th forecast value.* U_i *and* L_i *represent the upper and lower limits, respectively, of the prediction interval.* C_i *represents the number of true values contained in the construction interval* [U_i, L_i] and is the *i*-th prediction interval.

TABLE 5 | Optimal input characteristics based on the PACF.

Site	Input combination
EU ETS	$x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}, x_{t-9}$
SZ	$x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}, x_{t-9}, x_{t-10}$
BJ	$x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}, x_{t-9}, x_{t-10}$

From the results in **Table 6**, it can be seen that the use of data preprocessing technology can effectively ameliorate the prediction ability of the prediction model. MAPE, MAE, RMSE, and IA were adopted to evaluate the prediction ability of the model based on the prediction accuracy and the fitting situation. For the data on the three carbon trading markets, the MAPE, the MAE, and the RMSE of the prediction model are significantly lower than those of the model that was directly trained on the original dataset, regardless of which data preprocessing technology is adopted. When the original dataset is used directly, the $MAPE_{actual}$ of the prediction model under the EU ETS, SZ, and BJ datasets is between 3 and 14%, while the $MAPE_{ICEE}$of the prediction model is reduced to 1–5% when ICEEMDAN noise reduction technology is used. This is sufficient to indicate the necessity for data preprocessing.

In addition, for the three components of EPS, different data preprocessing methods are used as model inputs. The experimental results show that ICCEEMDAN is more effective than the other methods. For the EU ETS dataset, the prediction system using ICEEMDAN noise reduction technology has the highest prediction accuracy, and the average RMSE value of the three component models, $\overline{\boldsymbol{RMSE}}_{ICEE}^{EUETS} = 0.1081$, is the best; the prediction performance of the model based on EEMD is the worst: $\overline{\boldsymbol{RMSE}}_{EEMD}^{EUETS} = 0.1313$ in the SZ dataset, $\overline{\boldsymbol{MAE}}_{ICEE}^{SZ} = 1.2518$, $\overline{\boldsymbol{RMSE}}_{ICEE}^{SZ} = 1.9671$, and $\overline{\boldsymbol{MAPE}}_{ICEE}^{SZ} = 4.8261\%$. In the BJ dataset, the average values of these three indicators are 0.5573, 1.2169, and 1.0573%, obviously better than those obtained using other data processing technologies.

IA is an effective index that can be used to measure the correlation and consistency between the predicted values and the original data. The higher the index value is, the better the fitting effect of the model is. The ICEEMDAN feature extraction technology proposed in this article achieves the highest IA of all the models tested under the three carbon price datasets. In the SZ dataset, the IA values are $\boldsymbol{IA}_{ICEE-ELM}^{SZ} = 0.9644$, $\boldsymbol{IA}_{ICEE-CNN}^{SZ} = 0.9529$, and $\boldsymbol{IA}_{ICEE-BiLSTM}^{SZ} = 0.9696$. These values are 0.0138, 0.0025, and 0.0173 units higher than $\mathrm{IA}_{SSA-ELM}^{SZ}$, $\boldsymbol{IA}_{SSA-CNN}^{SZ}$, and $\boldsymbol{IA}_{SSA-GBiLSTM}^{SZ}$, respectively, and 0.0628, 0.0504 and 0.0657 units higher than $\boldsymbol{IA}_{EEMD-ELM}^{SZ}$, $\boldsymbol{IA}_{EEMD-CNN}^{SZ}$和$\boldsymbol{IA}_{EEMD-GBiLSTM}^{SZ}$, respectively. In summary, compared with other data preprocessing technologies, ICEEMDAN data preprocessing is more effective for data feature extraction and has incomparable advantages in improving the performance of the prediction model.

Key Finding: Compared with the original carbon price series and other classical data preprocessing techniques, ICEEMDAN preprocessing technology can extract the data characteristics of carbon prices more effectively, significantly enhances the prediction accuracy of the prediction model, and is a more reliable data preprocessing tool.

Experiment 2: Point Forecasting: Comparison of the EPS With Reference Models

To verify the effectiveness of EPS in carbon price prediction, the traditional single forecast model and the classical hybrid prediction model are compared with EPS. These models include the traditional statistical models ARIMA and ICEEMDAN-ARIMA, the traditional single neural network models BP, ELM, GRNN, the deep learning models LSTM, CNN, and the classical hybrid prediction models GWO-BP and ICEEMDAN-GWO-BP. In addition, to explore the expansibility of the model, the experimental content of multistep point prediction is included in the experiment. In the multistep prediction, rolling prediction is adopted. The specific method used to perform multistep prediction is shown in **Figure 3**. The experimental results are shown in

TABLE 6 | Comparison of the performances of prediction models based on different data feature extraction techniques.

Model	EU ETS				SZ				BJ			
—	MAPE	RMSE	MAE	IA	MAPE	RMSE	MAE	IA	MAPE	RMSE	MAE	IA
ELM	3.1190	0.2148	0.1583	0.9574	13.3955	4.4118	3.5126	0.7965	3.9351	4.0281	2.5107	0.7225
CNN	3.1581	0.1659	0.1659	0.9544	12.5333	4.4030	3.3666	0.8007	3.7483	3.8029	2.4362	0.7236
GBiLSTM	3.3159	0.2241	0.1744	0.9523	13.1742	4.3402	3.4707	0.7955	3.7406	3.7846	2.4355	0.7239
EEMD-ELM	1.8531	0.1235	0.0942	0.9857	7.8973	2.8796	2.2638	0.9016	2.9211	2.2281	1.5135	0.8443
EEMD-CNN	1.8892	0.1249	0.0962	0.9880	7.8426	2.9683	2.2237	0.9025	2.9058	2.2111	1.5042	0.8483
EEMD-GBiLSTM	2.2015	0.1456	0.1117	0.9829	7.6835	2.5719	2.0024	0.9029	2.8966	2.2051	1.4991	0.8495
SSA-ELM	1.4597	0.1021	0.0734	0.9922	5.8985	2.2233	1.6648	0.9506	1.5922	1.4362	0.8198	0.9318
SSA-CNN	1.5684	0.1100	0.0795	0.9905	5.9112	2.2256	1.6727	0.9504	1.5435	1.4568	0.8079	0.9139
SSA-GBiLSTM	2.0231	0.1293	0.1009	0.9861	5.8622	2.2085	1.6519	0.9513	1.5479	1.4577	0.8085	0.9137
ICEE-ELM	1.2806	0.0915	0.0647	0.9934	4.5032	1.7141	1.2424	0.9644	1.0571	1.2013	0.5561	0.9397
ICEE-CNN	1.5162	0.1075	0.0771	0.9910	5.6835	2.5719	1.3624	0.9529	1.0624	1.2485	0.5596	0.9359
ICEE-GBiLSTM	1.8746	0.1253	0.0941	0.9878	4.2916	1.6153	1.1508	0.9686	1.0525	1.2008	0.5561	0.9399
EPS	**1.2657**	**0.0904**	**0.0640**	**0.9936**	**4.0156**	**1.6096**	**1.1372**	**0.9687**	**1.0064**	**1.2049**	**0.5312**	**0.9402**

Note: *The best indicator values are shown in bold type.*

Table 7. The detailed experimental analysis is described below.

1) In comparison with the traditional single prediction model, we find that EPS displays incomparable advantages in the four indicators in both one-step and multiple-step prediction. This shows that the EPS developed by us is effective in predicting carbon prices. In addition, the MAPE values of GBiLSTM in the three datasets are $MAPE^{EUETS}_{GBiLSTM} = 3.3159$, $MAPE^{SZ}_{GBiLSTM} = 13.1740$, and $MAPE^{BJ}_{GBiLSTM} = 3.7406$; these values are better than those obtained using a single LSTM, proving the effectiveness of the GBiLSTM. The other two prediction components, ELM and CNN, have outstanding prediction performance in all single prediction models, so it is reasonable to choose them as the submodes of the EPS. In addition, we can see that for the traditional statistical model ARIMA, the average value of MAPE of the three stations is $\overline{\text{MAPE}}_{ARIMA} = 11.0072$in one-step prediction; this is not as high as the prediction accuracy achieved using other prediction models, indicating that the traditional linear statistical model is not suitable for the prediction of carbon price series with high volatility and complexity.
2) We can observe that under different datasets, different prediction models have different prediction performances. Under the EU ETS dataset, the neural network ELM performs best, yielding $RMSE^{EUETS}_{ELM} = 0.2148$; this is better than the RMSE values of the deep learning algorithms CNN and GBiLSTM, which are $RMSE^{EUETS}_{CNN} = 0.2218$ and $RMSE^{EUETS}_{BiLSTM} = 0.2241$, respectively. However, under the SZ and BJ datasets, the deep learning algorithms CNN and GBiLSTM achieve better prediction results than ELM. The same phenomenon occurs in multistep forecasting, and there the forecast advantage of the deep learning framework is more obvious. However, the prediction accuracy of EPS remains the highest under any of the tested datasets. The RMSE values in the one-step forecast are $RMSE^{EUETS}_{EPS} = 0.0904$, $RMSE^{SZ}_{EPS} = 1.6096$, and $RMSE^{BJ}_{EPS} = 1.2046$. The RMSE values in the two-step forecast are $RMSE^{EUETS}_{EPS} = 0.1829$, $RMSE^{SZ}_{EPS} = 3.3244$, and $RMSE^{BJ}_{EPS} = 1.7567$. This shows that the combination strategy retains the forecasting advantages obtained by using different forecasting components and that it compensates for each model's defects; as a result, EPS has strong robustness and wide adaptability.

In comparison with the classic hybrid forecasting models ICEE-GWO-BP, SSA-GRNN, and GWO-BP, several sets of hybrid forecasting methods have achieved good forecasting performance; however, because the index values used in these models are very similar, it is not easy to intuitively present the predictive ability of the model. In this case, we measure the percentage of improvement in the evaluation index to make the analysis more intuitive. The percentage of improvement in the evaluation index is a measure of the degree of improvement achieved by EPS compared with the index value of the comparison model; it can be expressed as follows:

$$\boldsymbol{P}^{model}_{Index} = \left|\frac{\boldsymbol{Index}_{model} - \boldsymbol{Index}_{EPS}}{\boldsymbol{Index}_{model}}\right| \times 100\% \tag{46}$$

where P^{model}_{Index} is the percentage of improvement indicators, Index_{model} stands for the index value of the comparison model, and Index_{EPS} is the index value of the EPS.

In the EU ETS dataset, the improved MAPE values for one-step prediction of the three hybrid models are $P^{ICEE-GWO-BP}_{MAPE} = 1.9673\%$, $P^{SSA-GRNN}_{MAPE} = 60.4950\%$, and $P^{GWO-BP}_{MAPE} = 59.5377\%$. The improved MAPE values of the two-step prediction are $P^{ICEE-GWO-BP}_{MAPE} = 9.2410\%$, $P^{SSA-GRNN}_{MAPE} = 30.7285\%$ and $P^{GWO-BP}_{MAPE} = 51.8867\%$. Under the BJ dataset, the improved IA values predicted for the three hybrid models using the one-step method are $P^{ICEE-GWO-BP}_{IA} = 0.9064\%$, $P^{SSA-GRNN}_{IA} = 48.1563\%$, and $P^{GWO-BP}_{IA} = 33.3896\%$. The improved IA values obtained by two-step prediction are $P^{ICEE-GWO-BP}_{MAPE} = 1.4313\%$, $P^{SSA-GRNN}_{MAPE} = 97.2385\%$, and $P^{GWO-BP}_{MAPE} = 104.3018\%$. The index improvement percentage more intuitively shows the improvement in prediction performance obtained using EPS. Compared with the classical hybrid prediction model, the EPS shows significant improvement in both prediction accuracy and fitting consistency.

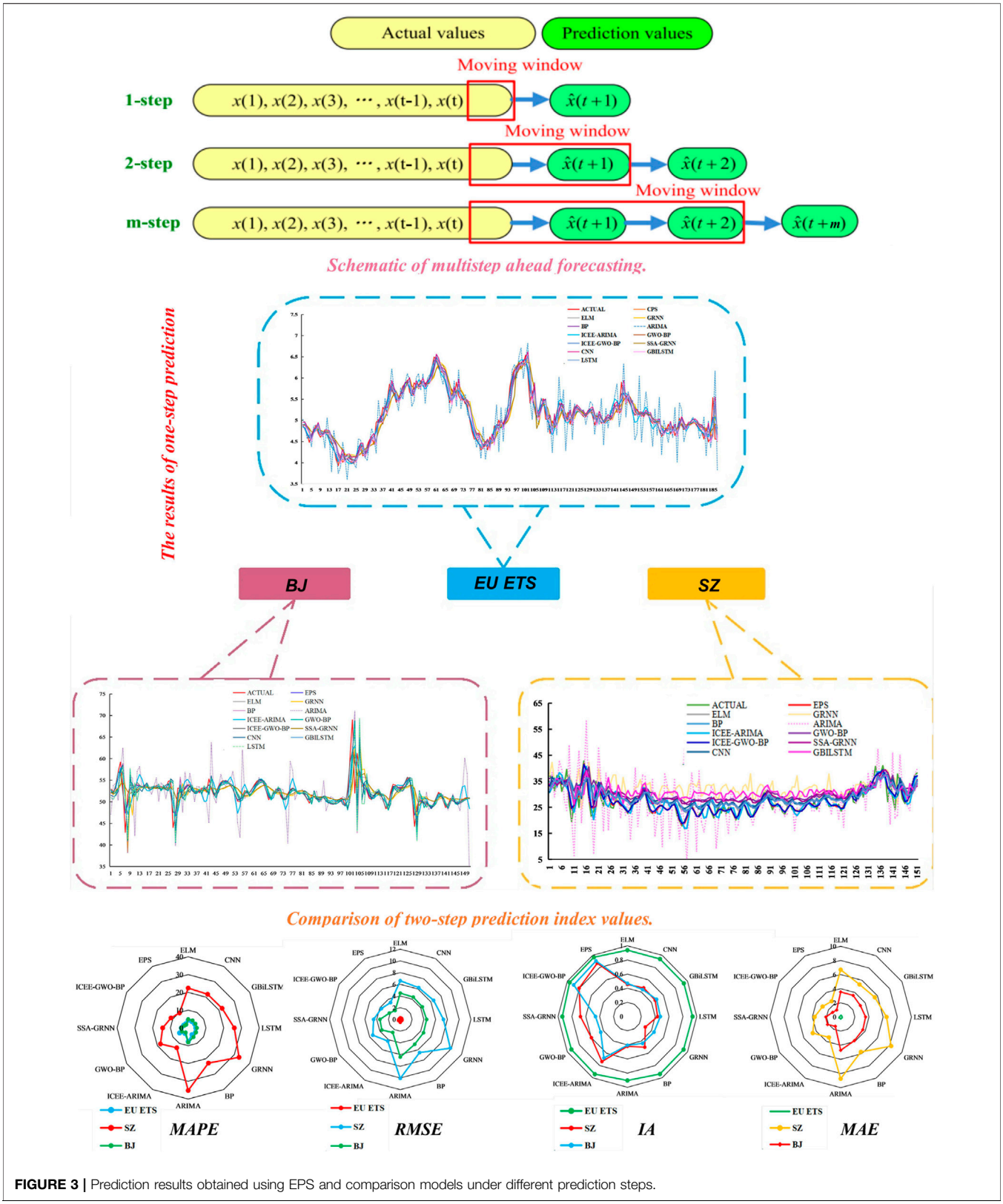


FIGURE 3 | Prediction results obtained using EPS and comparison models under different prediction steps.

Figure 3 shows a comparison of the prediction results obtained using EPS and the comparison model when different numbers of prediction steps are used.

Key finding: The difference in the prediction results between the EPS system and other prediction models is significant. Specifically, under each dataset and for each prediction step,

TABLE 7 | Comparison of the prediction ability of the proposed system with those of some traditional benchmark models and classic hybrid models.

Dataset	Model	ONE-STEP				TWO-STEP			
		MAPE	RMSE	MAE	IA	MAPE	RMSE	MAE	IA
EU ETS	**ELM**	3.119	0.2148	0.1583	0.9574	4.2962	0.2842	0.2169	0.9386
	CNN	3.1581	0.2218	0.1659	0.9544	4.1304	0.2805	0.2112	0.9392
	GBiLSTM	3.3159	0.2241	0.1744	0.9523	4.1862	0.2794	0.2133	0.9390
	LSTM	3.5642	0.2470	0.1815	0.9407	4.2575	0.2827	0.2164	0.9388
	GRNN	3.5405	0.2432	0.1799	0.9418	4.7743	0.3232	0.2432	0.9341
	BP	3.2941	0.2235	0.1704	0.9516	4.3052	0.2903	0.2193	0.9380
	ARIMA	5.8692	0.3789	0.2947	0.9327	7.6365	0.4814	0.3011	0.9036
	ICEE-ARIMA	1.9758	0.1323	0.1000	0.9862	3.6109	0.2401	0.1839	0.9403
	GWO-BP	3.1281	0.2149	0.1584	0.9571	5.5972	0.3500	0.2828	0.9322
	SSA-GRNN	3.2039	0.2170	0.1621	0.9548	3.8876	0.2632	0.1986	0.9385
	ICEE-GWO-BP	1.2911	**0.0891**	0.0644	0.9862	2.9672	0.1988	0.1506	0.9617
	EPS	**1.2657**	0.0904	**0.0640**	**0.9936**	**2.6930**	**0.1829**	**0.1378**	**0.9674**
SZ	**ELM**	13.3955	4.4118	3.5126	0.7965	22.4148	6.6057	6.6813	0.4538
	CNN	12.5333	4.4030	3.3666	0.8007	21.8746	6.2643	5.3190	0.4686
	GBiLSTM	13.1740	4.3402	3.4707	0.7955	22.034	6.5272	5.5683	0.4617
	LSTM	15.1983	6.8455	4.5637	0.7514	25.9514	7.3299	6.4834	0.4352
	GRNN	13.7152	5.1715	3.8576	0.7654	33.0502	9.8132	8.2422	0.2973
	BP	14.1428	5.8032	4.2144	0.7608	22.8251	6.5453	5.7117	0.5004
	ARIMA	21.8403	7.8476	5.9355	0.6595	35.1742	10.0714	8.7415	0.4211
	ICEE-ARIMA	8.9814	3.2795	2.5983	0.8943	12.8818	4.2559	3.3362	0.7351
	GWO-BP	12.3461	4.1508	3.2907	0.8017	18.0382	5.3858	4.5591	0.5993
	SSA-GRNN	5.9875	2.2372	1.6791	0.9494	14.3684	4.5285	3.6886	0.6773
	ICEE-GWO-BP	4.6025	1.8286	1.3027	0.9631	11.1825	3.7085	2.9526	0.8039
	EPS	**4.0156**	**1.6096**	**1.1372**	**0.9687**	**9.7600**	**3.3244**	**2.5621**	**0.8701**
BJ	**ELM**	3.9351	4.0281	2.5107	0.7225	4.8978	4.4902	3.5380	0.4689
	CNN	3.7483	3.8029	2.4362	0.7236	4.8724	4.4065	3.5240	0.4410
	GBiLSTM	3.7406	3.7846	2.4355	0.7239	4.6629	4.3132	3.2137	0.4852
	LSTM	3.8462	3.9033	2.4871	0.7231	4.8787	4.4753	3.5231	0.4749
	GRNN	4.1345	4.3247	2.9737	0.6273	5.0379	4.5516	3.6191	0.4416
	BP	4.0350	4.2179	2.8450	0.6355	5.9108	4.9031	3.8741	0.4379
	ARIMA	5.3120	4.6312	3.7397	0.5627	8.1714	6.4412	4.7140	0.4058
	ICEE-ARIMA	2.4705	2.0481	1.2635	0.9045	2.9320	2.5669	1.5213	0.6763
	GWO-BP	2.9202	3.0563	1.5062	0.7113	4.0738	3.6183	2.1437	0.4440
	SSA-GRNN	3.0901	2.6153	1.5929	0.6346	3.9989	3.4553	2.0979	0.4599
	ICEE-GWO-BP	1.1544	1.2383	0.5966	0.9488	2.4931	2.1482	1.3095	0.8943
	EPS	**1.0064**	**1.2049**	**0.5312**	**0.9402**	**2.1558**	**1.7567**	**1.1272**	**0.9071**

Note*: ICEE is an abbreviation for ICEEMDAN. The best indicator values are shown in bold type.*

the EPS has better prediction performance. Therefore, it is concluded that the advanced ensemble prediction system has better carbon price forecasting ability and potential than the traditional single models and classical hybrid models.

Experiment 3: Interval Forecasting: Uncertainty Analysis of Carbon Price

In Experiment 2, the accuracy and stability of the prediction system were discussed through the prediction method of certainty point estimation. However, the point prediction results do not reflect the uncertainty in the dataset. To further prove that the ESP system has a wider range of adaptability than other predictive models, this section uses the interval prediction method to mine the uncertainty of carbon prices. Unlike point prediction, interval prediction can provide the upper and lower bounds of the observed value, making it possible to construct the prediction interval under a given significance level. It can provide additional information for carbon price market policymakers and can help them analyze the carbon price market.

Distribution Function of Prediction Error

In previous studies, most of the prediction errors of the prediction model defaulted to obey the normal distribution. However, the normal distribution function does not effectively reflect the distribution of forecast model errors. Therefore, this research develops five fitting distribution functions and uses the MLE method to conduct an in-depth investigation of the prediction error to obtain a distribution function (DF) with better fitting performance. The most suitable probability distribution for further interval prediction is selected.

In this section, five DFs, namely, extreme value, normal, logistic, stable, and t location-scale, are used to represent the distribution of carbon price prediction errors. **Table 1** shows the relative PDF of these DFs. **Table 8** lists the five DF parameters estimated by the MLE method. These parameters can be used to

TABLE 8 | Parameter values of the different distribution functions determined by MLE.

Datasets	Extreme value		Logistic		Normal		Stable				T Location-scale		
—	μ	σ	μ	σ	μ	σ	α_1	α_2	β	δ	μ	σ	υ
EU ETS	0.0197	0.0342	0.0028	0.0189	0.0030	0.0334	1.9347	–0.0264	0.0227	0.0031	0.0030	0.0314	19.0078
SZ	0.2878	0.7713	–0.0660	0.3496	–0.0542	0.6631	1.7226	0.2460	0.3768	–0.092	–0.0691	0.5031	4.7444
BJ	0.1150	0.4266	–0.0642	0.1978	–0.0839	0.4106	1.2249	–0.1238	0.1469	-0.047	–0.0531	0.1665	1.5469

TABLE 9 | R^2 and RMSE values of the five distribution functions by MLE.

Datasets	EU ETS		SZ		BJ	
—	R^2	RMSE	R^2	RMSE	R^2	RMSE
Extreme value	0.9086	1.1897	0.8115	0.0941	0.6103	0.2913
Logistic	0.9546	0.8382	0.9685	0.0384	0.8780	0.1630
Normal	0.9760	0.6095	0.9512	0.0479	0.7207	0.2466
Stable	0.9771	0.5956	0.9507	0.0481	0.9668	0.0851
T location-scale	0.9877	0.4355	0.9791	0.0313	0.9671	0.0846

describe the scale and location of these DFs. In addition, the coefficient of determination ($0 \leq R^2 \leq 1$) and RMSE are used to determine the degree of fit of these DFs. The larger the R^2 value is, the lower the RMSE value is, and the better is the fitting ability of the DFs. The index values reflecting the fitting abilities of the five DFs are shown in **Table 9** and **Figure 4**.

Table 9 shows that the t location-scale function fits the EPS prediction error best. Its R^2 value is higher than 0.96, and its RMSE value is the lowest, indicating that it can provide better estimates in all cases, followed by stable distribution, normal distribution, logistic distribution, and extreme value distribution. In addition, although the stable distribution has a slightly worse fitting effect than the t location-scale distribution, it is still better than the normal distribution that the previous prediction error hypothesis obeys; this further proves the necessity of fitting the distribution of the prediction error. In addition, the motivation for estimating the distribution function of the carbon price dataset in this section is to prepare for further research on the establishment of carbon price interval predictions, as discussed in Section *Interval Prediction of Carbon Price*.

Interval Prediction of Carbon Price

Unlike the deterministic information given by the point forecast, the interval forecast can provide the forecast range, a confidence level, and other uncertain information on future values; this information is helpful to decision-makers who are attempting to analyze and supervise the reasonable operation of the carbon price market. Owing to the generalization ability of the forecasting model, the complex patterns of carbon price series fluctuations and other factors inevitably produce forecast errors, and the ability to effectively transform the uncertainty caused by forecast errors into measurable features is of great significance. Therefore, in this study, a new interval prediction scheme based on modeling of the prediction error distribution in the model training phase is proposed.

According to the point prediction results of the proposed EPS system, the t location-scale distribution function, which determines the prediction error in Section *Distribution Function of Prediction Error*, and the interval prediction method introduced in Section *Interval Prediction Theory*, the prediction interval is constructed under the given significance level α. To verify that the prediction interval constructed by the t location-scale model has the best fit, it is compared with the other four error distribution functions.

In addition, three evaluation indicators, FINAW, PICP, and AWD, listed in **Table 4**, are introduced in this section to present the effect of interval prediction. It is worth mentioning that the optimal interval prediction should satisfy the following conditions: under a given significance level α, the larger the PICP value is ($0 \leq PICP \leq 1$), the smaller the FINAW value is, and the better is the prediction performance of interval prediction. However, it is obvious that there is a contradiction between FINAW and PICP. When the PICP value increases, the FINAW of the response average bandwidth will certainly increase. Therefore, the AWD index is introduced as a supplement to the evaluation index system. **Table 10** shows the prediction intervals of the EPS system as evaluated based on the three carbon price markets using five different error distributions.

In theory, when the PICP is greater than the significance level, the constructed prediction interval is effective. As seen from **Table 10**, the models satisfying this condition are EPS-TLS and EPS-Extreme value; these models are effective at the 95, 90, and 80% significance levels. However, if the value of the PICP is the only goal, the result will become meaningless, as increased PICP inevitably leads to a larger FINAW. Based on different αconditions, the value of $FINAW_{EPS-Extreme\ value}$is significantly higher than the FINAW value obtained through modeling of other distributions. At the same time, considering that in Section *Distribution Function of Prediction Error*, the fitting of extreme value distribution to EPS prediction error is very bad, it can be considered that the prediction interval constructed by EPS-Extreme value is not reasonable.

The PICP of EPS-TLS in the data from the three carbon trading markets is higher than the significance levelα. In addition, under differentα, the FINAW values of the three markets are $FINAW^{EUETS}_{EPS-TLS} = 0.0542$, $FINAW^{SZ}_{EPS-TLS} = 0.1225$, and $FINAW^{BJ}_{EPS-TLS} = 0.1397$. The FINAW value is not optimal in any of the five interval prediction models, but with only a small increase in the FINAW value, the other index values achieve better results. All things considered, this can be accepted.

For the prediction interval constructed using a normal distribution, in the BJ dataset, the PICP values of EPS-Normal

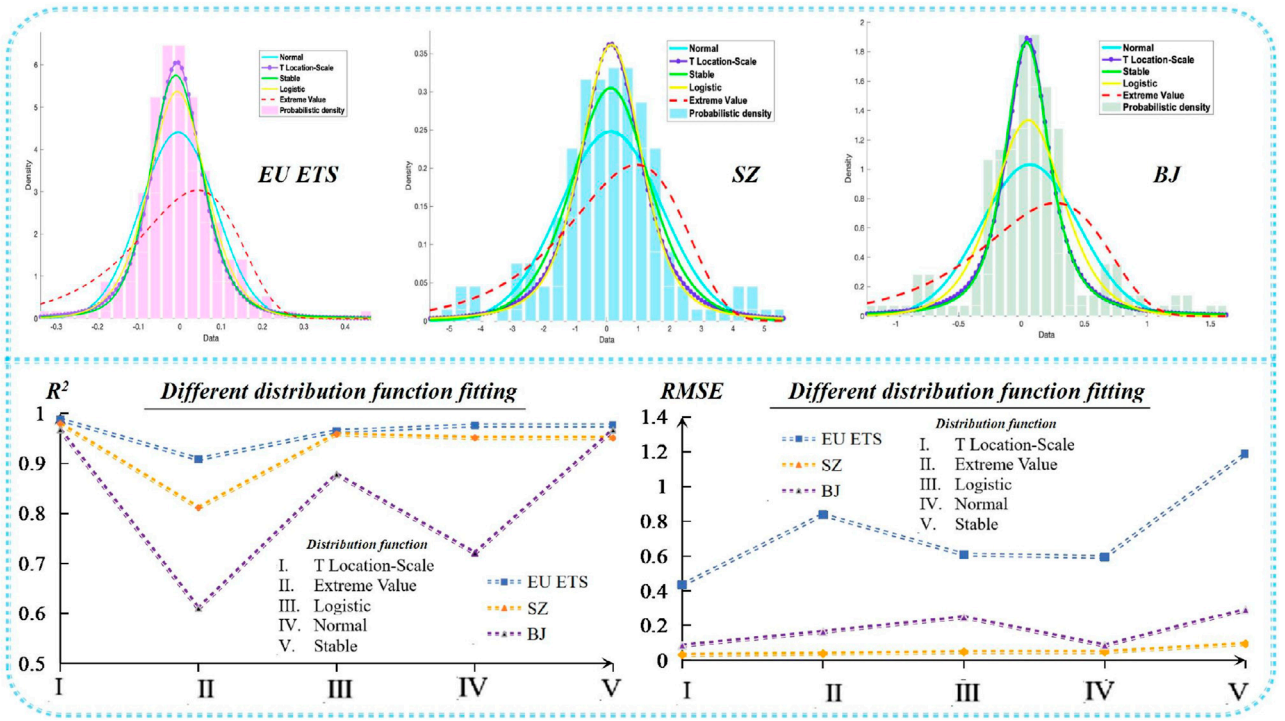


FIGURE 4 | Five distribution functions fit the distribution of EPS error.

are $PICP^{BJ}_{EPS-Normal}$ = 92.0530 and $PICP^{BJ}_{EPS-Normal}$ = 88.0795 under significance levels of 0.95 and 0.90, respectively; they fail to meet the condition of a level of α that is greater than significance. This also reflects the necessity of detailed research on error distribution. For AWD indicators, although $AWD_{EPS-TLS}$ is not all better than the benchmark model, there is little difference. Considering the comprehensive performance of the three indicators of the proposed scenario, EPS-TLS still has obvious advantages over the four benchmark models in constructing the prediction interval.

In addition, the carbon price prediction intervals generated by the three proposed schemes of the three carbon trading markets are shown in **Figure 5**. It can be observed that EPS-TLS has a smaller bandwidth and is surrounded by these constructed prediction intervals in most target values. The constructed confidence interval is very effective.

DISCUSSION

In this section, we will discuss the robustness, application, and further development of EPS in the carbon price market.

Robustness Discussion

Because the results of both deep learning and metaheuristic optimization algorithms are always accompanied by randomness and probability mechanisms, the results of each experiment will still have deviations even when the parameters are set exactly the same. At the same time, in the actual forecast, the actual values of the future carbon price cannot be predicted in advance; thus, it is impossible to use the evaluation index to verify the future value in advance. Therefore, the stability of the EPS is also an important factor that affects the prediction.

The standard deviation is an effective measure of system stability. It can be indicated as $SD_{(M)} = \sqrt{\sum_k^n (M_k - \overline{M})^2/n}$, where n is the number of training iterations, M_K is the predicted value of the K-th training result, and M is the average of the N-th results (Xiao et al., 2017). The smaller the value of SD, the higher the stability of the model.

To evaluate the stability of the different models, the $SD_{(M)}$ values of four evaluation indices were calculated in 30 prediction experiments using three carbon price datasets.

Table 11 shows a comparison of the stability test results of different prediction systems based on ICEEMDAN processing. In the EU ETS dataset, ICEE-ELM has good stability ($SD^{ICEE-ELM}_{(MAPE)}$ = 0.0105), but the stability is still slightly lower than that of the EPS prediction system ($SD^{EPS}_{(MAPE)}$ = 0.0101). In the BJ and SZ datasets, CNN has good prediction accuracy in previous experiments, but its robustness is not good, and the prediction results fluctuate greatly. In contrast, EPS obtains a smaller SD value regardless of which dataset is used. This further shows that different single prediction systems have different robustness when used with different datasets and indicates that EPS with a combination weighting strategy can be considered the prediction method that obtains the best prediction results.

It is worth mentioning that the average prediction stability of GBiLSTM in the three prediction datasets is better than that of the traditional LSTM model; that is,$SD^{\overline{GBiLSTM}}_{(RMSE)}$ = 0.0496, but $SD^{\overline{LSTM}}_{(RMSE)}$ = 0.0566. The results show that the proposed

TABLE 10 | Carbon price range prediction results based on EU ETS, SZ, and BJ under different significance levels.

Site	PINC	Distribution	PICP	FINAW	AWD
EU ETS	95%	EPS-*Extreme value*	97.3822	0.0701	0.0027
		EPS-*Logistic*	83.2461	0.0387	0.0354
		EPS-*Normal*	95.2880	0.0516	0.0083
		EPS-*Stable*	95.8115	0.05361	0.0073
		EPS-*TLS*	95.8115	0.0542	0.0070
	90%	EPS-*Extreme value*	96.8586	0.0573	0.0062
		EPS-*Logistic*	74.8691	0.0316	0.0584
		EPS-*Normal*	90.0109	0.0412	0.0194
		EPS-*Stable*	90.5759	0.0443	0.0162
		EPS-*T LS*	90.0524	0.0426	0.0188
	80%	EPS-*Extreme value*	90.5759	0.0434	0.0155
		EPS-*Logistic*	62.3037	0.0240	0.1170
		EPS-*Normal*	80.5340	0.0331	0.0460
		EPS-*Stable*	79.5812	0.0342	0.0419
		EPS-*T LS*	81.675	0.0352	0.0378
Site	**PINC**	**Distribution**	**PICP**	**FINAW**	**AWD**
SZ	95%	EPS-*Extreme value*	98.0132	0.1811	0.0029
		EPS-*Logistic*	94.7020	0.1191	0.0115
		EPS-*Normal*	95.3642	0.1240	0.0105
		EPS-*Stable*	95.3642	0.1207	0.0112
		EPS-*TLS*	95.3642	0.1225	0.0100
	90%	EPS-*Extreme value*	95.3642	0.1477	0.0051
		EPS-*Logistic*	92.0530	0.0970	0.0224
		EPS-*Normal*	90.0514	0.1028	0.0183
		EPS-*Stable*	92.0530	0.0927	0.0242
		EPS-*TLS*	92.0530	0.0967	0.0222
	80%	EPS-*Extreme value*	90.7285	0.1121	0.0158
		EPS-*Logistic*	81.4570	0.0724	0.0507
		EPS-*Normal*	81.4570	0.0706	0.0547
		EPS-*Stable*	81.4570	0.0682	0.0585
		EPS-*TLS*	86.7550	0.0801	0.0375
Site	**PINC**	**Distribution**	**PICP**	**FINAW**	**AWD**
BJ	95%	EPS-*Extreme value*	92.0530	0.1546	0.0110
		EPS-*Logistic*	89.4040	0.1054	0.0199
		EPS-*Normal*	92.0530	0.1177	0.0137
		EPS-*Stable*	96.6887	0.1550	0.0054
		EPS-*TLS*	96.0265	0.1397	0.0065
	90%	EPS-*Extreme value*	92.0530	0.1263	0.0175
		EPS-*Logistic*	85.4305	0.0848	0.0401
		EPS-*Normal*	88.0795	0.0987	0.0258
		EPS-*Stable*	86.0927	0.0900	0.0354
		EPS-*TLS*	91.7951	0.1017	0.0224
	80%	EPS-*Extreme value*	88.7417	0.0958	0.0332
		EPS-*Logistic*	84.1060	0.0632	0.0793
		EPS-*Normal*	85.4305	0.0769	0.0524
		EPS-*Stable*	75.4967	0.0520	0.1179
		EPS-*TLS*	80.7947	0.0522	0.1134

Note: *In the table above, FICP, FINAW, and AWD are selected to verify the prediction performance of different models, where* $FICP = \sum_{i=1}^{N} c_i \times 100\%/N$, $FINAW = \sum_{i=1}^{N} (U_i - L_i) \times 100\%$, *and* $AWD = \sum_{i=1}^{N} AWD_i/N$.

GBiLSTM not only has better prediction accuracy than LSTM but also that its robustness is significantly improved.

Application of the Proposed Ensemble Prediction System

1) A stable carbon price prediction system plays a prominent role in the initial allocation of the carbon quota, in transaction pricing and in effective monitoring of market risk. The proposed EPS system not only shows accurate point prediction performance but also reasonably analyzes and mines the potential uncertainty of carbon prices by constructing the carbon price prediction interval based on error distribution fitting.
2) The proposed EPS system combines a deep learning framework with a traditional neural network and thereby provides a new idea for carbon price prediction and an effective reference tool that policymakers can use to research the volatility of the carbon market.
3) Comparing the EU ETS market data with the carbon price markets in Shenzhen and Beijing, it is helpful for China to analyze the evolution of the mature carbon trading market price in the EU; this will help the regulatory authorities adjust the policy and ensure the steady development of China's carbon market.
4) EPS has high practical value and strong expansibility and can easily fit highly volatile nonlinear time series. It thus provides a new intelligent supervision scheme for building a sound global carbon trading market in the future. At the same time, the use of a deep learning integrated forecasting system with high accuracy and strong stability is expected to become a new direction of energy and financial markets in the future.

Suggestions on Further Improvement of Carbon Price Market

More accurate prediction of carbon prices can provide some effective suggestions through which governments and enterprises can build and improve the carbon price market in the future. These are outlined below.

Improvement of the Initial Allocation Mechanism of Carbon Emission Rights

In the initial allocation of carbon quotas, we should pay attention to the fairness of allocation. First, the government should formulate incentive policies to encourage regional governments and local enterprises to reduce emissions and should give appropriate incentives or policy support to the regions and enterprises that use emission reduction technologies. Second, the initial allocation of carbon emission rights requires effective operation and an effective regulatory system; both of these components directly affect the efficiency and fairness of carbon quota allocation. Strengthening the construction of a carbon emission rights regulatory system will help achieve efficiency and fairness of resource allocation as China's total emission reduction target is being met.

Rationalization of the Carbon Trading Pricing System

Owing to the imperfect development of the carbon trading market and the carbon trading pricing system, the carbon trading price is easily affected by monopoly enterprises. At present, there is a certain monopoly phenomenon in the carbon trading market in some regions of China. Some small buyer enterprises can only passively accept the carbon price, and this allows monopoly enterprises to disproportionately influence the supply and demand of the market and reduces total social welfare. The establishment of a reasonable pricing system that

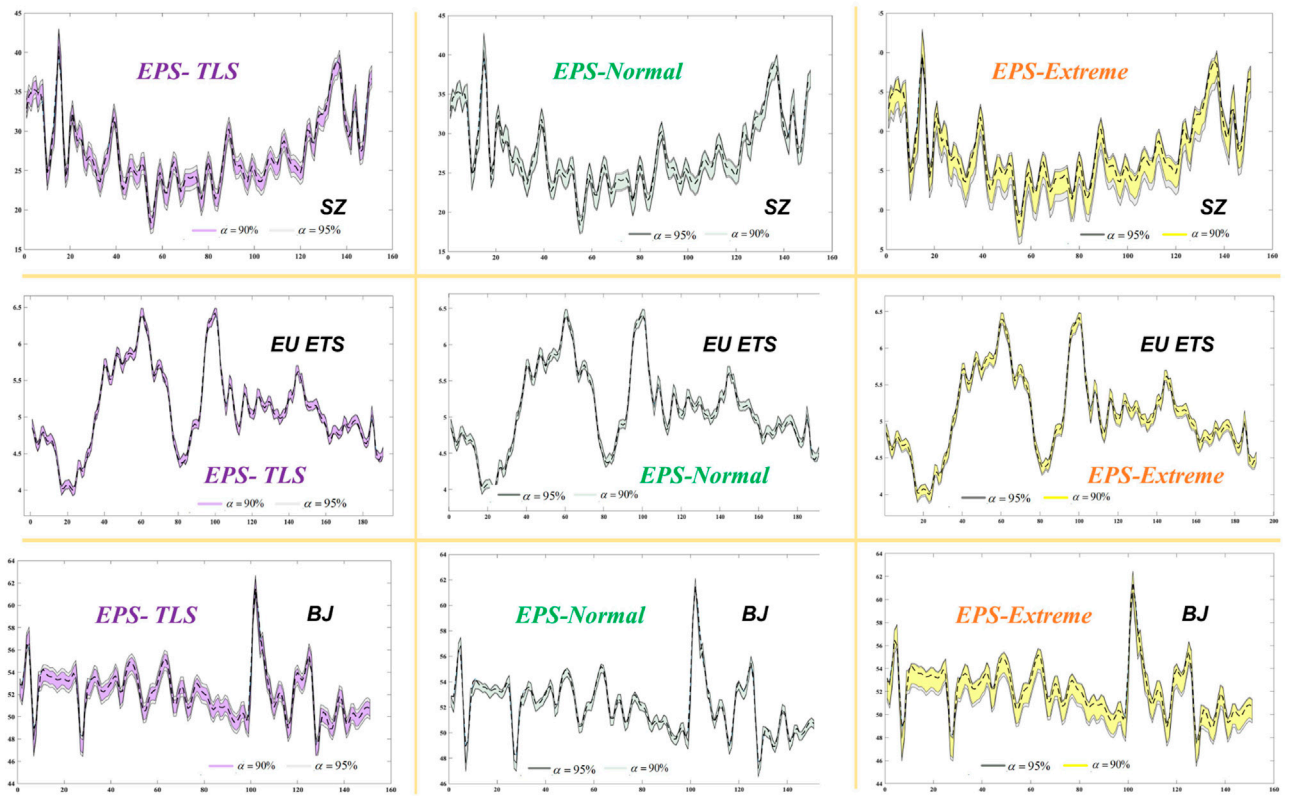


FIGURE 5 | Carbon price prediction intervals generated by the three proposed schemes.

TABLE 11 | Certainty of different forecasting methods.

Model	ICEE-ELM	ICEE-GBiLSTM	ICEE-CNN	EPS	ICEE-LSTM	ICEE-GWO-BP
EU ETS						
$SD_{(MAPE)}$	0.0105	0.0243	0.0902	**0.0101**	0.3277	0.0583
$SD_{(RMSE)}$	0.0004	0.0007	0.0041	**0.0003**	0.0042	0.0007
$SD_{(MAE)}$	**0.0005**	0.0008	0.0044	**0.0005**	0.0033	**0.0005**
$SD_{(IA)}$	**0.0001**	**0.0001**	0.0007	**0.0001**	0.0007	**0.0001**
BJ						
$SD_{(MAPE)}$	0.0255	0.0204	0.0757	**0.0147**	0.0438	0.0171
$SD_{(RMSE)}$	0.0091	0.0117	0.0224	**0.0106**	0.0212	0.0107
$SD_{(MAE)}$	0.0129	0.0123	0.0379	**0.0114**	0.0321	0.0819
$SD_{(IA)}$	**0.0011**	**0.0011**	0.0019	0.0012	0.0033	0.0038
SZ						
$SD_{(MAPE)}$	0.1845	0.1130	0.7846	**0.1124**	0.3831	0.1695
$SD_{(RMSE)}$	**0.0344**	0.1363	0.1745	0.1352	0.1443	0.2029
$SD_{(MAE)}$	0.0448	0.0062	0.1934	**0.0054**	0.0179	0.0348
$SD_{(IA)}$	0.0021	0.0061	0.0076	0.0018	0.0067	**0.0015**

Note: *ICEE is an abbreviation for ICEEMDAN. The best indicator values are shown in bold type.*

avoids monopoly price manipulation is conducive to the return of carbon prices to real value levels and to the optimal allocation of resources.

Improvement of the Carbon Market Risk Management and Control System

In the process of price fluctuation risk control, an accurate and effective price forecasting model can be used to monitor price fluctuations. Using the relevant data, such a model can be used to predict long-term and short-term carbon trading prices, predict future fluctuation trends, and establish an effective carbon trading price risk early warning index system to effectively monitor the volatility risk caused by market price fluctuations. Through prediction of the carbon trading market price, we can grasp the fluctuations in carbon prices and take measures in advance to exercise macro control and reduce the level of risk when large market price fluctuations occur.

CONCLUSION

The availability of a reliable carbon price forecasting system is significant in the emission trading market because it can help decision-makers evaluate climate policies and adjust the emission ceiling to effectively maintain the reliable operation of the market. In this study, the EPS system, which adopts advanced data feature extraction and selection methods, combines the three optimal submodels through a multiobjective dragonfly optimization algorithm and explores the deterministic and uncertainty prediction of carbon price series. This study has several important implications: 1) ICEEMDAN is better than traditional signal decomposition at extracting data features. This can improve the accuracy of the prediction system. 2) The deep learning algorithm has a better ability than other algorithms to forecast carbon price series. The developed GBiLSTM model has better predicted performance and stability than the traditional LSTM. 3) Unlike previous studies in which it was assumed that the prediction error obeys a Gaussian distribution, this study explores five fitting distribution functions of prediction error, finds a more accurate error distribution function, and constructs a more reasonable carbon price prediction interval. The experimental results indicate that the EPS prediction system achieves the best prediction performance, with MAPE values of 1.2657, 4.0156, and 1.0064% for the three datasets. In addition, according to the optimal distribution fitting function of EPS prediction error, the carbon price prediction interval is constructed to mine the uncertainty of carbon price fluctuation. At various significance levels, the constructed prediction interval contains most of the observations, showing that the interval prediction has good performance. Therefore, the system is an effective supplement to the existing carbon price prediction research framework and contributes to the ability of the government to reduce market risk and stabilize the market.

Although the combined prediction system proposed in this article achieves good prediction performance, there are still some limitations due to practical factors. Future research will analyze the carbon price trend from two perspectives, historical carbon price series and external factors, to obtain more accurate and stable prediction results.

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

YY: Conceptualization, Methodology. HG: Writing-Reviewing and Editing, Data curation. YJ: Formal analysis, Supervision. AS: Software, Validation and Visualization, Investigation.

FUNDING

This work was supported by the Major Project of the National Social Science Fund (Grant No. 17ZDA093).

REFERENCES

Arouri, M. E. H., Jawadi, F., and Nguyen, D. K. (2012). Nonlinearities in Carbon Spot-Futures price Relationships during Phase II of the EU ETS. *Econ. Model.* 29 (3), 884–892. doi:10.1016/j.econmod.2011.11.003

Atsalakis, G. S. (2016). Using Computational Intelligence to Forecast Carbon Prices. *Appl. Soft Comput.* 43, 107–116. doi:10.1016/j.asoc.2016.02.029

Benz, E. A., and Trück, S. (2009). Modeling the price Dynamics of CO2 Emission Allowances. *Energ. Econ.* 31, 4–15. doi:10.1016/j.eneco.2008.07.003

Byun, S. J., and Cho, H. (2013). Forecasting Carbon Futures Volatility Using GARCH Models with Energy Volatilities. *Energ. Econ.* 40, 207–221. doi:10.1016/j.eneco.2013.06.017

Chen, C., and Liu, H. (2021). Dynamic Ensemble Wind Speed Prediction Model Based on Hybrid Deep Reinforcement Learning. *Adv. Eng. Inform.* 48, 101290. doi:10.1016/J.AEI.2021.101290

Cheng, Z., and Wang, J. (2020). A New Combined Model Based on Multi-Objective Salp Swarm Optimization for Wind Speed Forecasting. *Appl. Soft Comput.* 92, 106294. doi:10.1016/j.asoc.2020.106294

Chevallier, J. (2009). Carbon Futures and Macroeconomic Risk Factors: A View from the EU ETS. *Energ. Econ.* 31 (4), 614–625. doi:10.1016/j.eneco.2009.02.008

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," in *NIPS 2014 Workshop on Deep Learning*. ArXiv e-prints Available at: https://arxiv.org/abs/1412.3555.

Colominas, M. A., Schlotthauer, G., and Torres, M. E. (2014). Improved Complete Ensemble EMD: a Suitable Tool for Biomedical Signal Processing. *Biomed. Signal Process. Control.* 14, 19–29. doi:10.1016/j.bspc.2014.06.009

Du, P., Wang, J., Yang, W., and Niu, T. (2020). Point and Interval Forecasting for Metal Prices Based on Variational Mode Decomposition and an Optimized Outlier-Robust Extreme Learning Machine. *Resour. Pol.* 69, 101881. doi:10.1016/J.RESOURPOL.2020.101881

Fan, X., Li, S., and Tian, L. (2015). Chaotic Characteristic Identification for Carbon price and an Multi-Layer Perceptron Network Prediction Model. *Expert Syst. Appl.* 42 (8), 3945–3952. doi:10.1016/j.eswa.2014.12.047

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Huang, G-B., Zhu, Q-Y., and Siew, C-K. (2006). Extreme Learning Machine: Theory and Applications. *Neurocomputing* 1. doi:10.1016/j.neucom.2005.12.126

Jiang, H., Tao, C., Dong, Y., and Xiong, R. (2020). Robust Low-Rank Multiple Kernel Learning with Compound Regularization. *Eur. J. Oper. Res.* 295, 634–647. doi:10.1016/J.EJOR.2020.12.024

Jiang, H., Luo, S., and Dong, Y. (2021a). Simultaneous Feature Selection and Clustering Based on Square Root Optimization. *Eur. J. Oper. Res.* 289, 214–231. doi:10.1016/j.ejor.2020.06.045

Jiang, H., Zheng, W., and Dong, Y. (2021b). Sparse and Robust Estimation with ridge Minimax Concave Penalty. *Inf. Sci.* 571, 154–174. doi:10.1016/J.INS.2021.04.047

Jiang, P., Liu, Z., Niu, X., and Zhang, L. (2021). A Combined Forecasting System Based on Statistical Method, Artificial Neural Networks, and Deep Learning Methods for Short-Term Wind Speed Forecasting. *Energy* 217, 119361. doi:10.1016/J.ENERGY.2020.119361

Jin, Y., Guo, H., Wang, J., and Song, A. (2020). A Hybrid System Based on LSTM for Short-Term Power Load Forecasting. *Energies* 13 (23), 6241. doi:10.3390/en13236241

Liu, Z., Jiang, P., Zhang, L., and Niu, X. (2020). A Combined Forecasting Model for Time Series: Application to Short-Term Wind Speed Forecasting. *Appl. Energ.* 259, 114137. doi:10.1016/j.apenergy.2019.114137

Liu, Z., Jiang, P., Wang, J., and Zhang, L. (2021). Ensemble Forecasting System for Short-Term Wind Speed Forecasting Based on Optimal Sub-model Selection

and Multi-Objective Version of Mayfly Optimization Algorithm. *Expert Syst. Appl.* 177, 114974. doi:10.1016/J.ESWA.2021.114974

Lu, H., Ma, X., Huang, K., and Azimi, M. (2020). Carbon Trading Volume and price Forecasting in China Using Multiple Machine Learning Models. *J. Clean. Prod.* 249, 119386. doi:10.1016/j.jclepro.2019.119386

Mirjalili, S. (2016). Dragonfly Algorithm: a New Meta-Heuristic Optimization Technique for Solving Single-Objective, Discrete, and Multi-Objective Problems. *Neural Comput. Applic* 27, 1053–1073. doi:10.1007/s00521-015-1920-1

Niu, T., Wang, J., Lu, H., Yang, W., and Du, P. (2020). Developing a Deep Learning Framework with Two-Stage Feature Selection for Multivariate Financial Time Series Forecasting. *Expert Syst. Appl.* 148, 113237. doi:10.1016/j.eswa.2020.113237

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., and Woo, W. C. (2015). "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in Advances in Neural Information Processing Systems (NIPS 2015). ArXiv e-prints Available at: https://arxiv.org/abs/1506.04214.

Song, J., and Li, S. (2017). Elite Opposition Learning and Exponential Function Steps-Based Dragonfly Algorithm for Global Optimization. *IEEE Int. Conf. Inf. Autom. ICIA*. doi:10.1109/ICInfA.2017.8079080

Song, Y., Qin, S., Qu, J., and Liu, F. (2015). The Forecasting Research of Early Warning Systems for Atmospheric Pollutants: A Case in Yangtze River Delta Region. *Atmos. Environ.* 118, 58–69. doi:10.1016/j.atmosenv.2015.06.032

Sun, Y., Wang, X., Chen, Y., and Liu, Z. (2018). A Modified Whale Optimization Algorithm for Large-Scale Global Optimization Problems. *Expert Syst. Appl.* doi:10.1016/j.eswa.2018.08.027

Tian, C., and Hao, Y. (2020). Point and Interval Forecasting for Carbon price Based on an Improved Analysis-Forecast System. *Appl. Math. Modelling(C).* doi:10.1016/j.apm.2019.10.022

Torres, M. E., Colominas, M. A., Schlotthauer, G., and Flandrin, P. (2011). "A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise," in ICASSP, IEEE International Conference on Acoustics, Speech and signal processing- Proceedings, 4144e7. doi:10.1109/ICASSP.2011.5947265

Wang, J., Niu, T., Du, P., and Yang, W. (2020). Ensemble Probabilistic Prediction Approach for Modeling Uncertainty in Crude Oil price. *Appl. Soft Comput.* 95, 106509. doi:10.1016/j.asoc.2020.106509

Wang, J., Niu, X., Zhang, L., and Lv, M. (2021). Point and Interval Prediction for Non-ferrous Metals Based on a Hybrid Prediction Framework. *Resour. Pol.* 73, 102222. doi:10.1016/J.RESOURPOL.2021.102222

Wang, R., Li, J., Wang, J., and Gao, C. (2018). Research and Application of a Hybrid Wind Energy Forecasting System Based on Data Processing and an Optimized Extreme Learning Machine [J]. *Energies* 11, 1712. doi:10.3390/en11071712

Wang, Y., Zhang, N., Tan, Y., Hong, T., Kirschen, D. S., and Kang, C. (2019). Combining Probabilistic Load Forecasts. *IEEE Trans. Smart Grid* 10, 3664–3674. doi:10.1109/TSG.2018.2833869

Wang, G. (2018). Design of Damage Identification Algorithm for Mechanical Structures Based on Convolutional Neural Network. *Concurrency Computat Pract. Exper* 30, e4891. doi:10.1002/cpe.4891

Wang, J. (2020). OCT Image Recognition of Cardiovascular Vulnerable Plaque Based on CNN. *IEEE Access* 8, 140767–140776. doi:10.1109/access.2020.3007599

Wei, S., Chongchong, Z., and Cuiping, S. (2018). Carbon Pricing Prediction Based on Wavelet Transform and K-ELM Optimized by Bat Optimization Algorithm in China ETS: the Case of Shanghai and Hubei Carbon Markets. *Carbon Manage.* 9 (6), 605–617. doi:10.1080/17583004.2018.1522095

Xiao, L., Qian, F., and Shao, W. (2017). Multi-step Wind Speed Forecasting Based on a Hybrid Forecasting Architecture and an Improved Bat Algorithm. *Energ. Convers. Manage.* 143, 410–430. doi:10.1016/j.enconman.2017.04.012

Zhang, X., Wang, J., and Zhang, K. (2017). Short-term Electric Load Forecasting Based on Singular Spectrum Analysis and Support Vector Machine Optimized by Cuckoo Search Algorithm. *Electric Power Syst. Res.* 146, 270–285. doi:10.1016/j.epsr.2017.01.035

Zhang, X., and Wang, J. (2018). A Novel Decomposition-ensemble Model for Forecasting Short-term Load-time Series with Multiple Seasonal Patterns. *Soft Comput.* 65, 478–494. doi:10.1016/j.asoc.2018.01.017

Zhu, B., and Wei, Y. (2013). Carbon price Forecasting with a Novel Hybrid ARIMA and Least Squares Support Vector Machines Methodology. *Omega* 41, 517–524. doi:10.1016/j.omega.2012.06.005

Zhu, B., Yuan, L., and Ye, S. (2019). Examining the Multi-Timescales of European Carbon Market with Grey Relational Analysis and Empirical Mode Decomposition. *Physica A: Stat. Mech. its Appl.* 517, 392–399. doi:10.1016/j.physa.2018.11.016

Zhu, J., Wu, P., Chen, H., Liu, J., and Zhou, L. (2019). Carbon price Forecasting with Variational Mode Decomposition and Optimal Combined Model. *Physica A: Stat. Mech. its Appl.* 519, 140–158. doi:10.1016/j.physa.2018.12.017

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.

Copyright © 2021 Yang, Guo, Jin and Song. 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.

APPENDIX

APPENDIX A1 | The list of abbreviations in this study.

List of terminologies (method and indices)			
EMD	Empirical model decomposition	**EEMD**	Ensemble empirical mode decomposition
ICEEMDAN	Improved complementary ensemble empirical mode decomposition with adaptive noise		
GRU	Gated recurrent unit	**CNN**	Convolutional neural networks
ELM	Extreme learning machine	**BP**	Back propagation neural network
CDF	Cumulative density function	**ARIMA**	Autoregressive interval moving average model
GWO	Grey wolf optimization algorithm	**MODA**	Multiobjective dragonfly optimization algorithm
GWO-BP	BP after GWO algorithm	**GBiLSTM**	Bidirectional long short-term memory-gated recurrent unit
BiLSTM	Bidirectional long short-term memory	**SSA**	Singular spectrum analysis
FINAW	Forecast interval normalized average width	**FICP**	Forecast interval coverage probability
AWD	Accumulated width deviation of testing dataset	**IMFs**	Intrinsic mode functions
DF	Distribution function	**CDF**	Cumulative distribution function
MAPE	Mean absolute percentage error	**MAE**	Mean absolute error
RMSE	Root mean square error	**IA**	Concordance index
TLS	T location-scale function	**GRNN**	Generalized regression neural network
R2	Coefficient of determination	**DL**	Deep learning probability density function
		PDF	
ANNs	Artificial neural networks	**LSTM**	Long short-term memory

frontiers in Environmental Science

ORIGINAL RESEARCH
published: 18 October 2021
doi: 10.3389/fenvs.2021.747101

An Outlier-Robust Point and Interval Forecasting System for Daily $PM_{2.5}$ Concentration

Ziqi Yin[1] *and Xin Fang*[2]*

[1]Faculty of Information Technology, Macau University of Science and Technology, Macau, China, [2]School of Business, Macau University of Science and Technology, Macau, China

Air pollution forecasting, particularly of $PM_{2.5}$ levels, can be used not only to deliver effective warning information to the public but also to provide support for decisions regarding the control and treatment of air pollution problems. However, there are still some challenging issues in air pollution forecasting that urgently need to be solved, such as how to handle and model outliers, improve forecasting stability, and correct forecasting results. In this context, this study proposes an outlier-robust forecasting system to attempt to tackle the abovementioned issues and bridge the gap in the current research. Specifically, the system developed consists of two parts that deal with point and interval forecasting, respectively. For point forecasting, a data preprocessing module is proposed based on outlier handling and data decomposition to mitigate the negative influences of outliers and noise, which can also help the model capture the main characteristics of the original time series. Meanwhile, an outlier-robust forecasting module is designed for better modeling of the preprocessed data. For the model to further improve its accuracy, a nonlinear correction module based on an error ensemble strategy is developed that can provide more accurate forecasting results. Finally, the interval forecasting part of the system is based on a newly proposed artificial intelligence–based distribution evaluation and the results of the point forecasting part to present the range of future changes. Experimental results and analysis utilizing daily $PM_{2.5}$ concentration from two provincial capital cities in China are discussed to verify the superiority and effectiveness of the system developed, which can be considered an effective technique for point and interval forecasting of daily $PM_{2.5}$ concentration.

Keywords: $PM_{2.5}$ concentration, point forecasting, interval forecasting, outlier handling and modeling, forecasting system

OPEN ACCESS

Edited by:
Pei Du,
Xi'an Jiaotong University, China

Reviewed by:
Song Ding,
Zhejiang University of Finance and Economics, China
Kevin Li,
University of Windsor, Canada

***Correspondence:**
Xin Fang
xfang@must.edu.mo

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: *25 July 2021*
Accepted: *19 August 2021*
Published: *18 October 2021*

Citation:
Yin Z and Fang X (2021) An Outlier-Robust Point and Interval Forecasting System for Daily $PM_{2.5}$ Concentration. Front. Environ. Sci. 9:747101. doi: 10.3389/fenvs.2021.747101

INTRODUCTION

Urbanization, industrialization, and energy consumption have caused the issue of air pollution to become increasingly serious. The air pollution issue is considered a major concern (Andrade et al., 2015) and regarded as the single largest health risk (Wendel, 2014). It can have adverse effects on human beings and bring great economic losses as well as problems for society, affecting areas such as public health (Li et al., 2017), corporate cash holdings (Li et al., 2021), and the tourism industry (Hao et al., 2021). In this context, to solve air pollution issues and accelerate ecological progress, air pollution forecasting, particularly of $PM_{2.5}$ (particulate matter with an aerodynamic diameter of 2.5 μm or less) levels, has been acknowledged as a promising technique for air pollution control and

treatment (Liu et al., 2019a). However, the performance of air pollution forecasting suffers from many factors, and how to develop a model that can improve forecasting effectiveness is not only a challenging research topic but also a matter of growing concern. As a result, developing a more effective model for modeling $PM_{2.5}$ concentration has become an imperative task that cannot be postponed (Yang et al., 2019a).

Over the past few decades, many studies have been conducted to propose an effective air pollution forecasting model. In general, the current air pollution forecasting models mainly belong to three categories, namely, the chemical transport method (CTM), the traditional statistical method, and the artificial intelligence method. In the first category, Timmermans et al. (2017) employed a kind of CTM method to analyze $PM_{2.5}$ in China. Zhang et al. (2019) used the community multiscale air quality (CMAQ) method to analyze $PM_{2.5}$ pollution events in Qingdao, China. Similarly, Fan et al. (2015) employed third-generation air quality modeling system Models-3/CMAQ to analyze air pollution episodes in one region of China. However, the forecasting performance of CTM methods is easily influenced by the scale and quality of the emissions data (Feng et al., 2015). Furthermore, Stern et al. (2008) suggested that due to incomplete descriptions of physical and chemical processes and limited knowledge of pollution sources, CTM models may produce poor forecasting results. In the traditional statistical method category, the widely employed statistical model mainly consists of an autoregressive integrated moving average (ARIMA) and regression method. Vlachogianni et al. (2011) employed a regression model for forecasting nitrogen oxides (NOx) and PM_{10} (particulate matter with an aerodynamic diameter of 10 μm or less). Zafra et al. (2017) developed an ARIMA model using hourly PM_{10} concentration data. However, the traditional statistical model is unable to extract the complex and nonlinear features of pollutant concentration data, which may result in undesirable results (Wang et al., 2020a). Fortunately, with the rapid development of advanced technologies, new methods, including artificial neural networks (ANN), support vector regression (SVR), extreme learning machines (ELM), and other artificial intelligence methods, have been proposed and are being widely employed in different forecasting fields, including air pollution forecasting. These novel methods are acknowledged as promising solutions to air pollution forecasting issues (Yang et al., 2019b).

However, all individual models, including artificial intelligence models, have their advantages and disadvantages. For instance, they may ignore the importance of data preprocessing and optimization and thus be unsuited to meet the requirements of decision-making and management. As a result, to overcome the deficiency of individual artificial intelligence methods and improve air pollution forecasting effectiveness, researchers have started to pay increasing attention to the development of hybrid models by hybridizing multifarious methods from individual forecasting models, mainly including data decomposition techniques (empirical mode decomposition (EMD) (Zhu et al., 2017), ensemble EMD (EEMD) (Bai et al., 2019), complementary EEMD (CEEMD) (Yang and Wang, 2017), fast EEMD (FEEMD) (Luo et al., 2018), complete EEMD with adaptive noise (CEEMDAN) (Hao and Tian, 2019), improved CEEMDAN (ICEEMDAN) (Sharma et al., 2020), variational mode decomposition (VMD) (Wu and Lin, 2019a), wavelet transform (WT) (Cheng et al., 2019), discrete WT (DWT) (Siwek and Osowski, 2012), stationary WT (SWT) (Bai et al., 2016), maximum overlap DWT (MODWT) (Prakash et al., 2011), wavelet packet decomposition (WPD) (Liu et al., 2019b), and empirical WT (EWT) (Liu and Chen, 2020), and so on) and artificial intelligence optimization (whale optimization algorithm (WOA) (Xu et al., 2017), bat algorithm (BA) (Wu and Lin, 2019b), modified grey wolf optimization (MGWO) (Xing et al., 2019), cuckoo search (CS) (Sun and Sun, 2017), multi-objective Harris hawks optimization (MOHHO) (Du et al., 2020), and so on). For example, Jiang et al. (2019) designed a hybrid system based on ICEEMDAN, imperialist competitive algorithm (ICA), and backpropagation neural network (BPNN) for pollutant forecasting. Similarly, Du et al. (2020) devised a hybrid model using ELM, MOHHO, and ICEEMDAN for air quality forecasting. The abovementioned hybrid forecasting studies prove the superiority of hybrid modeling, which has emerged as the most promising research direction in the air pollution forecasting field.

Although many hybrid models have been proposed for air pollution forecasting, there are still some challenging issues that urgently need to be solved. Specifically, as far as we are aware, most previous studies have employed data decomposition to improve forecasting performance while ignoring the significance of handling and modeling outliers in air pollution data, which may lead to the hybrid model being unable to further enhance the forecasting performance. Moreover, artificial intelligence optimization algorithms are used to search for the optimal parameters of methods in a hybrid model, but the forecasting ability can only be improved to a certain degree. Significantly, however, these studies ignore the time-consuming issues caused by incorporating artificial intelligence optimization algorithms, which may be unable to completely overcome the model's limitations, i.e., the instability of the final results. Furthermore, most previous studies have emphasized the contribution of advanced data decomposition and optimization algorithms while ignoring the significance of mining the characteristics of the original air pollution time series and correcting forecasting results to further improve the model's forecasting performance, despite the growing importance of air pollution forecasting performance. Moreover, another issue with air pollution forecasting, especially daily air pollution forecasting, is that it is mainly focused on point forecasting and thus can only provide deterministic information that is insufficient for real application and cannot provide uncertainty information. As far as we know, interval forecasting can make up for the defects of point forecasting, but this method has been ignored by the relevant researchers despite it being a novel research area that is especially deserving of attention.

In this study, to develop an effective model that overcomes the abovementioned limitations of most previous studies, a novel outlier-robust point and interval forecasting system is proposed for forecasting daily $PM_{2.5}$ concentration, which is composed of two parts. The point forecasting part proposes a novel forecasting

model based on a data preprocessing module, an outlier-robust forecasting module, and a nonlinear correction module to obtain future deterministic information, whereas the interval forecasting part is designed based on the newly proposed artificial intelligence–based distribution evaluation and point forecasting results to obtain future uncertainty information. Specifically, first, a data preprocessing module is proposed, which takes into consideration the significance of outlier handling and data decomposition, that can be employed to mitigate the negative effects of outlier and noise information so that the model developed can capture the main features and achieve better performance. Second, an outlier-robust forecasting module is designed, based on outlier-robust ELM (ORELM), to forecast preprocessed data. Next, a nonlinear correction module based on an error ensemble strategy is developed to mine information in the forecasting results and further improve the model's forecasting performance. Finally, an artificial intelligence–based distribution evaluation method is designed in the interval forecasting part, which can be combined with the point forecasting results to provide a range of future changes. Case studies utilizing daily $PM_{2.5}$ concentrations from two cities in China are designed to validate the developed system.

The main novelty and contribution of this study to current research can be summarized as follows:

(1) ***The forecasting focus of the system developed is not only on point forecasting but also on interval forecasting.*** Previous daily $PM_{2.5}$ concentration forecasting studies mainly focused on point forecasting, which can provide deterministic information and is insufficient for real application, while ignoring the research about interval forecasting, which can provide uncertainty information. Therefore, an outlier-robust forecasting system that consists of point and interval forecasting is successfully proposed in this study, which can make up for the defects of point forecasting and is validated well in two cities in China.

(2) ***An improved data preprocessing module is designed to solve the outlier data and noise information issues simultaneously.*** Most previous studies only employed data decomposition for data preprocessing while ignoring the significance of outlier handling and modeling and failed to further enhance the forecasting performance. As a result, this study develops an improved data preprocessing module based on outlier handling and data decomposition that can effectively overcome the limitations caused by outlier and noise information.

(3) ***Point forecasting performance is further enhanced by proposing an outlier-robust forecasting module and a nonlinear correction module.*** In the daily $PM_{2.5}$ concentration forecasting field, outlier modeling and forecasting result correction are of great importance but have been ignored by the relevant researchers. Thus, on the one hand, the outlier-robust forecasting module is designed to further solve the outlier modeling issue; on the other hand, a nonlinear correction module is developed based on an error ensemble strategy to mine information in the forecasting results and further enhance the model's forecasting performance.

(4) ***Convincing experiments are designed to verify the effectiveness and superiority of the system developed.*** For system evaluation, five evaluation metrics are employed in the evaluation of point forecasting, whereas two typical metrics are used in the evaluation of interval forecasting. For model comparison, four experiments are designed to prove the effectiveness and superiority of outlier handling and modeling, data decomposition, nonlinear correction module, and the system developed in daily $PM_{2.5}$ concentration forecasting.

The methods are presented in the *Methodology* section, next the construction of the outlier-robust point and interval forecasting system are discussed, then the *Experimental Analysis* section presents the experiments, and the final section draws the conclusions of this study.

METHODOLOGY

This section introduces the detailed methods used in the system developed, which includes a data preprocessing module, an outlier-robust forecasting module, a nonlinear correction module, artificial intelligence–based distribution evaluation, and interval forecasting theory.

Data Preprocessing Module

The data preprocessing module is proposed on the basis of outlier handling and data decomposition, which can overcome the limitations caused by outlier and noise information.

Outlier Handling

Outliers in the original time series data will have a negative influence on the development of a model, which may bring poor results. Therefore, in this study, an outlier handling algorithm, the Hampel filter (HF), is introduced into the data preprocessing module. HF, developed by Hampel (1974), is robust against outliers (Liu et al., 2004). A brief explanation of HF can be defined as follows (Wu et al., 2021).

Given a sequence $x_1, x_2, \ldots, x_n$, W_i^K is a set of numbers within a moving window and m_i is the median value from the moving window, which are defined as follows:

$$W_i^K = (x_{i-K}, \ldots, x_i, \ldots, x_{i+K}), \tag{1}$$

$$m_i = median(x_{i-K}, \ldots, x_i, \ldots, x_{i+K}), \tag{2}$$

where K is the sliding window's half-width.

The new data obtained after using HF to handle the original sequence can be defined as

$$y_i = \begin{cases} x_i, & |x_i - m_i| \le tS_i \\ m_i, & |x_i - m_i| > tS_i \end{cases}, \tag{3}$$

where t denotes a positive integer and S_i denotes the median absolute deviation (MAD), which can be given by

$$S_i = 1.4826 \times median(|x_{i-K} - m_i|, \ldots, |x_{i+K} - m_i|). \tag{4}$$

Data Decomposition

Data decomposition, which has been acknowledged as a promising data preprocessing technique, has been widely used in forecasting fields to solve complex nonstationary, nonlinear time series forecasting issues. Numerous studies have shown that the data decomposition algorithm has a significant influence on forecasting performance. Thus, a suitable decomposition method should be selected to identify and extract the inner characteristics of the original time series. In previous studies, such as Lin et al. (2017) and Yang et al. (2019a), VMD, proposed by Dragomiretskiy and Zosso (2014), has been shown to be superior to other algorithms, including EMD, EEMD, and CEEMD. Therefore, VMD is introduced into the data preprocessing module to mitigate the negative noise information influence. The main procedure of VMD is as follows.

Step 1: setting parameters of VMD, while each mode $\widehat{y}_k^1$, center pulsation $\widehat{w}_k^1$, and Lagrangian multipliers λ are initialized.

Step 2: y_k and w are updated by

$$\widehat{y}_k^{n+1} = \widehat{f}(w) - \sum_{i \neq k} \frac{\widehat{y}_i(w) + \frac{\widehat{\lambda}(w)}{2}}{1 + 2\alpha(w - w_k)^2}, \tag{5}$$

$$w_k^{n+1} = \frac{\int_0^\infty w\left|\widehat{y}_k^{n+1}(w)\right|^2 \mathrm{d}w}{\int_0^\infty \left|\widehat{y}_k^{n+1}(w)\right|^2 \mathrm{d}w}, \tag{6}$$

where n is the iterations number, $f(t)$ and y_k are the original time series and the kth component, whereas $\widehat{f}(w)$, $\widehat{y}_i(w)$, $\widehat{\lambda}(w)$, and $\widehat{y}_k^{n+1}(w)$ denote the Fourier transforms of $f(t)$, $y_i(t)$, $\lambda(t)$, and $y_k^{n+1}(t)$, respectively.

Step 3: λ can be updated by

$$\widehat{\lambda}^{n+1}(w) \leftarrow \widehat{\lambda}^n(w) + \pi\left[\widehat{f}(w) - \sum_k \widehat{u}_k^{n+1}(w)\right]. \tag{7}$$

Step 4: if $\sum_k \left\|\widehat{y}_k^{n+1} - \widehat{y}_k^n\right\|_2^2 / \left\|\widehat{y}_k^n\right\|_2^2 < e$, the VMD algorithm is stopped, and a series of band-limited modes is returned; otherwise, return to Step 2 to repeat the iteration.

Outlier-Robust Forecasting Module

The basic forecasting model is the important foundation of a hybrid model, which can make a significant difference in forecasting results. If outliers are present within the dataset, the performance of the model developed will be significantly affected. Considering the significance of outlier modeling and outlier robustness, the ORELM model is acknowledged as a potential contributor for modeling data with outliers. Therefore, the ORELM model is introduced into daily $PM_{2.5}$ concentration forecasting to design an outlier-robust forecasting module. The original version of the ORELM model is ELM, developed by Huang et al. (2004), which has many merits, such as its simple structure, better performance, fast computation speed, and the fact that it does not need a large number of samples. Furthermore, previous studies have revealed that ELM methods are superior to some typical ANN methods in solving forecasting issues (Yang et al., 2019a), and it has become one of the most promising approaches.

Given a training dataset with M samples, i.e., (x_t, y_t), $t = 1, \cdots, M$, the ELM model for input data x_t and output data y_t can be presented as

$$\widehat{y} = \sum_{i=1}^{L} \beta_i g_i(x_i) = \sum_{i=1}^{L} \beta_i G(w_i \cdot x_t + b_i,), \tag{8}$$

where L denotes the number of hidden layer nodes, w_i and b_i denote the input weight and hidden bias, G is the excitation function, β_i represents the connected weight between the ith hidden layer node and the output layer, and $\widehat{y}$ represents the forecasting results.

By defining the hidden layer output matrix, i.e., H,

$$H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_1) \end{bmatrix} = \begin{bmatrix} G(w_1 \cdot x_1 + b_1) & \cdots & G(w_L \cdot x_1 + b_L) \\ \vdots & \cdots & \vdots \\ G(w_1 \cdot x_M + b_1) & \cdots & G(w_L \cdot x_M + b_L) \end{bmatrix}_{M \times L}. \tag{9}$$

The ELM model presented in **Eq. 8** can be rewritten as

$$H\beta = Y, \tag{10}$$

where $\beta = [\beta_1 \cdots \beta_L]^T$, $Y = [y_1 \cdots y_M]^T$.

The optimal solution of β can be obtained by solving $\min_\beta = \left\|H\beta - Y\right\|^2$; the corresponding formula is

$$\widehat{\beta} = H^\dagger Y, \tag{11}$$

where $H^\dagger$represents the Moore–Penrose generalized inverse matrix of H; the corresponding formula is

$$H^\dagger = \left[H^T H\right]^{-1} H^T. \tag{12}$$

As mentioned above, to enhance the ELM model's robustness when modeling data with outliers, the ORELM model is developed by Zhang and Luo (2015). The core idea is redefining the minimum problem as

$$\begin{cases} \min\limits_{(\beta)} \|e\|_1 + \dfrac{1}{k}\|\beta\|_2^2 \\ s.t. e = Y - H\beta \end{cases}, \tag{13}$$

where e represents training error and k is the regularization parameter.

To solve the newly defined problem, the augmented Lagrange multiplier (ALM) algorithm is adopted, and the corresponding iteration process is defined as

$$\begin{cases} \beta_{t+1} = \arg\min\limits_{\beta} L_\mu(e_t, \beta, \lambda_t) \\ e_{t+1} = \arg\min\limits_{e} L_\mu(e, \beta_{t+1}, \lambda_t) \\ \lambda_{t+1} = \lambda_t + \mu(Y - H\beta_{t+1} - e_{t+1}) \end{cases}, \tag{14}$$

where λ represents the Lagrange multiplier vector, μ is the penalty parameter, and β_{t+1} and e_{t+1} are defined as

$$\begin{cases} \beta_{t+1} = \left(H^T H + \dfrac{2}{k\mu I}\right)^{-1} H^T\left(Y - e_t + \dfrac{\lambda_t}{\mu}\right) \\ e_{t+1} = shrink\left(Y - H\beta_{t+1} + \dfrac{\lambda_t}{\mu}, \dfrac{1}{\mu}\right) \end{cases}. \tag{15}$$

Nonlinear Correction Module

For a forecasting model, forecasting error is inevitable, but the short-term trend in the variation of the forecasting error can be anticipated by establishing a nonlinear model (Vukicevic, 1991). To further improve the performance of the system developed, a nonlinear correction module based on an error ensemble strategy is developed to mine information in the forecasting results, which is composed of three steps.

Step 1: generating the error sequence

Defining the actual value of the t th datum as $A(t)$ and forecasting the value of the t th datum as $F(t)$, the forecasting error value of the t th datum can be obtained by

$$E(t) = F(t) - A(t). \tag{16}$$

Step 2: developing the error forecasting model

Defining the error value of the $t–d$ th datum as $E(t–d)$, according to the detailed error sequence, the forecasting model can be developed and denoted as f. The error forecasting value of the t th datum, named $EF(t)$, can be obtained by

$$EF(t) = f[E(t-1), E(t-2), ..., E(t-d)]. \tag{17}$$

Step 3: correcting the forecasting results

To obtain a final result on the basis of the original forecasting results and corresponding error forecasting results, an error ensemble strategy based on ORELM is proposed, which fully exploits the advantages of ORELM and is equipped with outlier robustness. By developing an outlier-robust ensemble model, denoted as *En.f*, the final forecasting results of the t th datum, i.e., $FF(t)$, are

$$FF(t) = En.f[EF(t), F(t)]. \tag{18}$$

Artificial Intelligence–Based Distribution Evaluation

Distribution evaluation plays a vital role in many fields, such as wind energy evaluation, time series analysis, and interval forecasting. In recent years, in order to further mine data characteristics, researchers have focused on applying different distribution functions to fit the experimental data and obtain a suitable distribution; then, the interval forecasting results can be obtained according to the interval forecasting theory and point forecasting results. However, to the best of our knowledge, the related research is well validated in many fields, but so far, few studies have involved research on or application to daily $PM_{2.5}$ concentration forecasting. In this context, four typical distributions, i.e., Weibull, Gamma, Rayleigh, and Lognormal, are introduced in this study to fit the daily $PM_{2.5}$ concentration data. In general, the goodness of fit ($0 \le R^2 \le 1$) was employed to measure the fitting performance of one distribution. Traditionally, the maximum likelihood estimation (MLE) method is used to estimate the distribution function's parameters. However, the MLE method may not obtain the optimal distribution parameters. To the best of our knowledge, the larger the R^2 value, the more optimal the distribution. As a result, the optimal distribution determination problem can be converted into solving the maximum value problem. Inspired by Wang et al. (2020b), Schwarz et al. (2020), and Ließ et al. (2021), artificial intelligence optimization can be considered a promising technique for searching for the optimal distribution parameters. Based on this idea and considering the limitations of the traditional method, the artificial intelligence–based distribution evaluation is proposed to obtain the optimal distribution in this study. In order to obtain the optimal distribution, specifically, an advanced optimization algorithm named grey wolf optimizer (GWO) is adopted to search for the optimal parameters of specific distribution by maximizing the values of R^2. In this study, the minus R^2 is defined as the objective function of GWO-based distribution evaluation. Finally, the distribution with the best R^2 value among all distributions is selected as the optimal distribution of $PM_{2.5}$, which can be combined with interval forecasting theory to achieve interval forecasting.

Interval Forecasting Theory

Given the significance level α, actual value A_t, and lower and upper limits (L, U), the probability can be given by

$$P(L \le A_t \le U) = 1 - 2\alpha. \tag{19}$$

For a random variable time series, **Eq. 19** can be rewritten as

$$P(L \le A_t \le U) = P(L \le A_t \le U | E(A_t) = \hat{a}) \times P(E(A_t) = \hat{a}). \tag{20}$$

Supposing that the forecasting value has a similar distribution function, the estimated variance can be determined, and then the following conditional probability formula can be obtained as

$$\begin{cases} P(L \le A_t \le U | E(A_t) = \hat{a}) = \int_L^{\hat{a}} f(z|\Theta)dz + \int_{\hat{a}}^U f(z|\Theta)dz \\ \int z f(z|\Theta)dz = \hat{a} \\ \int (z - E(z))^2 f(z|\Theta)dz = S^2. \end{cases} \tag{21}$$

The lower and upper limits can be obtained by

$$\left\{ \left(\hat{L}, \hat{U}\right) \middle| \hat{L} \le A_t \le \hat{U}, \int_{\hat{a}}^{\hat{U}} f(z|\Theta)dz + \int_{\hat{L}}^{\hat{a}} f(z|\Theta)dz = 1 - 2\alpha \right\}. \tag{22}$$

CONSTRUCTION OF THE OUTLIER-ROBUST FORECASTING SYSTEM

The outlier-robust point and interval forecasting system is constructed in this section; the details of the system design and system evaluation are as follows.

System Design

The system design is composed of the point forecasting part and the interval forecasting part, which can provide deterministic

information and uncertainty information in the future, respectively.

Point Forecasting

The point forecasting part is developed based on three modules, i.e., a data preprocessing module, an outlier-robust forecasting module, and a nonlinear correction module, and consists of three phases as follows.

- **Phase I: data preprocessing**. Considering the negative influence of outlier data and noise information in original daily $PM_{2.5}$ concentration data, a data preprocessing module is designed based on outlier handling and data decomposition that can effectively solve the outlier data and noise information issues simultaneously. Specifically, on the one hand, the HF algorithm is employed to detect and correct the outliers in the original data, which can eliminate the outliers' negative influence on the model's development from the perspective of improving data quality. On the other hand, the advanced data decomposition method named VMD is performed to decompose the processed data into a number of modes, which can eliminate the noise's negative effect on the model's performance from the perspective of signal denoising and helps the model effectively capture the main features of the daily $PM_{2.5}$ concentration data.
- **Phase II: forecasting preprocessed data**. In this phase, some modes obtained in Phase I are forecasted, and then, the forecasting results can be obtained. As a result, it is necessary to select a suitable model for preprocessed data forecasting. Most previous studies tended to develop an optimized ANN model for mode forecasting while ignoring the instability parameter setting problem of artificial intelligence optimization algorithms, which not only cannot guarantee the forecasting model's stability but also bring time-consuming issues. In this context, the ORELM model, with its simple structure, high forecasting performance, and fast calculation speed, better forecasting ability than some typical ANN models, and better outlier robustness than the original ELM model, is selected to design the outlier-robust forecasting module to obtain the future value of each mode, and finally, by summing the forecasting results of each mode, the results of daily ahead $PM_{2.5}$ concentration forecasting can be achieved.
- **Phase III: forecasting results correction**. Most previous studies have focused on using data decomposition and optimization to enhance forecasting ability while ignoring the significance of forecasting results correction. Therefore, in this phase, a nonlinear correction module based on an error ensemble strategy is designed to mine information in the forecasting results and thereby enhance the model's performance, which is composed of three steps, i.e., generating an error sequence, developing an error forecasting model, and correcting the forecasting results. Following these three steps, the results of the point forecasting can be obtained; these results can provide deterministic information in the future.

Interval Forecasting

The interval forecasting part is developed based on the previous point forecasting results, artificial intelligence–based distribution evaluation, and interval forecasting theory and consists of two phases as follows.

TABLE 1 | Three typical metrics.

Abbreviation	Full name	Equation
MAE	Mean absolute error	$\mathbf{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert F_i - A_i\rvert$
RMSE	Root mean square error	$\mathbf{RMSE} = \sqrt{\frac{1}{N}\times\sum_{i=1}^{N}(F_i - A_i)^2}$
MAPE	Root mean square error	$\mathbf{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left\lvert\frac{A_i - F_i}{A_i}\right\rvert \times 100\%$

F_i and A_i denote the forecasting value and actual value, respectively, at time i, *and* N *is the length of the time series.*

- **Phase I: distribution evaluation**. Distribution evaluation is the crucial basis of the interval forecasting part, whereas the traditional MLE method may not fit the optimal distribution for a specific $PM_{2.5}$ concentration dataset. In order to solve this issue, an artificial intelligence–based distribution evaluation is proposed to obtain the optimal distribution. Specifically, the advanced optimization algorithm named GWO is selected to search the parameters of four typical distributions, i.e., Weibull, Gamma, Rayleigh, and Lognormal. Finally, the distribution function with the best R^2 is considered the optimal distribution.
- **Phase II: obtaining interval forecasting results**. According to the point forecasting results, the interval forecasting results are estimated using the optimal distribution determined in Phase I and interval forecasting theory, which can provide uncertainty information in the future.

System Evaluation

This section is designed to provide system evaluation metrics, including point forecasting evaluation and interval forecasting evaluation.

Point Forecasting Evaluation

To evaluate the forecasting performance of the system developed for daily $PM_{2.5}$ concentration point forecasting, the three typical metrics listed in **Table 1** are selected in this study.

In addition to evaluating the forecasting accuracy using these three typical evaluation metrics, to measure the similarity of the forecasting value curve and actual value curve, grey relational analysis (GRA) (Wang et al., 2015) is introduced into point forecasting evaluation, which provides a new metric named grey relational degree (GRD) for further analysis. Moreover, forecasting stability is another important metric in practical application. Thus, a metric named variance ratio (VR) (Yang and Wang, 2017) is introduced into the point forecasting evaluation, which can measure the forecasting stability of different models by considering the variance between the actual and the forecast values.

Interval Forecasting Evaluation

To evaluate the interval forecasting performance, two widely used metrics named forecasting interval coverage probability (FICP) and forecasting interval normalized average width (FINAW) are introduced into the interval forecasting evaluation. Specifically, the FICP indicator is selected to measure the probability that the

actual observation value falls into the forecasting interval, which can reflect the reliability of the interval forecasting results. The larger the FICP value, the higher the interval forecasting accuracy. Furthermore, to avoid the excessive pursuit of the reliability of interval forecasting while ignoring the effective measure of interval forecasting width for interval forecasting results, FINAW is employed to express the width of interval forecasting results. The FICP and FINAW can be calculated by

$$FICP = \frac{1}{N}\sum_{i=1}^{N} c_i \times 100\%, \tag{23}$$

where $c_i = \begin{cases} 1, & A_i \in [L_i, U_i] \\ 0, & A_i \notin [L_i, U_i] \end{cases}$ and N is the length of the time series,

$$FINAW = \frac{1}{NR}\sum_{i=1}^{N} (U_i - L_i), \tag{24}$$

where U_i and L_i denote the forecasting interval's upper and lower limits, respectively, at time i, and R is the range of actual values.

EXPERIMENTAL ANALYSIS

The experiments and analysis are presented in this section, which is mainly composed of data description, experiment design, and four detailed experiments.

Data Description

To validate the ability of the outlier-robust system developed to perform point forecasting and interval forecasting of daily $PM_{2.5}$ pollution, Jinan and Zhengzhou are considered as the study areas; two datasets collected from these two study areas are used as illustrative empirical studies in this study. Jinan, the capital city of Shandong Province, is located in the middle of China. Zhengzhou, the capital city of Henan Province, is located in the middle part of the Yellow River. Specifically, two daily $PM_{2.5}$ concentration datasets, covering 1 yr from July 1, 2017, to June 30, 2018, are employed in this study. In the experiment, the data, from July 1, 2017, to May 31, 2018, are employed as training data for the development of the proposed system, whereas the data from June 1, 2018, to June 30, 2018, are considered as testing data to test the forecasting performance of the system developed.

Experiment Design

As mentioned above, in this study, an outlier-robust point and interval forecasting system is developed, which is composed of a data preprocessing module, an outlier-robust forecasting module, a nonlinear correction module, artificial intelligence–based distribution evaluation, and interval forecasting theory to obtain future deterministic information and uncertainty information about daily $PM_{2.5}$ pollution. To verify the forecasting superiority of the system developed, sufficient empirical research should be carried out. In addition to comparing the performance of the system developed with that of the other types of forecasting models, the contribution of each component proposed or employed in the system developed should also be proved by designing appropriate comparative studies. For this purpose, this study designs four experiments to conduct a convincing evaluation of the system developed. Specifically, in Experiment I, the effectiveness of outlier handling and modeling in the system developed is verified from the perspectives of data preprocessing and model selection. In Experiment II, the effectiveness of data decomposition in the system proposed is compared with other decomposition algorithm–based models and a model without a decomposition preprocess. In Experiment III, a nonlinear correction module is developed to correct the forecasting results, which is designed to compare the proposed system with the model without correcting the process and the model with a simple error-addition strategy. It should be noted that the experiments for each model in Experiments I–III are carried out 100 times in this study, and the average values of the forecasting results are considered the final forecasting results for practical application and model comparison, which can ensure that the system developed is more reliable, accurate, and independent of random factors to some extent. In Experiment IV, different distributions of daily $PM_{2.5}$ concentration are compared to obtain the optimal distribution, and the interval forecasting results based on point forecasting are obtained and evaluated by two typical metrics.

Experiment I: The Effectiveness of Outlier Handling and Modeling

To evaluate the effectiveness of outlier handling and modeling, eight models, i.e., ELM, regularized ELM (RELM), weighted RELM (WRELM), ORELM, HF-ELM, HF-RELM, HF-WRELM, and HF-ORELM, are proposed and tested. The MAE, RMSE, MAPE, VR, and GRD values of these eight models are shown in **Table 2**. Meanwhile, the results of the different models in the two cities are depicted in **Figures 1**, **2**, which indicate that the ORELM model is superior to ELM, RELM, and WRELM, whereas the HF-ORELM model is superior to the seven other models. As shown in **Table 2**, two types of comparison can be designed based on these eight models. Comparison I compares the forecasting results of the ELM (HF-ELM), RELM (HF-RELM), WRELM (HF-WRELM), and ORELM (HF-ORELM) models. Meanwhile, Comparison II compares the forecasting results of the ORELM and HF-ORELM models (or ELM and HF-ELM, or RELM and HF-RELM, or WRELM and HF-WRELM). In other words, transverse comparison and longitudinal comparison can be conducted according to the metric values in **Table 2**. The detailed comparisons are as follows:

1) In Comparison I, by comparing the ORELM (HF-ORELM) with ELM (HF-ELM), RELM (HF-RELM), and WRELM (HF-WRELM), it can be observed that the ORELM model is superior to the ELM, RELM, and WRELM models, whereas the HF-ORELM model is superior to the HF-ELM, HF-RELM, and HF-WRELM models. For example, for daily $PM_{2.5}$ concentration forecasting in Jinan, the ORELM model has a lower MAPE value of 28.6266%, compared to the MAPE values of 31.5740%, 31.8239%, and 30.9697% for the ELM, RELM, and WRELM models, respectively.

TABLE 2 | Forecasting results of four individual models and HF-based models.

Model	MAE ($\mu g/m^3$)	RMSE ($\mu g/m^3$)	MAPE (%)	VR (–)	GRD (–)	Model	MAE ($\mu g/m^3$)	RMSE ($\mu g/m^3$)	MAPE (%)	VR (–)	GRD (–)
Site: Jinan											
ELM	7.3582	10.0165	31.5740	0.3709	0.6769	HF-ELM	6.9967	9.4329	30.1351	0.3342	0.6712
RELM	7.4260	10.0446	31.8239	0.3649	0.6729	HF-RELM	6.9388	9.3872	29.9125	0.3385	0.6726
WRELM	7.2311	9.8402	30.9697	0.3761	0.6733	HF-WRELM	6.7626	9.2276	29.0110	0.3529	0.6693
ORELM	6.8231	9.2232	28.6266	0.3514	0.6560	HF-ORELM	6.4068	8.2988	26.1079	0.4033	0.6598
Model	**MAE ($\mu g/m^3$)**	**RMSE ($\mu g/m^3$)**	**MAPE (%)**	**VR (–)**	**GRD (–)**	**Model**	**MAE ($\mu g/m^3$)**	**RMSE ($\mu g/m^3$)**	**MAPE (%)**	**VR (–)**	**GRD (–)**
Site: Zhengzhou											
ELM	8.8045	11.0005	33.2161	0.4607	0.6804	HF-ELM	8.2973	10.6228	30.4850	0.5829	0.6689
RELM	8.7766	10.9252	33.0157	0.4571	0.6778	HF-RELM	8.3243	10.6302	30.5535	0.5942	0.6672
WRELM	8.6697	10.7735	32.3432	0.5422	0.6836	HF-WRELM	7.9918	10.1581	28.8852	0.6194	0.6714
ORELM	8.0171	9.7210	28.4607	0.5481	0.6977	HF-ORELM	7.6229	9.6868	25.9379	0.7379	0.6804

(–) indicates no measurement unit.

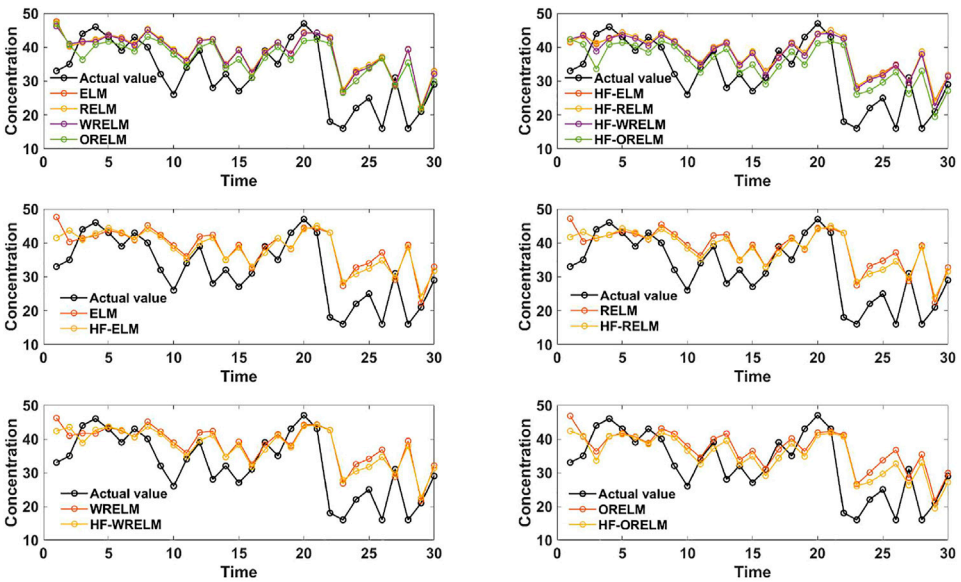


FIGURE 1 | Forecasting results of different comparative studies in Jinan.

Furthermore, for daily $PM_{2.5}$ concentration forecasting in Zhengzhou, the HF-ORELM model achieves the best MAPE value of 25.9379% compared to the MAPE values of 30.4850%, 30.5535%, and 28.8852% for the HF-ELM, HF-RELM, and HF-WRELM models, respectively. The differences in the model forecasting results compared illustrate that the ORELM model is more powerful and robust than the other models for daily $PM_{2.5}$ concentration forecasting. Therefore, we can reasonably conclude that the ORELM model will make a great contribution to the final successful modeling; therefore, it can be selected as the basic forecasting model for the outlier-robust forecasting system.

2) In Comparison II, the forecasting results of models with outlier handling and those of models without outlier handling can be compared to evaluate the effectiveness of outlier handling in the proposed data preprocessing module. By comparing the forecasting results of the ORELM and HF-ORELM model (or ELM and HF-ELM, or RELM and HF-RELM, or WRELM and HF-WRELM), it can be observed that the HF-based model is superior to the models without outlier handling. For example, the MAPE values of HF-ELM, HF-RELM, HF-WRELM, and HF-ORELM for daily $PM_{2.5}$ concentration forecasting in Jinan are 30.1351%, 29.9125%, 29.0110%, and 26.1079%, whereas the corresponding models without outlier handling have larger MAPE values, i.e., ELM (31.5740%), RELM (31.8239%), WRELM (30.9697%), and ORELM (28.6266%). The differences between the forecasting results of the ORELM and HF-ORELM models, ELM and HF-ELM, RELM and HF-RELM, WRELM and HF-WRELM reveal that outlier handling is of great importance for daily $PM_{2.5}$ concentration forecasting, which can be combined with data decomposition to design the data preprocessing module to further improve forecasting performance.

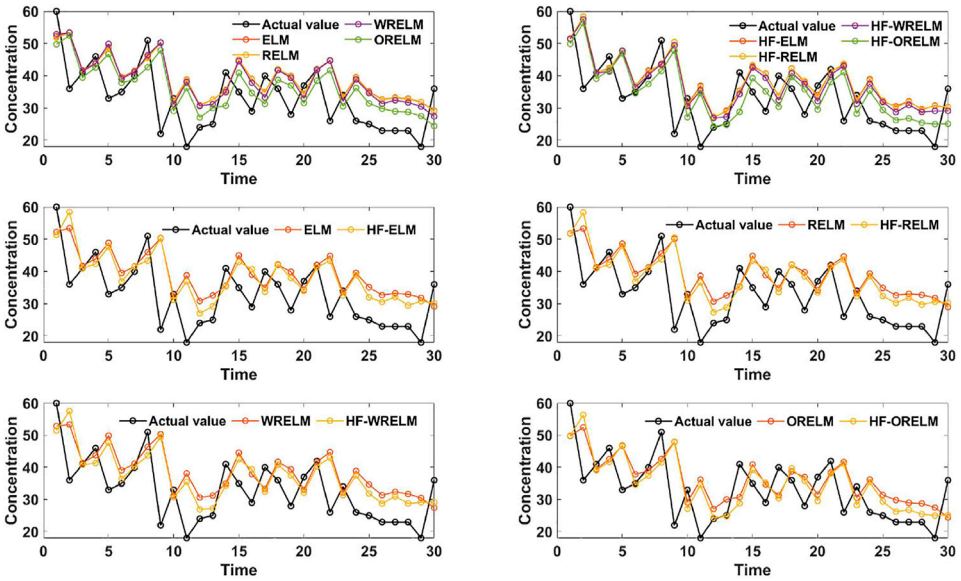


FIGURE 2 | Forecasting results of different comparative studies in Zhengzhou.

Summary: by taking Zhengzhou as an example, the improvement percentage values of MAPE between the different models are employed to summarize the contribution and effectiveness of outlier handling and modeling in this study. The detailed results are 8.2222% (HF-ELM vs ELM), 7.4577% (HF-RELM vs RELM), 10.6916% (HF-WRELM vs WRELM), 8.8642% (HF-ORELM vs ORELM), 14.3166% (ORELM vs ELM), 13.7965% (ORELM vs RELM), and 12.0041% (ORELM vs WRELM). It can be concluded that the HF algorithm and ORELM model are suitable for outlier handling and modeling, which make a great contribution to the success of the system developed in this study.

Experiment II: The Effectiveness of Data Decomposition

To verify the contribution of data decomposition in the proposed data preprocessing module and the superiority of the forecasting results of the outlier-robust forecasting module developed, four models, i.e., HF-ORELM, HF-EMD-ORELM-S, HF-EEMD-ORELM-S, and HF-VMD-ORELM-S, are developed and compared in Jinan and Zhengzhou. In detail, the HF-EMD-ORELM-S, HF-EEMD-ORELM-S, and HF-VMD-ORELM-S employ different data decomposition algorithms to decompose the data after outlier handling into some modes, and the simple addition way is used to add all modes' forecasting results to obtain the daily $PM_{2.5}$ concentration forecasting results. The MAE, RMSE, MAPE, VR, and GRD values of HF-ORELM, HF-EMD-ORELM-S, HF-EEMD-ORELM-S, and HF-VMD-ORELM-S are shown in **Table 3**. Moreover, the forecasting results of these four models in the two cities are shown in **Figure 3**, which indicates that the HF-VMD-ORELM-S model is superior to the original HF-ORELM model and the EMD- or EEMD-based HF-ORELM model. In this experiment, two comparisons can be designed as follows:

1) Comparison I is proposed to validate the superiority of the data decomposition algorithm in the system developed by comparing the HF-VMD-ORELM-S with other decomposition method–based forecasting models, i.e., HF-EMD-ORELM-S and HF-EEMD-ORELM-S. It can be observed that the HF-EMD-ORELM-S model obtains worse forecasting performance compared with the EEMD- and VMD-based models, whereas the VMD-based model achieves better forecasting performance compared with the EMD- and EEMD-based models. For example, for daily $PM_{2.5}$ concentration forecasting in Zhengzhou, the MAE, RMSE, MAPE, VR, and GRD values of HF-VMD-ORELM-S are 1.1259, 1.5228, 3.8169%, 0.9523, and 0.9222, respectively, whereas the metric values of HF-EMD-ORELM-S are 4.2089, 5.9923, 14.1646%, 0.9024, and 0.7705, and the values of HF-EEMD-ORELM-S are 2.1140, 2.7875, 7.0637%, 0.8161, and 0.8652. It is obvious that there are significant differences in the forecasting power of these three models, which further demonstrates the significance of selecting a suitable data decomposition algorithm for the data preprocessing module and the system developed. Therefore, in this study, the VMD algorithm is combined with outlier handling to design the data preprocessing module, which also makes great contributions to the success of the system developed.
2) Comparison II is designed to verify the superiority of the outlier-robust forecasting module developed in daily $PM_{2.5}$ concentration forecasting. In Experiment I, the superiority of HF-ORELM over ORELM has been proven, which means the outlier handling is an indispensable part of the forecasting system developed. Against this background, there is no need to compare the proposed outlier-robust forecasting module's results with those of the individual ORELM model. As a result,

TABLE 3 | Forecasting results of HF-ORELM and HF-ORELM with different data decompositions.

Model	MAE (μg/m^3)	RMSE (μg/m^3)	MAPE (%)	VR (–)	GRD (–)
Site: Jinan					
HF-ORELM	6.4068	8.2988	26.1079	0.4033	0.6598
HF-EMD-ORELM-S	4.2089	5.9923	14.1646	0.9024	0.7705
HF-EEMD-ORELM-S	2.1140	2.7875	7.0637	0.8161	0.8652
HF-VMD-ORELM-S	1.1259	1.5228	3.8169	0.9523	0.9222
Model	**MAE (μg/m^3)**	**RMSE (μg/m^3)**	**MAPE (%)**	**VR (–)**	**GRD (–)**
Site: Zhengzhou					
HF-ORELM	7.6229	9.6868	25.9379	0.7379	0.6804
HF-EMD-ORELM-S	5.5281	6.4669	18.1868	0.9350	0.7335
HF-EEMD-ORELM-S	2.6744	3.6666	8.3559	0.7184	0.8596
HF-VMD-ORELM-S	1.4782	2.7491	5.0868	0.9365	0.9151

(–) indicates no measurement unit.

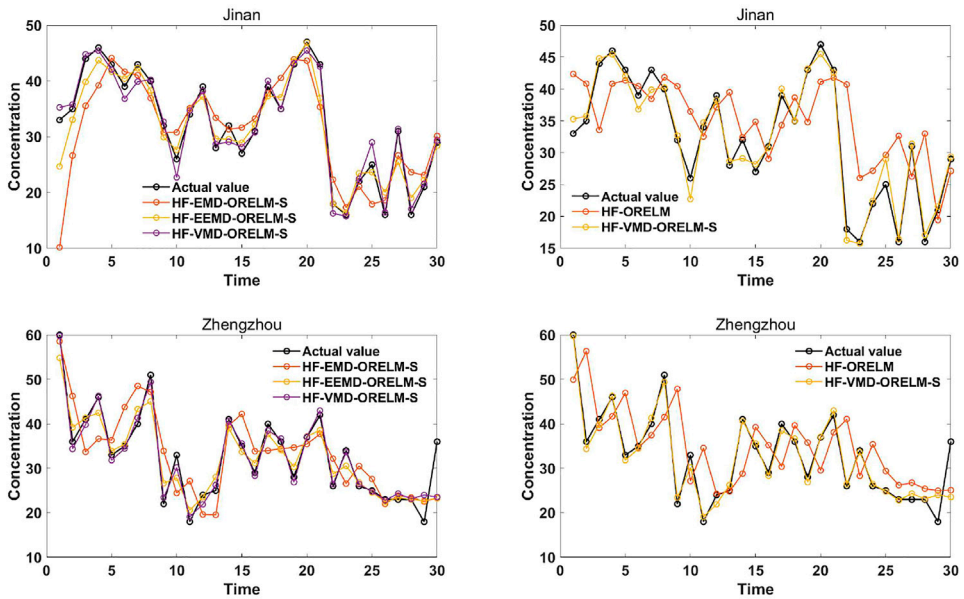


FIGURE 3 | Forecasting results of HF-ORELM with/without data decomposition.

in Comparison II, the outlier-robust forecasting module is compared with the HF-ORELM model to prove the contribution of the VMD algorithm. From **Table 3**, it can be observed that the HF-VMD-ORELM-S model performs better than the HF-ORELM model in terms of MAE, RMSE, MAPE, VR, and GRD. For example, in comparison with the HF-ORELM model, the proposed HF-VMD-ORELM-S model presents an improvement from (6.4068, 8.2988, 26.1079%, 0.4033, and 0.6598) to (1.1259, 1.5228, 3.8169%, 0.9523, and 0.9222) in terms of MAE, RMSE, MAPE, VR, and GRD in Jinan. The results prove that the forecasting ability of the HF-VMD-ORELM-S model is superior to the benchmark model, which can provide a better basis for the success of the system developed.

Summary: by taking Jinan as an example, the improvement percentage values of MAPE between different models are employed to summarize the contribution and effectiveness of data decomposition in this study. The detailed results are 85.3803% (HF-VMD-ORELM-S vs HF-ORELM), 73.0532% (HF-VMD-ORELM-S vs HF-EMD-ORELM-S), 45.9646% (HF-VMD-ORELM-S vs HF-EEMD-ORELM-S), 45.7459% (HF-EMD-ORELM-S vs HF-ORELM), and 72.9442% (HF-EEMD-ORELM-S vs HF-ORELM). It can be concluded that the VMD algorithm is superior to the EMD and EEMD algorithms and is a promising technique for daily $PM_{2.5}$ concentration decomposition, which can also make a great contribution to the success of the system developed.

Experiment III: The Effectiveness of the Nonlinear Correction Module

As mentioned above, the third module, named the nonlinear correction module, is proposed to correct the results of the

TABLE 4 | Metrics of the system developed and compared models.

Model	MAE (μg/m^3)	RMSE (μg/m^3)	MAPE (%)	VR (–)	GRD (–)
Site: Jinan					
HF-VMD-ORELM-S	1.1259	1.5228	3.8169	0.9523	0.9222
HF-VMD-ORELM+Error	1.1472	1.5203	3.9048	0.9529	0.9224
HF-VMD-ORELM+EnError	1.0744	1.5525	3.5736	0.9648	0.9228
Model	**MAE (μg/m^3)**	**RMSE (μg/m^3)**	**MAPE (%)**	**VR (–)**	**GRD (–)**
Site: Zhengzhou					
HF-VMD-ORELM-S	1.4782	2.7491	5.0868	0.9365	0.9151
HF-VMD-ORELM+Error	1.4252	2.6989	5.0681	0.9350	0.9192
HF-VMD-ORELM+EnError	1.3841	2.7134	4.8328	0.9439	0.9193

(–) indicates no measurement unit.

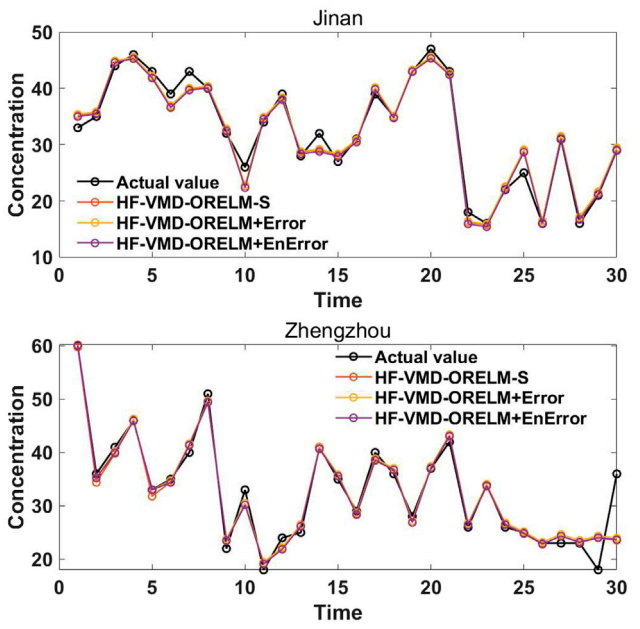


FIGURE 4 | Forecasting results of the system developed and compared models in Jinan and Zhengzhou.

outlier-robust forecasting module to further improve the daily $PM_{2.5}$ concentration forecasting performance. To prove the superiority and effectiveness of the proposed nonlinear correction module and the system developed for point forecasting, the performance of the point forecasting part developed, i.e., HF-VMD-ORELM+EnError, is compared with HF-VMD-ORELM-S and HF-VMD-ORELM+Error in this section. In detail, the HF-VMD-ORELM-S model without a correcting process is the best model in Experiment II, which can provide the results of the devised forecasting module, whereas the HF-VMD-ORELM+Error model is a model with a simple error-addition strategy. The MAE, RMSE, MAPE, VR, and GRD values of the system developed, HF-VMD-ORELM-S, and HF-VMD-ORELM+Error are listed in **Table 4**; meanwhile, the results of these three models are shown in **Figure 4**. Based on Experiment III, the following conclusions can be obtained:

1) The HF-VMD-ORELM+Error model performs better than HF-VMD-ORELM-S model in Jinan but performs worse than HF-VMD-ORELM+Error in Zhengzhou, which indicates that the simple error-addition strategy cannot guarantee the effectiveness of error correction. Therefore, how to correct the forecasting results is a challenging issue in forecasting fields. In other words, the method of correcting forecasting results plays a vital role in the success of the system developed. In this context, a nonlinear correction module based on an error ensemble strategy is presented to further improve the model's forecasting performance.
2) By comparing the HF-VMD-ORELM+EnError model and the HF-VMD-ORELM+Error model, it can be found that the main difference between these two models is the forecasting results correcting method. As shown in **Table 4**, the forecasting results obtained by the system developed are better than those obtained by the HF-VMD-ORELM+Error model, with the MAE, RMSE, MAPE, VR, and GRD values in Jinan and Zhengzhou being (1.0744, 1.5525, 3.5736%, 0.9648, and 0.9228), and (1.3841, 2.7134, 4.8328%, 0.9439, and 0.9193), respectively. Therefore, the HF-VMD-ORELM+EnError model performs better than the HF-VMD-ORELM+Error model, which proves the superiority of the nonlinear correction module based on an error ensemble strategy.
3) The contribution of the proposed nonlinear correction module and the superiority of the system developed can be measured by comparing the point forecasting results of the system developed with those of the HF-VMD-ORELM-S model. As shown in **Table 4**, in comparison with the HF-VMD-ORELM-S model, the proposed system presents an improvement from (1.4782, 2.7491, 5.0868%, 0.9365, and 0.9151) to (1.3841, 2.7134, 4.8328%, 0.9439, and 0.9193) in terms of MAE, RMSE, MAPE, VR, and GRD in Zhengzhou. In previous experiments, the superiority of the HF-VMD-ORELM-S model over other benchmark models has been proven. Therefore, considering the system developed performs better than HF-VMD-ORELM-S, we can conclude that the system designed in this study is superior to other models and can be widely employed in daily $PM_{2.5}$ concentration forecasting.

Summary: by taking the MAPE metric as an example, the improvement percentage values between different models are employed to summarize the contribution and effectiveness of

TABLE 5 | Distribution parameters and R^2 provided by MLE and GWO.

Site	Method	Weibull			Gamma			Rayleigh		Lognormal		
		λ	k	R^2	θ	k	R^2	σ	R^2	μ	σ	R^2
Jinan	MLE	64.9598	1.7146	0.9708	17.6411	3.2567	0.9839	48.2826	0.9432	3.8896	0.5554	0.9956
	GWO	59.1176	1.8952	0.9879	16.8538	3.1787	0.9936	41.7205	0.9873	3.8587	0.5689	0.9976
Site	**Method**	**Weibull**			**Gamma**			**Rayleigh**		**Lognormal**		
		λ	k	R^2	θ	k	R^2	σ	R^2	μ	σ	R^2
Zhengzhou	MLE	71.8142	1.4689	0.9325	25.1562	2.5516	0.9494	57.8859	0.7784	3.9533	0.6057	0.9818
	GWO	57.3570	1.5556	0.9779	23.2795	2.2547	0.9818	39.7569	0.9666	3.8587	0.5703	0.9879

the nonlinear correction module in this study. The detailed results for Jinan are 6.3743% (HF-VMD-ORELM+EnError vs HF-VMD-ORELM-S), 8.4819% (HF-VMD-ORELM+EnError vs HF-VMD-ORELM+Error), and −2.3029% (HF-VMD-ORELM+Error vs HF-VMD-ORELM-S), whereas the values for Zhengzhou are 4.9933% (HF-VMD-ORELM+EnError vs HF-VMD-ORELM-S), 4.6428% (HF-VMD-ORELM+EnError vs HF-VMD-ORELM+Error), and 0.3676% (HF-VMD-ORELM+Error vs HF-VMD-ORELM-S). It can be concluded that the proposed nonlinear correction module is not only effective for improving the final forecasting results but also is superior to the HF-VMD-ORELM+Error model. Furthermore, the HF-VMD-ORELM+Error model may perform worse than the HF-VMD-ORELM-S model. In other words, the proposed nonlinear correction module is suitable for correcting forecasting results, which can contribute to improving the performance of the system developed.

Experiment IV: Interval Forecasting

In the system developed, the interval forecasting can be achieved by the proposed interval forecasting part according to the results of the point forecasting part, artificial intelligence–based distribution evaluation, and interval forecasting theory. In order to perform the interval forecasting, a distribution evaluation of daily $PM_{2.5}$ concentration data is conducted. As mentioned above, the traditional MLE method may not obtain the optimal distribution function for a specific $PM_{2.5}$ concentration dataset, whereas artificial intelligence optimization is a powerful technique for determining the optimal distribution. Therefore, in this study, the advanced optimization algorithm named GWO is selected to fit four typical distributions, i.e., Weibull, Gamma, Rayleigh, and Lognormal. In order to prove the superiority of GWO, detailed distribution is also determined by MLE, and the parameters and R^2 values provided by MLE and GWO are presented in **Table 5**. Furthermore, the comparison is also depicted in **Figure 5**.

On the basis of **Table 5** and **Figure 5**, we find that the GWO-based distribution evaluation can obtain the best R^2 values for each distribution function, which indicates that the GWO-based distribution evaluation is superior to the MLE method and is suitable for fitting the detailed distribution. As a result, the results of the GWO-based distribution evaluation can be compared by R^2. As shown in **Table 5**, the R^2 values of Weibull, Gamma, Rayleigh, and Lognormal are (0.9879, 0.9936, 0.9873, and 0.9976) and (0.9779, 0.9818, 0.9666, and 0.9879) in Jinan and Zhengzhou, respectively. It can be observed that the Lognormal distribution achieves the largest R^2, which means that the Lognormal distribution can effectively fit the daily $PM_{2.5}$ concentration data in Jinan and Zhengzhou. Thus, the optimal Lognormal distribution obtained can be combined with the point forecasting results and interval forecasting theory to achieve the final interval forecasting.

The interval forecasting results under different significance levels are depicted in **Figure 6**, and the corresponding evaluation metric values are listed in **Table 6**. From **Table 6**, we can find that the interval forecasting performances for Jinan and Zhengzhou are different at the same significance level. For example, when α = 0.30, the FINAW and FICP values for Jinan and Zhengzhou are (0.6268, 100.0000%) and (0.4695, 96.6667%), respectively. The main reasons for this phenomenon are that the interval forecasting performance largely depends on the point forecasting performance. As the system developed has achieved excellent point forecasting performance, it has also achieved ideal interval forecasting results. Moreover, for the same dataset, the FINAW and FICP values under five significance levels are different. For instance, for the Zhengzhou dataset, the FINAW and FICP values for α = 0.20 and α = 0.25 are (0.7711, 100.0000%) and (0.6097, 96.6667%), respectively. Furthermore, the solid lines represent the actual values, and the shaded areas represent the forecasting intervals in **Figure 6**. Obviously, as most of the observations fall into the shaded area, the interval forecasting ability of the system established can be considered effective and good. According to the abovementioned analysis and discussion, we can reasonably conclude that the system developed can be a promising tool for daily $PM_{2.5}$ concentration interval forecasting.

CONCLUSION

Forecasting air pollution is not only a challenging research topic but also a growing concern. To model and forecast the complex $PM_{2.5}$ concentration time series, in this study, a novel outlier-robust point and interval forecasting system is developed, which attempts to mitigate or solve some of the challenges in current studies. In the system developed, the point forecasting part is designed to provide future deterministic information on daily $PM_{2.5}$ concentration, whereas the interval forecasting part is devised to present future uncertainty information. More specifically, three modules, named the data preprocessing

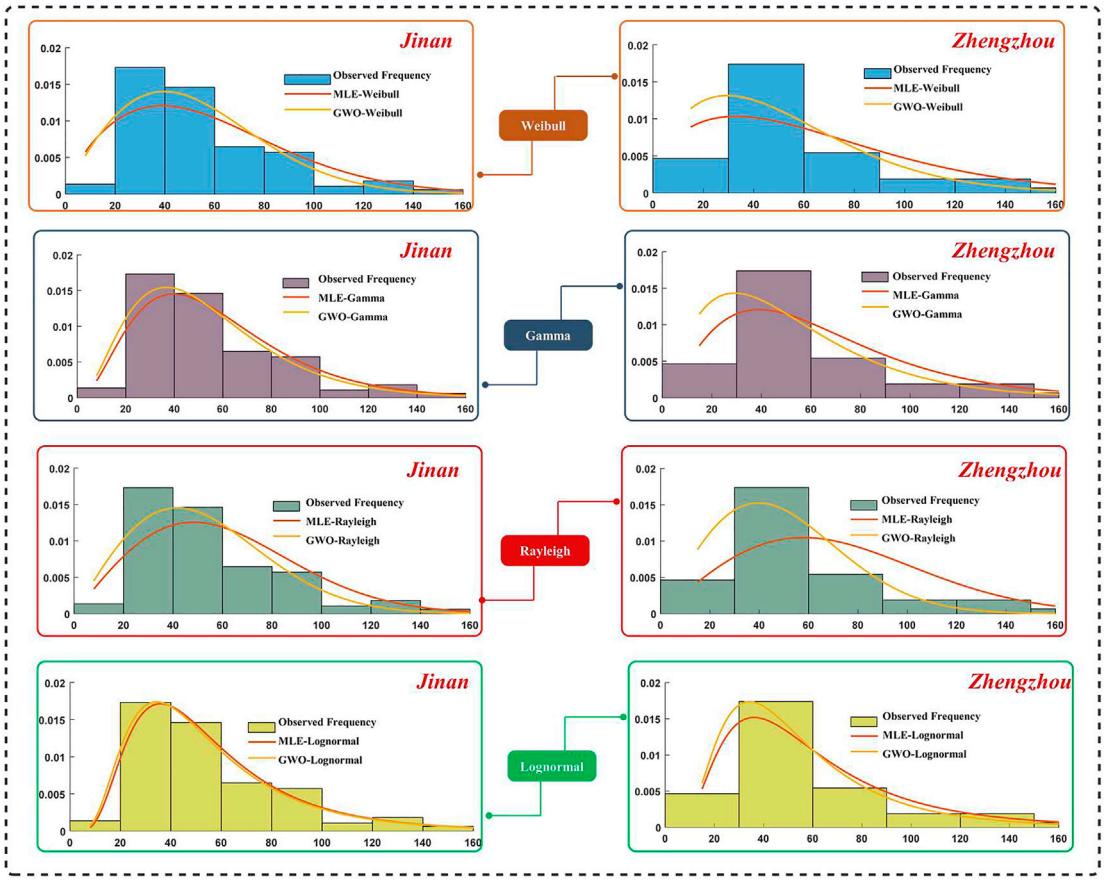


FIGURE 5 | Distribution results provided by MLE and GWO.

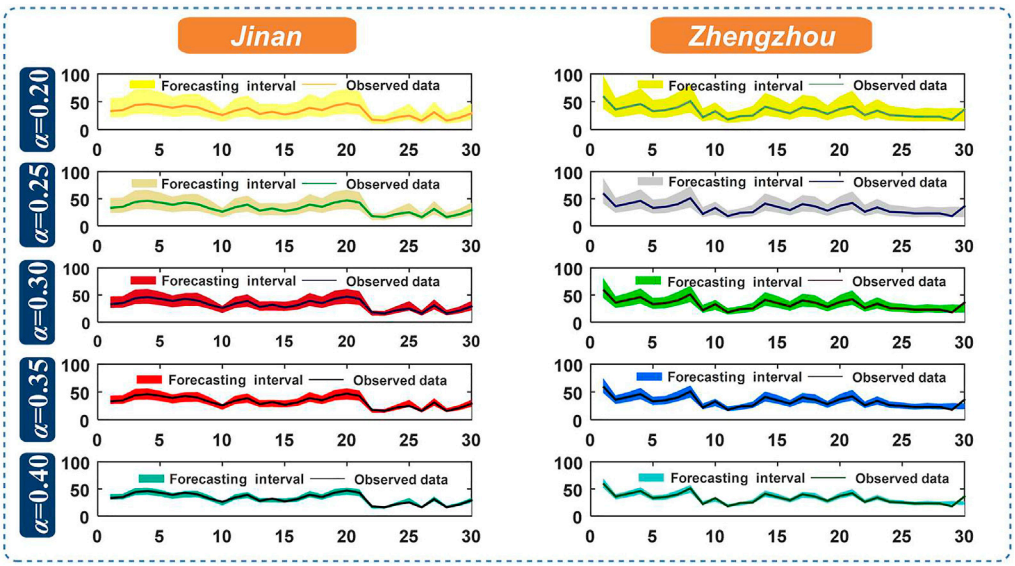


FIGURE 6 | Results of interval forecasting in Jinan and Zhengzhou.

TABLE 6 | Results of interval forecasting based on different significance levels.

α	Jinan		Zhengzhou	
	FINAW (–)	FICP (%)	FINAW (–)	FICP (%)
0.20	1.0295	100.0000	0.7711	100.0000
0.25	0.8140	100.0000	0.6097	96.6667
0.30	0.6268	100.0000	0.4695	96.6667
0.35	0.4574	100.0000	0.3426	93.3333
0.40	0.2994	96.6667	0.2243	93.3333

(–) indicates no measurement unit.

module, the outlier-robust forecasting module, and the nonlinear correction module, are proposed in the point forecasting part. The data preprocessing module, considering the negative influence of outliers and noise on the development of the model, is designed on the basis of outlier handling and data decomposition for the purpose of solving the outlier data and noise information issues simultaneously. Moreover, in order to obtain a forecasting model with outlier robustness, the ORELM model with superior performance is selected to design the outlier-robust forecasting module to forecast each mode. Furthermore, the nonlinear correction module is developed based on an error ensemble strategy, which can mine information in the forecasting results and further improve the model's forecasting performance. Afterward, the interval forecasting part is developed based on artificial intelligence–based distribution evaluation and interval forecasting theory, which can be incorporated with the point forecasting results to obtain the range of future changes. The experimental results illustrate that the system developed can not only perform better than other compared models in point forecasting but also provide uncertainty information in the future. Moreover, outside the field of daily $PM_{2.5}$ concentration forecasting, the system developed can also be employed to solve other challenging issues, including energy forecasting, economic forecasting, and financial forecasting.

Although the system developed shows better performance in daily $PM_{2.5}$ concentration forecasting, there are still some limitations that must be considered in future research. For example, other features neglected in this study may make the forecasting more reliable and practical and can further improve the forecasting performance, which can be a promising subject for future studies. Moreover, only 1-day-ahead forecasting is conducted by the system developed; how to achieve multi-day-ahead forecasting is of great importance and worth studying but is still a challenging task. As a result, further studies about multi-day-ahead point and interval forecasting for $PM_{2.5}$ concentration can be considered an important research direction in future studies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: China Air Quality Online Monitoring and Analysis platform (https://www.aqistudy.cn).

AUTHOR CONTRIBUTIONS

ZY and XF conducted the research.

ACKNOWLEDGMENTS

The authors are grateful to the reviewers and the editor for their constructive comments and suggestions for this paper.

REFERENCES

Andrade, M. d. F., Ynoue, R. Y., Freitas, E. D., Todesco, E., Vara Vela, A., Ibarra, S., et al. (2015). Air Quality Forecasting System for Southeastern Brazil. *Front. Environ. Sci.* 3, 9. doi:10.3389/fenvs.2015.00009

Bai, Y., Li, Y., Wang, X., Xie, J., and Li, C. (2016). Air Pollutants Concentrations Forecasting Using Back Propagation Neural Network Based on Wavelet Decomposition with Meteorological Conditions. *Atmos. Pollut. Res.* 7, 557–566. doi:10.1016/j.apr.2016.01.004

Bai, Y., Zeng, B., Li, C., and Zhang, J. (2019). An Ensemble Long Short-Term Memory Neural Network for Hourly $PM_{2.5}$ Concentration Forecasting. *Chemosphere* 222, 286–294. doi:10.1016/j.chemosphere.2019.01.121

Cheng, Y., Zhang, H., Liu, Z., Chen, L., and Wang, P. (2019). Hybrid Algorithm for Short-Term Forecasting of $PM_{2.5}$ in China. *Atmos. Environ.* 200, 264–279. doi:10.1016/j.atmosenv.2018.12.025

Dragomiretskiy, K., and Zosso, D. (2014). Variational Mode Decomposition. *IEEE Trans. Signal. Process.* 62, 531–544. doi:10.1109/TSP.2013.2288675

Du, P., Wang, J., Hao, Y., Niu, T., and Yang, W. (2020). A Novel Hybrid Model Based on Multi-Objective Harris Hawks Optimization Algorithm for Daily $PM_{2.5}$ and PM_{10} Forecasting. *Appl. Soft Comput.* 96, 106620. doi:10.1016/j.asoc.2020.106620

Fan, Q., Lan, J., Liu, Y., Wang, X., Chan, P., Hong, Y., et al. (2015). Process Analysis of Regional Aerosol Pollution During Spring in the Pearl River Delta Region, China. *Atmos. Environ.* 122, 829–838. doi:10.1016/j.atmosenv.2015.09.013

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., and Wang, J. (2015). Artificial Neural Networks Forecasting of $PM_{2.5}$ Pollution Using Air Mass Trajectory Based Geographic Model and Wavelet Transformation. *Atmos. Environ.* 107, 118–128. doi:10.1016/j.atmosenv.2015.02.030

Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation. *J. Am. Stat. Assoc.* 69, 383–393. doi:10.1080/01621459.1974.10482962

Hao, Y., Niu, X., and Wang, J. (2021). Impacts of Haze Pollution on China's Tourism Industry: A System of Economic Loss Analysis. *J. Environ. Manage.* 295, 113051. doi:10.1016/j.jenvman.2021.113051

Hao, Y., and Tian, C. (2019). The Study and Application of a Novel Hybrid System for Air Quality Early-Warning. *Appl. Soft Comput.* 74, 729–746. doi:10.1016/j.asoc.2018.09.005

Huang, G. B., Zhu, Q. Y., and Siew, C. K. (2004). Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks. IEEE International Conference on Neural Networks. doi:10.1109/IJCNN.2004.1380068

Jiang, P., Li, C., Li, R., and Yang, H. (2019). An Innovative Hybrid Air Pollution Early-Warning System Based on Pollutants Forecasting and Extenics Evaluation. *Knowledge-Based Syst.* 164, 174–192. doi:10.1016/j.knosys.2018.10.036

Li, B., He, M., Gao, F., and Zeng, Y. (2021). The Impact of Air Pollution on Corporate Cash Holdings. *Borsa Istanbul Rev.* 21, S90–S98. doi:10.1016/j.bir.2021.04.007

Li, L., Lei, Y., Wu, S., Chen, J., and Yan, D. (2017). The Health Economic Loss of Fine Particulate Matter (PM 2.5) in Beijing. *J. Clean. Prod.* 161, 1153–1161. doi:10.1016/j.jclepro.2017.05.029

Ließ, M., Gebauer, A., and Don, A. (2021). Machine Learning with GA Optimization to Model the Agricultural Soil-Landscape of Germany: An Approach Involving Soil Functional Types with Their Multivariate Parameter Distributions Along the Depth Profile. *Front. Environ. Sci.* 9, 692959. doi:10.3389/fenvs.2021.692959

Lin, Y., Luo, H., Wang, D., Guo, H., and Zhu, K. (2017). An Ensemble Model Based on Machine Learning Methods and Data Preprocessing for Short-Term Electric Load Forecasting. *Energies* 10, 1186. doi:10.3390/en10081186

Liu, H., and Chen, C. (2020). Prediction of Outdoor $PM_{2.5}$ Concentrations Based on a Three-Stage Hybrid Neural Network Model. *Atmos. Pollut. Res.* 11, 469–481. doi:10.1016/j.apr.2019.11.019

Liu, H., Jin, K., and Duan, Z. (2019b). Air $PM_{2.5}$ Concentration Multi-step Forecasting Using a New Hybrid Modeling Method: Comparing Cases for Four Cities in China. *Atmos. Pollut. Res.* 10, 1588–1600. doi:10.1016/j.apr.2019.05.007

Liu, H., Shah, S., and Jiang, W. (2004). On-line Outlier Detection and Data Cleaning. *Comput. Chem. Eng.* 28, 1635–1647. doi:10.1016/j.compchemeng.2004.01.009

Liu, H., Xu, Y., and Chen, C. (2019a). Improved Pollution Forecasting Hybrid Algorithms Based on the Ensemble Method. *Appl. Math. Model.* 73, 473–486. doi:10.1016/j.apm.2019.04.032

Luo, H., Wang, D., Yue, C., Liu, Y., and Guo, H. (2018). Research and Application of a Novel Hybrid Decomposition-Ensemble Learning Paradigm with Error Correction for Daily PM 10 Forecasting. *Atmos. Res.* 201, 34–45. doi:10.1016/j.atmosres.2017.10.009

Prakash, A., Kumar, U., Kumar, K., and Jain, V. K. (2011). A Wavelet-Based Neural Network Model to Predict Ambient Air Pollutants' Concentration. *Environ. Model. Assess.* 16, 503–517. doi:10.1007/s10666-011-9270-6

Schwarz, N., Hoffmann, F., Knapp, S., and Strauch, M. (2020). Synergies or Trade-Offs? Optimizing a Virtual Urban Region to foster Plant Species Richness, Climate Regulation, and Compactness Under Varying Landscape Composition. *Front. Environ. Sci.* 8, 16. doi:10.3389/fenvs.2020.00016

Sharma, E., Deo, R. C., Prasad, R., and Parisi, A. V. (2020). A Hybrid Air Quality Early-Warning Framework: An Hourly Forecasting Model with Online Sequential Extreme Learning Machines and Empirical Mode Decomposition Algorithms. *Sci. Total Environ.* 709, 135934. doi:10.1016/j.scitotenv.2019.135934

Siwek, K., and Osowski, S. (2012). Improving the Accuracy of Prediction of PM10 Pollution by the Wavelet Transformation and an Ensemble of Neural Predictors. *Eng. Appl. Artif. Intelligence* 25, 1246–1258. doi:10.1016/j.engappai.2011.10.013

Stern, R., Builtjes, P., Schaap, M., Timmermans, R., Vautard, R., Hodzic, A., et al. (2008). A Model Inter-comparison Study Focussing on Episodes with Elevated PM10 Concentrations. *Atmos. Environ.* 42, 4567–4588. doi:10.1016/j.atmosenv.2008.01.068

Sun, W., and Sun, J. (2017). Daily PM 2.5 Concentration Prediction Based on Principal Component Analysis and LSSVM Optimized by Cuckoo Search Algorithm. *J. Environ. Manage.* 188, 144–152. doi:10.1016/j.jenvman.2016.12.011

Timmermans, R., Kranenburg, R., Manders, A., Hendriks, C., Segers, A., Dammers, E., et al. (2017). Source Apportionment of $PM_{2.5}$ Across China Using LOTOS-EUROS. *Atmos. Environ.* 164, 370–386. doi:10.1016/j.atmosenv.2017.06.003

Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., and Kukkonen, J. (2011). Evaluation of a Multiple Regression Model for the Forecasting of the Concentrations of NOx and PM10 in Athens and Helsinki. *Sci. Total Environ.* 409, 1559–1571. doi:10.1016/j.scitotenv.2010.12.040

Vukicevic, T. (1991). Nonlinear and Linear Evolution of Initial Forecast Errors. *Mon. Weather Rev.* 119, 1602–1611. doi:10.1175/1520-0493(1991)119<1602:NALEOI>2.0.CO

Wang, J.-Z., Wang, Y., and Jiang, P. (2015). The Study and Application of a Novel Hybrid Forecasting Model - A Case Study of Wind Speed Forecasting in China. *Appl. Energ.* 143, 472–488. doi:10.1016/j.apenergy.2015.01.038

Wang, J., Du, P., Hao, Y., Ma, X., Niu, T., and Yang, W. (2020a). An Innovative Hybrid Model Based on Outlier Detection and Correction Algorithm and Heuristic Intelligent Optimization Algorithm for Daily Air Quality index Forecasting. *J. Environ. Manage.* 255, 109855. doi:10.1016/j.jenvman.2019.109855

Wang, J., Yang, W., Du, P., and Niu, T. (2020b). Outlier-robust Hybrid Electricity price Forecasting Model for Electricity Market Management. *J. Clean. Prod.* 249, 119318. doi:10.1016/j.jclepro.2019.119318

Wendel, J. (2014). Air Pollution Ranks as Largest Health Risk. *Eos Trans. AGU* 95, 120. doi:10.1002/2014eo140007

Wu, D.-C., Bahrami Asl, B., Razban, A., and Chen, J. (2021). Air Compressor Load Forecasting Using Artificial Neural Network. *Expert Syst. Appl.* 168, 114209. doi:10.1016/j.eswa.2020.114209

Wu, Q., and Lin, H. (2019b). A Novel Optimal-Hybrid Model for Daily Air Quality Index Prediction Considering Air Pollutant Factors. *Sci. Total Environ.* 683, 808–821. doi:10.1016/j.scitotenv.2019.05.288

Wu, Q., and Lin, H. (2019a). Daily Urban Air Quality index Forecasting Based on Variational Mode Decomposition, Sample Entropy and LSTM Neural Network. *Sustain. Cities Soc.* 50, 101657. doi:10.1016/j.scs.2019.101657

Xing, Y., Yue, J., Chen, C., Xiang, Y., Chen, Y., and Shi, M. (2019). A Deep Belief Network Combined with Modified Grey Wolf Optimization Algorithm for $PM_{2.5}$ Concentration Prediction. *Appl. Sci.* 9, 3765. doi:10.3390/app9183765

Xu, Y., Yang, W., and Wang, J. (2017). Air Quality Early-Warning System for Cities in China. *Atmos. Environ.* 148, 239–257. doi:10.1016/j.atmosenv.2016.10.046

Yang, W., Wang, J., Lu, H., Niu, T., and Du, P. (2019a). Hybrid Wind Energy Forecasting and Analysis System Based on Divide and Conquer Scheme: A Case Study in China. *J. Clean. Prod.* 222, 942–959. doi:10.1016/j.jclepro.2019.03.036

Yang, W., Wang, J., Niu, T., and Du, P. (2019b). A Hybrid Forecasting System Based on a Dual Decomposition Strategy and Multi-Objective Optimization for Electricity Price Forecasting. *Appl. Energ.* 235, 1205–1225. doi:10.1016/j.apenergy.2018.11.034

Yang, Z., and Wang, J. (2017). A New Air Quality Monitoring and Early Warning System: Air Quality Assessment and Air Pollutant Concentration Prediction. *Environ. Res.* 158, 105–117. doi:10.1016/j.envres.2017.06.002

Zafra, C., Ángel, Y., and Torres, E. (2017). ARIMA Analysis of the Effect of Land Surface Coverage on PM 10 Concentrations in a High-Altitude Megacity. *Atmos. Pollut. Res.* 8, 660–668. doi:10.1016/j.apr.2017.01.002

Zhang, K., and Luo, M. (2015). Outlier-robust Extreme Learning Machine for Regression Problems. *Neurocomputing* 151, 1519–1527. doi:10.1016/j.neucom.2014.09.022

Zhang, Q., Xue, D., Liu, X., Gong, X., and Gao, H. (2019). Process Analysis of $PM_{2.5}$ Pollution Events in a Coastal City of China Using CMAQ. *J. Environ. Sci.* 79, 225–238. doi:10.1016/j.jes.2018.09.007

Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., and Che, J. (2017). Daily Air Quality Index Forecasting with Hybrid Models: A Case in China. *Environ. Pollut.* 231, 1232–1244. doi:10.1016/j.envpol.2017.08.069

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.

Copyright © 2021 Yin and Fang. 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 in Environmental Science

ORIGINAL RESEARCH
published: 22 October 2021
doi: 10.3389/fenvs.2021.761058

Comprehensive Evaluation of Water Carrying Capacity in Hebei Province, China on Principal Component Analysis

Huadong Wang[1,2], *Yunhong Xu*[1,3], *Rini Suryati Sulong*[3], *HuiLing Ma*[4] *and Lifeng Wu*[1]*

[1]School of Management Engineering and Business, Hebei University of Engineering, Hebei, Handan, [2]School of Business, Chosun University, Gwangju, South Korea, [3]Faculty of Business, Economics and Accountancy, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia, [4]School of Landscape and Ecological Engineering, Hebei University of Engineering, Hebei, Handan

Water is an important basic resource for social and economic development and also a necessity for the life and produce of people. The unbalanced development of water resources in Hebei Province of China and the obvious contradiction between supply and demand, affected by geography and natural environment change, has seriously influenced the Hebei village renewal process. This paper presents a comprehensive evaluation method of water carrying capacity—principal component analysis (PCA)—and constructs the evaluation index system of water carrying capacity in Hebei Province from water resources, water management, industrial development, agricultural development, social development, environmental protection, and other aspects. Based on the economic and water statistical data of Hebei province from 2009 to 2018, this paper adopts principal component analysis as an evaluation method to comprehensively evaluate the carrying capacity of water resources in Hebei Province across time and space. The results show that principal component analysis is an effective method for the comprehensive evaluation of water carrying capacity, which can reflect the local water carrying capacity objectively and comprehensively.

Keywords: principal component analysis, water carrying capacity, comprehensive evaluation, hebei province, natural environment change

OPEN ACCESS

Edited by:
Wendong Yang,
Shandong University of Finance and Economics, China

Reviewed by:
Yaoyao He,
Hefei University of Technology, China
Wang Jiefang,
North China University of Water Conservancy and Electric Power, China
Yingjie Yang,
De Montfort University, United Kingdom

***Correspondence:**
Lifeng Wu
wlf6666@126.com

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: *19 August 2021*
Accepted: *24 September 2021*
Published: *22 October 2021*

Citation:
Wang H, Xu Y, Suryati Sulong R, Ma H and Wu L (2021) Comprehensive Evaluation of Water Carrying Capacity in Hebei Province, China on Principal Component Analysis. Front. Environ. Sci. 9:761058. doi: 10.3389/fenvs.2021.761058

INTRODUCTION

Water conservancy plays an irreplaceable supporting role in implementing the rural revitalization strategy. Water is an important basic resource for social and economic development as well as a necessity for the life and produce of people. As an important component of natural resources, water resources are the key objects of ecological protection. How to measure the carrying capacity of water resources scientifically, how to predict the carrying capacity of water resources in advance, and how to provide warning for the imminent danger of water resources, are the key issues in the research of water carrying capacity.

Rural revitalization will eventually achieve comprehensive revitalization. Water plays an important role in the rural revitalization of the Beijing-Tianjin-Hebei region. Affected by the region and natural environment, the water resources in Hebei Province of China are obviously not the same in Beijing and Tianjin. In terms of quantity and structure of water, there are different degrees of quantity shortage or structure imbalance, which seriously affects the process of rural

revitalization in Hebei Province. Water carrying capacity provides a new thinking and methods for the comprehensive utilization of water resources in Hebei Province. How to effectively evaluate the carrying capacity of regional water resources, reasonably develop the utilization potential of water resources, and achieve the coordinated development of the social economy and ecological environment is an important issue to be quickly solved in Hebei Province.

The study of water carrying capacity is helpful for the relevant departments to formulate water protection policies and control risks, contribute to the rational utilization of water resources, and provide water resource guarantees for rural revitalization and development in Hebei Province. The research on the carrying capacity of water resources at home and abroad is reviewed, its concept and characteristics is given, its evaluation methods are summarized and the theory of principal component analysis is elaborated upon in detail in this paper. It is an important part of examining the carrying capacity of water resources to establish its evaluation indexes. 16 evaluation indexes of water carrying capacity in Hebei province are established from the aspects of water, water management, industrial development, agricultural development, social development, and environmental protection on the basis of its influencing factors in the paper. The principal component analysis method is adopted to comprehensively evaluate the carrying capacity of the water resources of Hebei Province over time and space, and rank the carrying capacity of water resources in all regions of Hebei Province.

The rest of the paper is organized as follows: Literature review is introduced in *Literature Review*. The proposed method of Principal component analysis (PCA) is introduced in *Methods*. The case study is illustrated in *Data And Case Analysis*, and conclusions are discussed in *Conclusion*.

Literature Review

Research on the Concept of Water Carrying Capacity

At present, the research of water carrying capacity has a certain basis. The bearing capacity is a concept in mechanics, and it first appeared in the engineering field. It refers to the ability of the foundation to bear the load of buildings. Now it has been accepted and used in many fields, among which the most widely used is the study of environmental bearing capacity and resource bearing capacity in ecology. In 1921, Park and Burgess proposed the concept of ecological carrying capacity in the journal of Human Ecology. In the 1980s, UNESCO put forward the concept of carrying capacity (Nixon, et al., 2002). In 1999, United Research Service (URS Corp.)was commissioned by the United States Army Corps of Engineers and the Florida Society Office to study the carrying capacity of the Florida Keys Basin.

Taking India as an example, Joardor. (1998) studied the carrying capacity of urban water resources from the perspective of water supply, and incorporated it into urban development planning. Michiel A. Rijsberman et al. (2000) takes water carrying capacity as a measurement standard for urban water resource security.

In addition, the research on carrying capacity is more abundant in other fields. C. Bacher et al. (1997) researched ecosystem carrying capacity. Jonathan et al. (1999) researched the carrying capacity of water resources in agricultural production areas. Rees (1996) researched Urban water Supply capacity. Duarte et al. (2003) researched the Carrying capacity to coastal waters. Samuel Shephard et al. (2010) researched the Carrying Capacity of Marine ecosystems. Murray (2010) researched Population carrying capacity. Guangwei Huang (2012) has done extensive research on the carrying capacity of migratory waterfowl. France Salerno et al. (2013), based on the concept of environmental protection and sustainable development, established an environmental model to discuss the concept of tourism carrying capacity.

In China, Shi et al.(1989)first put forward the concept of the carrying capacity of water resources. Later on, Shi Yafeng (1992), Hui Yang He (2001), and Li Yunling et al. (2017) defined water carrying capacity from the maximum carrying capacity of water resources. Dictionary of Environmental Science (1991), Feng et al. (1997), Xia Jun (2002), Liu Jia-jun (2011), Duan Chunqing (2010), YANG Junfeng (2014), SUN Deliang (2018), and WANG Lili (2018) defined water carrying capacity from the maximum support scale. Xu Youpeng (1993), Gao Yanchun (1997), Hu Cheng (2013), and Song et al. (2011) defined the carrying capacity of water resources from the perspective of maximum development capacity of water resources. Tan Xiao (2018) believes that the carrying capacity of water resources is the embodiment of the sustainable development function of water resources-environment-economy-society system.

Research on the Evaluation Index System of Water Carrying Capacity

The index system of water carrying capacity is an important aspect of water carrying capacity research, but there are many factors affecting water carrying capacity, so scholars have established the evaluation index system of water carrying capacity from different perspectives.

Zhu et al. (2003) and Zhou Li (2016) established three subsystems covering water resources, ecological environment, social, and economic development. Liu et al. (2011), Zeng et al. (2013), Qu Xiao ‘e (2017), and Song et al. (2018) established an evaluation index system covering four aspects of water resources, society, economy, and ecological environment. Li et al. (2017) constructed the evaluation index system of water carrying capacity in the Yangtze River Economic Belt from four aspects: social economy, water resource quantity, water consumption, and wastewater discharge.

Research on Evaluation Methods of Water Carrying Capacity

At present, studies on the evaluation of water carrying capacity are mainly structured as follows:

Firstly, some evaluation method is used to comprehensively evaluate the index system, and then the water carrying capacity is ranked according to the evaluation results. Xu Youpeng (1993), for example, used a fuzzy comprehensive evaluation method to evaluate water carrying capacity. Fu et al. (1999) used principal component method to evaluate water carrying capacity. Zhu et al. (2003) used analytic hierarchy process to evaluate water carrying capacity. Zhou Li (2016), Huang Qiuxiang et al. (2016), Li et al. (2017), Liu et al. (2020), and Hong (2020) used principal component analysis and cluster analysis to evaluate the carrying capacity of water resources.

Secondly, some scholars established a mathematical model for quantification based on the interaction of some factors in water carrying capacity. For example, Qu Xiao 'e (2017) made a comprehensive evaluation of the water carrying capacity of relevant regions and cities by using the comprehensive evaluation method of TOPSIS. Li Yun et al. (2017) made an empirical analysis based on the technical route and evaluation standard of water carrying capacity evaluation. Song et al. (2018) proposed the improved abrupt progression method to evaluate the water carrying capacity of five provinces and cities in the lower reaches of the Yangtze River.

Review of Literature

To sum up, scholars have carried out different degrees of research on different cities, basins or regions, but a generally accepted viewpoint has not been formed in the systematic research on the concept of water carrying capacity, and there are still shortcomings:

1) The research on the concept of water carrying capacity has not yet formed a generally accepted theoretical system.

According to the existing studies, the concept of water carrying capacity can be divided into three categories: The maximum carrying capacity of water resources, the maximum supporting scale of water resources, and the maximum development capacity of water resources. Water carrying capacity is a comprehensive concept involving many elements such as society, economy, environment, and ecology. Existing definitions from one aspect, or from several aspects, do not fully cover the subject.

2) There are a few personal subjective factors involved in the selection of evaluation indicators of water carrying capacity. Therefore, it is necessary to adopt objective methods to reduce the impact of subjective factors and determine the importance of indicators for evaluation research.
3) The evaluation method of water carrying capacity is not comprehensive enough. The existing research methods are mainly based on comprehensive evaluation, but most scholars fail to consider the influence of index weight. Especially after the rural revitalization strategy is put forward, how Hebei province integrates the water resources of Beijing, Tianjin, and Hebei, is particularly important to the development of Hebei Province, but there is a lack of research on this aspect.

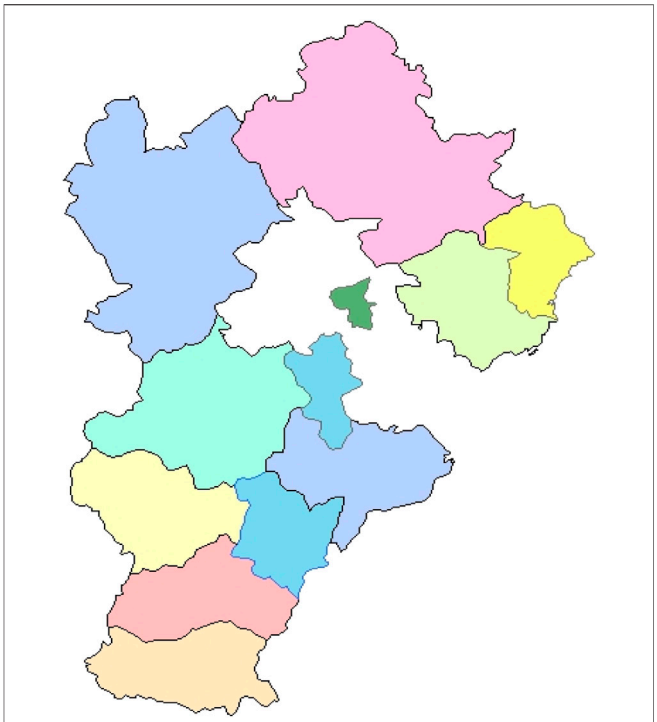


FIGURE 1 | Location map of research area.

METHODS

From the previous research results, the evaluation methods of water carrying capacity mainly include supply-demand balance method, analytic hierarchy process, fuzzy comprehensive evaluation method, principal component analysis method, systematic dynamic method, etc. In this paper, principal component analysis (PCA) is used to evaluate the carrying capacity of water resources in Hebei province. Principal component analysis (PCA) is an independent statistical analysis method that uses a small number of indicators to represent majority variable indicators and reflects the information reflected by majority variable indicators as much as possible through dimension reduction (France Salerno et al., 2013). The specific calculation steps are as follows.

Step1: In order to eliminate the impact of errors caused by order of magnitude and dimension, the original data are standardized.

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}, i = 1, 2, \cdots, n; j = 1, 2, \cdots p$$

Step2: Calculate the correlation coefficient matrix of the standardized samples.

$$R = \left[r_{ij}\right]_p Xp = \frac{Z^T Z}{n-1}$$

$$r_{ij} = \frac{\sum z_{kj} \cdot z_{kj}}{n-1}, i, j = 1, 2, \cdots, p$$

Step3: Calculate eigenvalues and eigenvectors.

TABLE 1 | Evaluation index system of water carrying capacity in Hebei Province.

The target layer	Rule layer	Index layer	Properties
Evaluation of water carrying capacity in Hebei Province	The economic development	X1 Total population at year end	+
		X2 Urbanization rate	+
		X3 Gross domestic product	+
		X4 Fixed asset investment	+
	The water resources	X5 Rainfall	+
		X6 Total water resources	+
		X7 The total water supply	+
	Water management	X8 Sewage discharge	-
	Industrial development	X9 Industrial water consumption	-
	Agricultural development	X10 Agricultural water consumption	-
		X11 Irrigation water per mu	-
		X12 Water for forestry, fishing and livestock	-
	The environmental protection	X13 Ecological water consumption	-
	Social development	X14 Domestic water consumption	-
		X15 Water consumption per capita	-
		X16 Water consumption per 10,000 yuan of GDP	-

*Note: "+" and "-" in **Table 1** are the positive and negative properties of indicators respectively. The positive index has a positive influence on the evaluation of water carrying capacity. The greater the index value is, the greater the water carrying capacity will be. The negative index has a negative impact on the evaluation of water carrying capacity. The greater the index value is, the smaller the water carrying capacity is.*

Step4: Calculate the contribution rate of principal component and the cumulative contribution rate. Under normal circumstances, determine the main component with the cumulative contribution rate greater than or equal to 85%.

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \geq 0.85$$

Step5: Calculate principal component load.

$$U_{ij} = z_i^T b_j^o, j = 1, 2, \cdots m$$

U_1 is called the first principal component, U_2 is called the second principal component, and Up is called the *p*th principal component.

Step6: Calculate comprehensive score and conduct comprehensive evaluation on regional water carrying capacity.

DATA AND CASE ANALYSIS

Study Area Status and Data Sources

Hebei is located between longitude 113°27′ and 119°50′ east and latitude 36°05′ and 42°40′ north. It is located in North China, north of Zhanghe river, east of Bohai Sea and inner Ring of Beijing and Tianjin, west of Taihang Mountain, North of Yanshan Mountain, North of Yanshan is Zhangbei Plateau and the rest is Hebei plain (As shown in **Figure 1**). It is the only province in China that has plateaus, mountains, hills, plains, lakes, and seashores. Hebei province covers an area of 188,800 square kilometers and has a permanent resident population of 75, 919, 700. It has jurisdiction over 11 prefecture-level cities, including Shijiazhuang city, Tangshan city, Handan city, Cangzhou city, Baoding city, Langfang city, Qinhuangdao city, Zhangjiakou city, Chengde city, Hengshui city, and Xingtai city. It has a temperate continental monsoon climate, and most of the four seasons are distinct.

The data used in this study are from Statistical Yearbook of Hebei Province and Hebei Water Resources Bulletin (2009-2018) which are calculated and sorted out.

Establish the Evaluation Index System

It can be seen from the definition of water carrying capacity that it is a comprehensive concept involving many factors such as society, economy, environment, ecology, etc. Therefore, in the evaluation and analysis of regional water carrying capacity, the selection of appropriate indicators should also involve several factors.

This article is based on the summary and reflection of the water carrying capacity system. According to the actual situation of Hebei Province, 16 factors were selected from the aspects of water resources, water management, industrial development, agricultural development, social development, and environmental protection to comprehensively evaluate the water carrying capacity of Hebei province from 2009 to 2018, as shown in **Table 1**.

Because Hebei province straddles Beijing and Tianjin, and has plateaus, mountains, hills, plains, lakes, and seashores, the distribution of water resources is inevitably uneven, which makes the water resources ineffectively used. In order to better understand the carrying capacity of water resources in Hebei Province, 11 cities in Hebei Province were also evaluated in this paper. Combined with the actual situation, the evaluation index system includes X_1 (The total population at the end of the year), X_2 (Gross regional product), X_3 (Per capita GDP), X_4 (Per capita disposable income of urban residents), X_5 (Per capita net income of farmers), X_6 (Per capita water consumption), X_7 (Total water resources), X_8 (Total water supply), X_9 (Industrial water consumption), X_{10} (Agricultural water consumption), X_{11} (Urban environmental water consumption), X_{12} (Water consumption per 10,000 yuan of GDP), X_{13} (Development and utilization rate of water resources), and X_{14} (Rainfall).

TABLE 2 | Economic and water resource status statistics of hebei province from 2009 to 2018.

year	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16
2009	7034.4	0.4374	17,026.6	12,311	463	141.16	193.72	20.51	23.71	134.2	150	12.96	2.69	20.15	275	114
2010	7194	0.4450	20,197.1	15,083	526	137.8	193.68	19.73	23.06	134.87	147	12.12	2.87	17.46	270	96
2011	7240.51	0.4560	24,228.2	16,404	493	157.29	195.97	21.49	25.72	132.01	168	11.6	3.62	18.63	271	81
2012	7287.51	0.4680	26,575	19,661	606	235.53	195.33	21.94	25.22	130.99	251	11.94	3.8	18.5	268	74
2013	7332.61	0.4812	28,301.4	23,194	531	175.86	191.29	21.93	25.23	126.35	187	11.29	4.65	18.79	261	67
2014	7383.75	0.4933	29,421.2	26,672	408	106.14	192.82	21.07	24.48	128.45	113	10.72	5.06	19.28	261	65
2015	7424.92	0.5133	29,806.1	29,448	511	135.09	187.19	31.1	22.53	124.18	141	11.05	5	19.5	252	62
2016	7470.05	0.5332	31,827.9	31,750	596	208.31	182.57	28.88	21.94	116.99	217	11.01	6.72	20.71	244	57
2017	7519.52	0.5501	35,964	33,407	479	138.34	181.56	25.37	20.33	114.31	148.97	11.78	8.17	21.75	241	50
2018	7556.3	0.5643	36,010.3	35,311	508	164.04	182.42	26.3	19.08	109.87	167.7	11.21	14.51	22.82	241	50

TABLE 3 | Correlation coefficient matrix of impact factors of water carrying capacity.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16
X1	1.000															
X2	0.957	1.000														
X3	0.985	0.955	1.000													
X4	0.976	0.984	0.966	1.000												
X5	0.108	0.055	0.075	0.040	1.000											
X6	0.072	0.040	0.099	0.012	0.912	1.000										
X7	0.824	0.931	0.810	0.893	0.105	0.017	1.000									
X8	0.701	0.741	0.649	0.758	0.280	0.131	0.763	1.000								
X9	0.652	0.803	0.622	0.723	0.027	0.137	0.885	0.574	1.000							
X10	0.904	0.980	0.915	0.938	0.100	0.106	0.945	0.693	0.831	1.000						
X11	0.038	0.000	0.070	0.026	0.906	0.998	0.022	0.090	0.179	0.065	1.000					
X12	0.753	0.595	0.700	0.703	0.013	0.047	0.389	0.518	0.132	0.489	0.077	1.000				
X13	0.793	0.874	0.801	0.809	0.010	0.040	0.778	0.484	0.821	0.910	0.008	0.407	1.000			
X14	0.639	0.825	0.688	0.740	0.136	0.029	0.834	0.527	0.832	0.880	0.069	0.168	0.862	1.000		
X15	0.941	0.987	0.928	0.974	0.109	0.046	0.966	0.774	0.815	0.970	0.008	0.573	0.818	0.788	1.000	
X16	0.981	0.903	0.982	0.939	0.139	0.140	0.736	0.658	0.503	0.845	0.113	0.793	0.712	0.553	0.882	1.000

TABLE 4 | Eigenvalues and contribution rates of principal components.

	The eigenvalue	Percentage of variance	Cumulative %
1	10.445	65.283	65.283
2	2.964	18.523	83.806
3	1.435	8.970	92.776

TABLE 5 | Factor loading matrix.

	1	2	3
X1	0.960	0.039	0.242
X2	0.998	0.027	0.014
X3	0.953	0.045	0.203
X4	0.985	0.035	0.139
X5	0.089	0.955	0.073
X6	0.062	0.981	0.100
X7	0.932	0.037	0.236
X8	0.761	0.151	0.069
X9	0.797	0.209	0.493
X10	0.981	0.020	0.154
X11	0.020	0.983	0.090
X12	0.613	0.020	0.750
X13	0.872	0.076	0.237
X14	0.809	0.171	0.464
X15	0.989	0.002	0.028
X16	0.907	0.113	0.362

Principal Component Analysis of Water Carrying Capacity in Hebei Province

The data of 16 factors reflecting the water carrying capacity of Hebei Province from 2009 to 2018 are shown in **Table 2**. The data in **Table 2** were standardized by SPSS25, and then the standardized data were analyzed by principal component analysis. The correlation coefficient matrix of the impact factors of water carrying capacity (**Table 3**) and the eigenvalue and contribution rate of principal components (**Table 4**) can be obtained.

As can be seen from **Table 3**, there is a certain correlation between the 16 factors. X_1With X_2, X_3, X_4 has a strong positive correlation, and with X_{10}、 X_{15}, X_{16} has a negative correlation; X_2With X_3, X_4 has a strong positive correlation, and with X_{10}, X_{15} has a negative correlation; X_3With X_4 has a strong positive correlation, and with X_{15}, X_{16} has a negative correlation; X_4With X_{10}, X_{14}, X_{15} has a negative correlation; X_5With X_6, X_{11} has a strong positive correlation. X_6With X_{11} has a strong positive correlation. X_7With X_{10}\X_{15} has a strong positive correlation. These explain the rationality of principal component analysis.

TABLE 6 | Component score coefficient matrix.

	1	2	3
X1	–0.029	0.190	–0.002
X2	0.083	0.050	–0.003
X3	–0.014	0.168	0.003
X4	0.018	0.133	–0.019
X5	0.002	–0.014	0.326
X6	0.011	–0.031	0.337
X7	–0.171	0.077	–0.011
X8	0.023	0.086	0.048
X9	–0.274	0.230	0.026
X10	–0.138	0.027	–0.024
X11	0.003	–0.028	0.336
X12	0.266	–0.449	0.065
X13	0.168	–0.082	–0.002
X14	0.261	–0.213	–0.016
X15	–0.087	–0.042	–0.006
X16	0.086	–0.255	–0.013

As can be seen from **Table 4**, the cumulative contribution rate of the first three principal components reaches 92.776%, which can be considered as the main factor affecting the water carrying capacity. Therefore, the first, second, and third principal components are selected to analyze the water carrying capacity of Hebei Province, and the loads of each variable on the first, second, and third principal components are calculated.

As can be seen from **Table 5**, the first principal component has a strong statistical significance with X_1, X_2, X_3, and X_4, and is negatively correlated withX_7, X_{15}, and X_{16}. This shows that the population and social and economic development level are the main factors affecting the carrying capacity of water resources in Hebei Province. The total population of Hebei increased from 70, 344, 000 in 2009 to 755,563,000 in 2018. With the increase of population, the demand for water resources also gradually increased, and the contradiction between water resources supply and demand intensified. Hebei's GDP in 2018 was 360.03 billion yuan, 111.49 % higher than 1702.66 billion yuan in 2009. Investment in fixed assets was 3.53109 trillion yuan in 2018, an increase of 186.84 percent over the 123.05 billion yuan in 2009. At the same time, with the rapid development of social economy, the consumption of water resources is increased, and the carrying capacity of water resources is under great pressure.

The second principal component and the third principal component have a strong positive correlation with X_5, X_6, X_{11}, and X_{13}, mainly reflecting the natural status of water resources. Hebei province is a big agricultural province with a large amount of agricultural water consumption. However, Hebei province is located in the semi-arid region of North China and is inherently deficient in water resources. In 2009, the total amount of water resources was 14.116 billion m^3. In 2012, the best year, the total water resources was 23.553 billion m^3, the total amount of water resources varies greatly from year to year, thus affecting the stability of the water carrying capacity.

Through factor analysis, the component scoring coefficient matrix (factor scoring coefficient) is obtained. **Table 6** lists the coefficient vectors of standardized variables in the analytical expressions of the three main components. We can write the expression of common factors (F1, F2, and F3 represent the three common factors, and ZX_1~ZX_6 respectively, represent the variables after the standard normal transformation):

$$F1= -0.029*ZX_1+0.083*ZX_2-0.014*ZX_3+0.018*ZX_4+0.002*ZX_5+0.011*ZX_6-0.171*ZX_7+0.023*ZX_8-0.274*ZX_9-0.138*ZX_{10}+0.003*ZX_{11}+0.266*ZX_{12}+0.168*ZX_{13}+0.261*ZX_{14}-0.087*ZX_{15}+0.086*ZX_{16}$$

Same thing with F2 and F3.

According to the formula, the comprehensive score of the water carrying capacity of Hebei Province can be obtained (see **Table 7**). The positive score means that the value is higher than the average level at the time of the study, while the negative score means that the value is lower than the average level. The larger the comprehensive score value is, the stronger the carrying capacity of water resources is, and conversely, the weaker the carrying capacity is.

As can be seen from **Table 7**, with the passage of time, the carrying capacity of water resources in Hebei province presents an increasing trend year by year, this is mainly due to the advancement of urbanization, the increase in Gross Domestic Product (GDP) and population, consumption, and increasing demand for water matched by increased water use efficiency and the ability to deal with sewage gradually, in addition, the constant improvement of the consciousness of water-saving among people, to some extent, also can improve the bearing capacity of water resources.

Comparative Analysis of Water Carrying Capacity of Various Cities in Hebei Province

Due to the serious uneven spatial and temporal distribution of water resources in Hebei Province, this uneven distribution has further reduced the effective supply of water resources. In order to better understand the carrying capacity of water resources in Hebei Province, this paper also selected the average data of the cities in Hebei province in the past 10 years, and compared and analyzed the differences of carrying capacity of water resources among cities in Hebei Province. The selected index system includes X_1 (total population at year end), X_2 (GDP), x_3 (Per capita GDP), x4 (The per capita disposable income of urban residents), X5 (The farmers' average net income), X6 (Per capita

TABLE 7 | Comprehensive scores of water carrying capacity in Hebei Province from 2009 to 2018.

year	F1	F2	F3	F	Ranking
2009	0.1680	1.7810	0.5065	0.2863	7
2010	0.5338	0.5681	0.2879	0.5169	9
2011	0.7925	0.1164	0.1279	0.5468	10
2012	0.5327	0.0941	1.9148	0.2085	5
2013	0.5986	0.5214	0.3820	0.2802	6
2014	0.7331	1.0468	1.6317	0.4646	8
2015	0.0821	0.8499	0.4256	0.0707	4
2016	0.4276	0.4943	1.2952	0.5248	3
2017	1.1706	0.1564	0.5332	0.7409	2
2018	1.5066	0.4292	0.0790	0.9668	1

TABLE 8 | Initial values of comprehensive evaluation indicators in Hebei province from 2009 to 2018.

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
Tangshan	758.63	5749.62	74,341.7	28,006.7	12,353.5	102.608	20.22	7.7163	1.565	5.082	0.1906	14.47	0.359	83.11
Shijiazhuang city	1049.665	4862.79	46,085.2	24,994.6	9926.2	52.903	17.452	5.88876	0.88439	1.5907	0.826	12.095	0.362	69.63
Cangzhou city	755.779	2974.474	73,432.3	23,640.05	9023.5	20.446	10.739	0.79037	0.25193	0.2081	0.045	2.895	0.088	76.13
Handan	932.56	2988.86	30,539.6	23,236	9796.4	18.423	10.172	1.9595	0.84264	0.164	0.27	6.34	0.189	60.63
Baoding	1062.795	2661.94	25,425.3	20,676.2	9178.5	20.704	22.408	1.72409	0.4191	0.534	0.203	6.772	0.079	116.61
Langfang	445.319	2075.65	46,018.5	27,969.8	11,750.8	39.883	6.642	1.7608	0.28686	0.884	0.105	9.955	0.299	34.67
Qinhuangdao	303.964	1204.296	39,725.4	25,097.3	9541.5	71.139	15.647	2.15639	0.61526	0.661	0.181	17.889	0.164	51.53
Zhangjiakou	440.18	1266.733	28,815.6	20,700.3	7134.8	33.709	14.622	1.54373	0.47891	0.764	0.039	12.757	0.106	154.24
Chengde	377.99	1238.334	32,919.6	19,668.54	6932.26	29.071	22.033	1.283	0.47766	0.233	0.019	9.599	0.054	199.75
Hengshui city	447.69	1130.71	25,316.1	19,948.1	7908	25.271	6.075	1.12486	0.33104	0.524	0.076	11.274	0.222	45.27
Xingtai	721.95	1653.724	22,874.1	20,316.5	8011.1	19.2	12.099	1.40052	0.3828	0.4715	0.2415	9.296	0.144	63.01

TABLE 9 | Characteristic values and contribution rates of principal components in each region of Hebei Province.

	The eigenvalue	Percentage of variance	Cumulative %
1	7.145	51.034	51.034
2	2.286	16.330	67.364
3	2.116	15.118	82.481
4	1.302	9.300	91.782

TABLE 10 | Factor loading matrix.

	1	2	3	4
x1	0.331	0.011	0.920	0.047
x2	0.879	0.067	0.436	0.148
x3	0.657	0.111	0.034	0.648
x4	0.816	0.378	0.194	0.204
x5	0.834	0.381	0.010	0.188
x6	0.850	0.218	0.456	0.007
x7	0.239	0.892	0.172	0.047
x8	0.952	0.241	0.067	0.123
x9	0.863	0.330	0.011	0.043
x10	0.890	0.238	0.166	0.131
x11	0.507	0.034	0.523	0.599
x12	0.411	0.245	0.704	0.503
x13	0.842	0.338	0.054	0.295
x14	0.334	0.860	0.021	0.217

water consumption), X_7 (Total water resources), X_8 (Total water supply), X_9 (Industrial water consumption), X_{10} (Agricultural water consumption), X_{11} (Urban environmental water consumption), X_{12} (Water consumption per 10,000 yuan OF GDP), X_{13} (Development and utilization rate of water resources), X_{14} (rainfall). The original data of different regions from 2009 to 2018 are shown in **Table 8**.

It can be seen from **Table 9**, the cumulative contribution rate of the first four principal components reached 91.454%. This can be considered as the main factor affecting the carrying capacity of water resources. Therefore, the first, second, third and fourth principal components are selected to analyze the carrying capacity of water resources in all regions of Hebei province, and the loads of each variable on the first, second, third and fourth principal components are calculated.

TABLE 11 | Component score coefficient matrix.

	1	2	3	4
x1	–0.014	0.358	–0.248	0.051
x2	0.149	0.147	–0.149	0.072
x3	0.365	–0.233	–0.261	0.053
x4	0.200	–0.107	–0.030	–0.145
x5	0.189	–0.036	–0.080	–0.138
x6	0.103	–0.099	0.215	0.055
x7	0.018	0.074	0.035	0.392
x8	0.045	0.135	0.112	0.076
x9	0.070	0.080	0.098	0.125
x10	0.158	–0.051	0.066	0.103
x11	–0.224	0.450	0.132	-0.087
x12	–0.159	–0.013	0.499	–0.028
x13	–0.023	0.143	0.159	–0.203
x14	0.034	–0.082	–0.029	0.404

As it can be seen from **Table 10**, there is a strong positive correlation between x_2、x_8、x_9、$x_{10,}$ indicating that agricultural, social, and economic development level are the main factors affecting the water carrying capacity of all cities in Hebei Province. The development of cities in Hebei province is based on agriculture, and agricultural development consumes large amounts of water resources and causes great pressure on the carrying capacity of water supply resources. The second principal component has a strong positive correlation with x_7 and x_{14}. There is a strong positive correlation between the third principal component and x_1.

Through factor analysis, it is concluded that component score coefficient matrix (coefficient of factor score), **Table 11** lists the four main composition analytic expressions of the standardized variable coefficient vector, we can write a common factor expression (F1, F2, F3, and F4 represent four common factors, ZX_1~ZX_{14} represent the standard normal after the transformation of variables):

$$F1= -0.014*ZX_1+0.149*ZX_2+0.365*ZX_3+0.200*ZX_4+0.189*ZX_5+0.103*ZX_6+0.018*ZX_7+0.045*ZX_8+0.070*ZX_9+0.158*ZX_{10}-0.224*ZX_{11}-0.159*ZX_{12}-0.023*ZX_{13}+0.034*ZX_{14}$$

Same thing with F2、F3, and F4.

According to the formula and principal component calculation, the comprehensive score of water carrying capacity in Hebei province can be obtained (see **Table 12**). The positive score means that the value is

TABLE 12 | Comprehensive scores of water carrying capacity of cities in Hebei Province.

Year	F1	F2	F3	F4	F	Ranking
Tangshan	2.4473	0.17586	0.92988	0.7411	1.6204	1
Shijiazhuang city	0.0208	2.31078	0.54503	0.167	0.4955	2
Cangzhou city	0.8872	0.7903	2.1027	0.181	0.0120	3
Handan	0.188	0.77635	0.7921	0.547	0.1525	6
Baoding	0.443	0.74457	0.9397	1.0044	0.1669	7
Langfang	0.6072	0.7807	0.0741	1.713	0.0130	4
Qinhuangdao	0.171	0.7973	1.55848	0.407	0.0214	5
Zhangjiakou	0.688	0.7271	0.52184	0.7982	0.3451	9
Chengde	0.571	0.8792	0.01533	1.7953	0.2895	8
Hengshui	0.934	0.3963	0.43077	1.01	0.6212	11
Xingtai	0.967	0.36329	0.0927	0.314	0.5203	10

higher than the average level at the time of the study, while the negative score means that the value is lower than the average level. The larger the comprehensive score value is, the stronger the carrying capacity of water resources is, and conversely, the smaller the comprehensice score value is, the weaker the carrying capacity is.

On the whole, Tangshan, Cangzhou, Langfang, Shijiazhuang, and other regions with relatively high economic development have relatively large industrial water consumption. However, with the reform of industrial technology, industrial water is reused, which reduces the industrial water consumption. Moreover, with the deepening of air pollution prevention and control, the government requires large water users such as metal smelters and chemical raw material manufacturers to stop production. The stronger their water resources development and utilization capacity is, the larger their water carrying capacity is.

Chengde city, Zhangjiakou city, and Baoding city are located in the vast Bashang grassland, with insufficient regional resources, but abundant precipitation, and the total water resources are in the forefront of the province. Their economic level of the province is in the middle level, but the development is strong and the demand for water resources is also large, so the carrying capacity of water resources is in the middle level. Due to the serious shortage of natural water resources, the total water resources and precipitation of Xingtai city and Hengshui city are relatively low, and the comprehensive score of water resources carrying capacity is relatively low.

CONCLUSION

1) Human activities are the main factors of water carrying capacity change in Hebei Province.

In recent years, the total amount of water resources in Hebei Province is gradually expanding, but the population is increasing year by year, the amount of water resources per capita is decreasing, and the water consumption per capita is increasing year by year. According to the change trend of water resources in Hebei Province from 2009 to 2018 and the water carrying capacity of various cities in Hebei Province, the total population, urbanization rate, GDP and fixed asset investment are the main factors affecting the water carrying capacity in Hebei Province. With the expansion of human activities and the development of the social economy to a certain stage, production activities and living behaviors have a great impact on the water environment, which is mainly manifested in the reduction of total water resources, large water consumption and serious water pollution.

2) The carrying capacity of water resources in Hebei Province showed a good trend.

According to the data analysis from 2009 to 2018, the water carrying capacity grade of Hebei Province is basically developing towards a good trend, and the carrying capacity of water resources is gradually improving. However, with the development of the social economy and the acceleration of urbanization, the contradiction between supply and demand of water resources will become increasingly prominent, and the comprehensive utilization of water resources should be strengthened.

3) There are differences in the carrying capacity of water resources in 11 cities of Hebei Province

Under the influence of natural conditions and policy factors, the carrying capacity of water resources in the eastern part of Hebei is higher than that in the northern and southern parts. During the 13th Five-Year Plan period, the measures on water resource environment optimization issued by the government are positive and effective. During the 14th Five-Year Plan period, it is necessary to continue to maintain a good momentum of development, introduce measures to boost the carrying capacity of water resources in various regions of Hebei, and narrow the differences between regions.

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

HW: Data curation, Investigation, Methodology, Project administration, Formal analysis, Resources, Validation, Visualization, Editing. YX:Writing-original draft. RS: Conceptualization. HM: Supervision. LW: Funding acquisition.

ACKNOWLEDGMENTS

The relevant research is supported by 2019 Hebei Provincial Social Science Fund project (subject No: HB19GL062), 2019 Handan Science and Technology Bureau project (Subject No: 19422303008–71), 2018 Hebei Provincial Water Ecological Civilization and Social Governance Research Center Project (Subject No: 2018SZX7).

REFERENCES

Bacher, C., Duarte, P., and Ferreira, J. G. (1997). Assessment and Comparison of the Marennes-Oléron Bay (France) and Carlingford Lough (Ireland) Carrying Capacity with Ecosystem Models. *Aquat. Ecol.* 31, 379–394. doi:10.1023/A:1009925228308

Duarte, P., Meneses, R., and Hawkins, A. J. S. (2003). Mathematical Modelling to Measure the Carrying Capacity for Mufti-Species Culture within Coastal Waters. *Ecol. Model.* doi:10.1016/S0304-3800(03)00205-9

Hong, X. (2020). Analysis of Water Resources Carrying Capacity in Xinjiang Based on Principal Component Analysis. *Water Resour. Plann. Des.* 9, 39–41+60. doi:10.3969/j.issn.1672-2469.2020.09.010

Joardar, S. D. (1998). Carrying Capacities and Standards as Bases towards Urban Infrastructure Planning in India: A Case of Urban Water Supply and Sanitation. *Plann. IndiaHabitat Intl* 22, 327–337. doi:10.1016/S0197-3975(98)00002-2

Murray, L. (2010). *The Carrying Capacity Imperative: Assessing Regional Carrying Capacity Methodologies for Sustainable Land-Use Planning*. Florida: Land use Policy. doi:10.1016/j.landusepol.2010.01.006

Nixon, S. W., Dalrymple, G. H., Deyle, R. E., Huber, W. C., Peterson, M. S., Polasky, S., et al. (2002). *A Review of the Florida Keys Carrying Capacity studyCommittee to Review the Florida Keys Carrying Capacity Study*. Florida: National Academy Press. http://www.nap.edu/books/030908346X/html.

Rees, William. E. (1996). Revisiting Carrying Capacity: Area-Based Indicators of Sustainability[J]. *Popul. Environ.* 17 (3), 195–215. doi:10.1007/BF02208489

Rijsberman, M. A., Frans, H., and van de Ven, M. (2000). *Different Approaches to Assessment of Design and Management of Sustainable Urban Water Systems*. Delft: Environment Impact Assessment Review, 333–345. doi:10.1016/S0195-9255(00)00045-7

Salerno, F., Viviano, G., and Manfredi, E. C., (2013). Multiple Carrying Capacities from a Management-Oriented Perspective to Operationalize Sustainable Tourism in Protected Areas. *J. Environ. Manage.*, 116–125. doi:10.1016/j.jenvman.2013.04.043

Shephard, S., Brophy, D., and David, G. (2010). Reid, Can Bottom Trawling Indirectly Diminish Carrying Capacity in a marine Ecosystem? *Mar. Biol.*, 2375–2381. doi:10.1007/s00227-010-1502-9

Zeng, H., Zhang, Z-w., Sun, X-z., and Li, Q-G. (2013). *Study of Water Resources Carrying Capacity in the Hanjiang River Basin of Hubei*. Oslo: South-to-North Water Transfers and Water Science & Technology, 22–25+30. doi:10.3724/SP.J.1201.2013.04022

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.

Copyright © 2021 Wang, Xu, Suryati Sulong, Ma and Wu. 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 in Environmental Science

ORIGINAL RESEARCH
published: 12 November 2021
doi: 10.3389/fenvs.2021.757533

Research on Influence Factors of Air Quality in the Host Area of the 24th Winter Olympic Games Based on Grey Relational Analysis

Wei Wang[1], *Yan Chen*[1], *Yuhan Xie*[1] *and Lifeng Wu*[1,2]*

[1]College of Management Engineering and Business, Hebei University of Engineering, Handan, China, [2]Hebei Key Laboratory of Intelligent Water Conservancy, Hebei University of Engineering, Handan, China

In recent years, the continuous development of the economy and science and technology of China has caused a certain degree of pollution to the atmospheric environment on which the people depend. The current air pollution problem is actively a concern by the government and all walks of life. Based on the 2015–2019 air quality indicators and some socioeconomic indicators, this paper uses the grey correlation analysis method to analyze the Beijing and Zhangjiakou cities that will host the Winter Olympics in 2022. The study found that the three factors most closely related to the Beijing Air Quality Index (AQI) are the permanent population (0.831), energy consumption (0.801), the number of motor vehicles (0.79), and the permanent population (0.916) and industrial added value (0.905). The total output value of agriculture, forestry, and animal husbandry and fishery (0.89) are the three factors most closely related to the air quality index (AQI) of Zhangjiakou City, and the permanent population is the common factor affecting the two cities. Considering that the factors that affect the air quality of the two cities are not exactly the same, this paper combines the development positioning of the two cities and their own characteristics, and puts forward specific suggestions and opinions on the different problems faced by the two cities. The aim is to promote the continuous improvement of air quality in the two cities to reach an excellent level through scientific and feasible air management programs before the opening of the 2022 Winter Olympic Games, and help the 2022 Winter Olympics to be held smoothly.

Keywords: air quality, grey relational analysis, permanent population, governance measures, 24th Winter Olympic Games

OPEN ACCESS

Edited by:
Pei Du,
Xi'an Jiaotong University, China

Reviewed by:
Hongfang Lu,
Southeast University, China
Xinping Xiao,
Wuhan University of Technology, China
Rana Muhammad Adnan Ikram,
Hohai University, China

***Correspondence:**
Lifeng Wu
wlf6666@126.com

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: *12 August 2021*
Accepted: *12 October 2021*
Published: *12 November 2021*

Citation:
Wang W, Chen Y, Xie Y and Wu L (2021) Research on Influence Factors of Air Quality in the Host Area of the 24th Winter Olympic Games Based on Grey Relational Analysis. Front. Environ. Sci. 9:757533. doi: 10.3389/fenvs.2021.757533

INTRODUCTION

In recent years, China's economy, production, industry, and other aspects have been on rapid development. The requirements of the people for material life and the pursuit of spiritual life are constantly improving. Therefore, air quality has become one of the most concerning ecological and environmental issues of the public. Due to industrial development and the continuous improvement of living standards, pressure on the ecological environment has also increased. For example, industrial exhaust, automobile exhaust, coal combustion exhaust, and many other harmful gases are emitted into the atmosphere, resulting in many areas in the north of China being repeatedly covered by haze. Haze is a particular concern in the Beijing–Tianjin–Hebei region, where the air

pollution problem is most prominent. Because the indicators for measuring air quality are affected by a variety of environmental factors, and the main air pollutants vary from region to region, analyzing the pollutants affecting air quality from multiple perspectives and selecting reasonable and valuable indicator data for correlation comparison has become the focus of research and drawn the attention of many scholars in China and internationally. Since the theory of the grey system was proposed by Professor Julong Deng in 1982, it has caused many scholars to study deeply and explore actively, making it possible to expand the depth and breadth of the grey system theory in continuous development to better serve national and social development. The advantage of the grey system theory is that it can perform a systematic analysis of the uncertainty of data with a small sample and sparse information and reveal the connections between, and potential value of, the data. Grey relational analysis, as an important branch of the grey system theory, is a new approach to factor analysis (Xiao, 1997). Grey relational analysis has now been applied to many research fields such as factor analysis, comprehensive evaluation, and program decision making (Deng, 1989). Wang et al. studied the sources and characteristics of PM2.5 in the areas where the Olympic Games were held and found that the main factors of PM2.5 pollution in the two areas were quite different, and the contribution of the surrounding areas to PM2.5 in the areas where the Olympic Games were held was nearly twice that of the remote areas (Wang et al., 2021). Liu et al. used a combination of hybrid simulation and risk analysis to study 243 cities and towns in Zhangjiakou City and found that 34 cities and towns may be affected by aviation hazards, but the construction site for the 2022 Winter Olympics will not be threatened by the acute air pollution in Zhangjiakou City (Liu et al., 2018). Zhan et al. studied the correlation between air pollutant indicators and economic and social indicators in the Guangdong–Hong Kong–Macao Greater Bay Area from 2006 to 2016 (Zhan et al., 2018). Wang and He studied the correlation between $PM_{2.5}$ ($PM_{2.5}$ is particulate matter with a diameter ≤2.5 μm) and six elements of air quality in Urumqi from 2014 to 2016 using grey relational analysis and found that CO concentration values had the greatest influence on $PM_{2.5}$ (Wang and He, 2018). Based on the grey model, Ni took Wuhan City as an example and found that the main factors influencing air quality were the proportion of secondary industry, total urban population, industrial SO_2 emissions, and public green space per capita (Ni, 2013). Gao and Wu used the entropy-weighted grey relation method to study the factors influencing air quality in Beijing and found that the proportion of secondary industry, average temperature, and industrial SO_2 emissions were the main factors affecting air quality in Beijing (Gao and Wu, 2017). Li et al. analyzed the influence factors of air quality indexes in the Beijing, Tianjin, and Hebei regions from 2013 to 2017 and showed that rainfall, wind speed, and air temperature were negatively correlated with air quality index, while relative humidity and air pressure varied from region to region (Li and Wang, 2019). Wang and Tian used grey relational analysis to investigate the influence factors of air quality in Yichang City and found that different air quality indexes had different influence factors (Wang and Tian, 2019). Nan and Sun used a grey correlation model to analyze the correlation between O_3 concentration and influencing index factors in Shanxi Province in 2015 and analyzed the factors that were more correlated with O_3 from different perspectives (Nan and Sun, 2017). Wang et al. used grey relational analysis to study the factors influencing air quality in Handan from 2014 to 2018 and found that the industrial value added, energy consumption, and motor vehicle ownership were the main factors influencing air pollution in Handan (Wang et al., 2019). Wang used a modified grey relational analysis to analyze the factors influencing haze in Zhengzhou City and found that vehicle emissions, smoke and dust emissions, and construction site dust were the main controllable influence factors (Wang, 2020). Zhang et al. showed that local air pollutant emissions, lagged $PM_{2.5}$ concentrations, wind speed, and $PM_{2.5}$ concentrations in nearby areas had significant effects on $PM_{2.5}$ concentrations in the Beijing–Tianjin–Hebei region (Zhang et al., 2018). Li et al. established an air quality research model using a particulate matter source analysis technique and concluded that the main sources of $PM_{2.5}$ in Beijing were both residential and industrial (Li et al., 2015). In their 2016 study, Yan et al. showed that $PM_{2.5}$ concentrations in 13 cities, including Beijing, Tianjin, and Hebei, were influenced by seasonality, and the seasons with high to low concentrations were winter, autumn, spring, and summer in order (Yan et al., 2018). Tian et al. showed that seasonal effects, land use, vehicle density, and emission intensity were the main influence factors that continued to affect air quality in Beijing (Tian et al., 2019). Chen et al. analyzed and discussed the spatial–temporal coupling relationship and restrictive factors among different cities in Ningxia Hui Autonomous Region by combining grey relational analysis with coupling coordination model and linear regression (Chen et al., 2021). Khuman et al. regarded the absolute grey correlation degree as a part of the novel R-fuzzy grey analysis framework (RfGAf). The framework based on subjective uncertainty has the advantages of capturing information comprehensively, timely, and with a wide range of adaptations (Khuman et al., 2019). Feng and Sun used the grey correlation analysis method to select important indicators for the economic, environmental, and social benefits of the environmental investment of China, and established a comprehensive evaluation model for the benefits of environmental investment (Feng and Sun, 2020). Wu et al. used a model combining grey correlation and entropy to analyze the relationship between water resources and the economy in the Beijing–Tianjin–Hebei region of China and found that the correlation between the two systems from 2015 to 2024 is weak (Wu et al., 2021). Xiao et al. combined the grey model with the coupling model to systematically study the coordination relationship between the two important factors of the economy and technology of China (Xiao et al., 2021). Zhu et al. studied the relationship between the regression coefficients and physical properties of coarse aggregates by means of grey relational analysis (Zhu et al., 2021). Fiaz et al. used the grey relational analysis method to study the first wave of COVID-19 epidemic data obtained from the worldometer website, and the research results can bring a lot of useful information to the health

systems of many countries (Qazi et al., 2021). Li and Ye chose the grey correlation analysis method to analyze the coordination of environmental policies in the 281 prefecture-level cities in China from 2007 to 2016 and found that policy coordination has a significant positive impact on haze governance (Li and Ye, 2020). Wu and Qu proposed a new type of grey relational model of dynamic weighting function that enables the problem of multi-factor and multi-attribute classification to be better solved (Wu and Qu, 2020). Xiao et al. proposed a grey correlation model based on language binary matrix to quantitatively evaluate the operation quality of regional industrial industries, and promoted the innovation and application prospect of the grey correlation model (Xiao et al., 2020). Wang et al. used the grey correlation method to sort the multiple stages of different alternatives in water pollution control to get the overall ranking, and then allocated resources according to priority (Wang et al., 2020). Gong et al. found that the main factors affecting the durability of mixed modified asphalt mixtures are aggregate gradation and aging temperature through the grey correlation analysis method (Gong et al., 2021). Luo et al. used the grey correlation analysis method to analyze the three aspects of energy, economy, and environment in Guangxi and Zhejiang in China, and found that energy consumption can promote economic development but inhibit environmental improvement (Luo et al., 2019).

As can be seen from the above studies, most of the studies on air quality are mainly based on meteorology, greenery, seasons, and energy consumption, while relatively few studies analyze air quality in conjunction with economic and social factors in the study region. Based on this, this paper uses grey relational analysis to study the main socioeconomic factors affecting air quality in Beijing and Zhangjiakou from 2015 to 2019, finds the key influence factors through comparative analysis, and proposes specific management measures to contribute to the successful holding of the 2022 Winter Olympic and Paralympic Winter Games.

DATA AND METHODS

Data Source and Processing

In the rapid economic development of modern China, air pollution is influenced by multiple factors. Agriculture, industry, energy consumed in winter for residential heating, and many pollutants such as carbon monoxide, carbon dioxide, nitrogen oxides, and hydrocarbons emitted from motor vehicle exhaust can cause serious air pollution. The air pollution in most regions is mainly influenced by the population base, ecological environment, industrial status, and energy consumption of the region. As a super-populous city and the economic center of northern China, many factors affect the air quality in Beijing. In Zhangjiakou, the secondary and tertiary industries are developing rapidly under the 13th Five-Year Plan, and the number of domestic and foreign tourists visiting Zhangjiakou is increasing year by year. As a result, the air pollution indicators in Beijing and Zhangjiakou are affected by many socioeconomic factors. Considering the availability, reference, and outstanding representativeness of data indicators, seven air pollution-related indicators were selected as a reference sequence, and 10 representative socioeconomic indicators were selected as a comparison sequence for grey relational analysis in Beijing, while six air pollution-related indicators were selected as a reference sequence and eight representative socioeconomic indicators were selected as a comparison sequence in Zhangjiakou. The air pollution data of Beijing were obtained from http://sthjj.beijing.gov.cn/. The socioeconomic data of Beijing were obtained from http://tjj.beijing.gov.cn/. The air pollution data of Zhangjiakou were obtained from http://hb.zjk.gov.cn/. The socioeconomic data of Zhangjiakou were obtained from http://tjj.zjk.gov.cn/. The air pollution data of Beijing for 2015–2019 are shown in **Table 1**, and the socioeconomic data of Beijing are shown in **Table 2**. The 2015–2019 air pollution data of Zhangjiakou are shown in **Table 3**, and the socioeconomic data of Zhangjiakou are shown in **Table 4**.

TABLE 1 | Air pollution data of Beijing.

Indicator	2015	2016	2017	2018	2019
AQI	7.4	6.8	5.87	5.35	4.7
$PM_{2.5}$ ($\mu g/m^3$)	80.6	73	58	51	42
PM_{10} ($\mu g/m^3$)	102	92	84	78	68
SO_2 ($\mu g/m^3$)	14	10	8	6	4
NO_2 ($\mu g/m^3$)	50	48	46	42	37
$CO-95_{per}$ (mg/m^3)	3.6	3.2	2.1	1.7	1.4
$O_3-8h-90_{per}$ ($\mu g/m^3$)	202.6	199	193	192	191

Methods

Since the air quality situation is closely related to the time point, in order to conduct a more effective analysis, this paper selected the data of air pollution and socioeconomic indicators from 2015 to 2019. However, as the sample size was relatively small, and the relationships between the data were not easily found directly, many existing statistical models were not suitable for conducting data analysis and research using such data. For example, the Pearson correlation coefficient proposed around the 1880s mainly studies the degree of correlation between two variables, but it becomes quite difficult to study the degree of correlation between multiple variables at the same time. Moreover, Pearson correlation coefficient method can only reflect the linear relationship between two variables, and the direction of correlation has some limitations. In 1982, The grey model proposed by the Chinese scholar Professor Deng Julong has been committed to solving the problem of uncertainty in terms of small samples and poor information. Among them, the grey correlation analysis model can process and mine the potential correlations between multiple indicators at the same time, and then obtain the correlation degree between multiple comparison indicators, which is a great theoretical and practical significance. In the approximately 40 years since the appearance of the grey correlation model, due to its good correlation analysis performance, it has been popularized and applied by many scholars in wide fields. Because other factors such as demographic factors, economic factors, industrial factors, and ecological factors can cause different degrees of air pollution, these complex factors can be understood as a grey system. Therefore, for this paper, it was appropriate to choose grey relational analysis for the research and

TABLE 2 | Socioeconomic data of Beijing.

Indicator	2015	2016	2017	2018	2019
Resident population (10,000 people)	2,170.5	2,172.9	2,170.7	2,154.2	2,153.6
Number of tourists (10,000 people)	27,279	28,531.5	29,746.2	31,093.6	32,209.9
Total output value (hundred million yuan)	24,779.1	27,041.2	29,883	33,106	35,371.3
Construction industry output value (hundred million yuan)	1,002.6	1,074.5	1,210.9	1,387.8	1,513.7
Industrial value added (hundred million yuan)	3,458.9	3,635.5	3,885.9	4,139.9	4,241.1
Output value of tertiary industry (hundred million yuan)	20,218.9	22,245.7	24,711.7	27,508.1	29,542.5
Energy consumption (10 thousand tons)	6,802.8	6,916.7	7,088.3	7,269.8	7,360.3
Private car ownership (10 thousand)	440.3	452.8	467.2	479	497.4
Artificial forestation area (hectares)	8,252	12,667	11,853	17,974	18,698
Park area (hectares)	29,503	30,069	31,019	32,619	35,157

TABLE 3 | Air pollution data of Zhangjiakou.

Indicator	2015	2016	2017	2018	2019
AQI	4.64	4.5	4.18	4.11	3.53
$PM_{2.5}$ ($\mu g/m^3$)	34	32	31	29	25
PM_{10} ($\mu g/m^3$)	78	83	70	69	56
SO_2 ($\mu g/m^3$)	31	20	16	14	11
NO_2 ($\mu g/m^3$)	26	27	25	23	22
$CO-95_{per}$ (mg/m^3)	1.6	1.4	1.3	1.4	1.1

TABLE 4 | Socioeconomic data of Zhangjiakou City.

Indicator	2015	2016	2017	2018	2019
Resident population (10,000 people)	442.17	442.51	443.3	443.4	442.33
Number of tourists (10,000 people)	3,848	5,193.77	6,259.8	7,354.8	8,605.06
Total output value (hundred million yuan)	1,363.54	1,461.05	1,555.6	1,536.6	1,551.06
Industrial added value (hundred million yuan)	404.7	373.9	372.5	423.9	456.96
Output value of tertiary industry (hundred million yuan)	574.13	651.86	726	791.6	861.87
Total output value of agriculture, forestry, animal husbandry, and fishery (hundred million yuan)	430.93	465.05	493.4	403.7	438.51
Artificial forestation area (hectares)	47,444	81,075	137,211	127,392	50,533
Urban green area (hectares)	3,791	4,163	3,848.7	3,989.7	4,081.22

analysis of the air pollution data of Beijing and Zhangjiakou, and the socio–economic indicators of the two regions. Grey relational analysis is used to determine the degree of relation between different data by the proximity of reference data and comparison data, and the specific steps of grey relational analysis are as follows.

Step 1: $Y_0 = (y_0(1), y_0(2), \cdots, y_0(n))$ as the reference sequence, and $Y_1 = (y_1(1), y_1(2), \cdots, y_1(n))$, ..., $Y_m = (y_m(1), y_m(2), \cdots, y_m(n))$ as the comparison sequence.

Step 2: Normalize data as in **Eq. 1**:

$$X_i(j) = \frac{y_i(j)}{y_i(1)}, i = 0, 1, 2, \cdots m; j = 1, 2, \cdots n \tag{1}$$

Step 3: Calculate the grey correlation coefficient using **Eq. 2**:

$$\xi_i(x_0(k), x_i(k)) = \frac{\min\limits_i \min\limits_k \left|X_0(k) - X_i(k)\right| + \rho \max\limits_i \max\limits_k \left|X_0(k) - X_i(k)\right|}{\left|X_0(k) - X_i(k)\right| + \rho \max\limits_i \max\limits_k \left|X_0(k) - X_i(k)\right|} \tag{2}$$

where $\rho \in [0, 1]$ is the differentiation coefficient. According to most of the literature on the use of the grey relation coefficient, $\rho = 0.5$. It is further confirmed that the correlation result obtained when the value of a is relatively better, in the *Grey relational analysis between socio–economic factors and air pollution in Beijing and Zhangjiakou* section of this paper, the value of a in the interval of 0–1 is selected for three equal points to verify the case again.

Step 4: Calculate the grey relation using **Eq. 3**:

$$\gamma_i\left(y_{(0)}, y_{(i)}\right) = \frac{1}{n}\sum_{k=1}^{n} \gamma\left(x_{(0)}(k), x_{(i)}(k)\right) \tag{3}$$

where $\gamma_i(y_{(0)}, y_{(i)})$ is the grey relation of Y_0 and Y_i. If the grey relation is closer to 1, it means that the reference sequence Y_i has more influence on the main sequence Y_0. Otherwise, it means that the influence is less.

Step 5:

$$\bar{\gamma} = \frac{1}{m}\sum_{i=1}^{m} \gamma_i\left(y_{(0)}, y_{(i)}\right) \tag{4}$$

$\bar{\gamma}$ is the mean value of grey correlation obtained by all control indicators corresponding to the same reference indicator.

GREY RELATIONAL ANALYSIS BETWEEN SOCIOECONOMIC FACTORS AND AIR POLLUTION IN BEIJING AND ZHANGJIAKOU

Taking the grey relation between the air quality index (AQI) and the selected socioeconomic pollution sources of Beijing as an

TABLE 5 | Relational analysis between socio–economic factors and air pollutants in Beijing.

Socio-economic factor	AQI	$PM_{2.5}$	PM_{10}	SO_2	NO_2	CO – 95_{per}	O_3 – 8h – 90_{per}
Resident population (10,000 people)	0.831	0.803	0.84	0.739	0.888	0.768	0.954
Number of tourists (10,000 people)	0.772	0.753	0.777	0.702	0.814	0.727	0.85
Total output value (hundred million yuan)	0.713	0.702	0.716	0.664	0.743	0.684	0.758
Construction industry output value (hundred million yuan)	0.708	0.697	0.71	0.66	0.736	0.68	0.749
Industrial value added (hundred million yuan)	0.757	0.74	0.762	0.693	0.795	0.716	0.825
Output value of tertiary industry (hundred million yuan)	0.706	0.696	0.709	0.659	0.734	0.678	0.747
Energy consumption (10 thousand tons)	0.801	0.779	0.809	0.721	0.851	0.748	0.901
Private car ownership (10 thousand)	0.79	0.768	0.796	0.714	0.836	0.739	0.88
Artificial forestation area (hectares)	0.564	0.565	0.564	0.555	0.573	0.561	0.561
Park area (hectares)	0.785	0.765	0.791	0.711	0.83	0.736	0.87

TABLE 6 | Relational analysis between socioeconomic factors and air pollutants in Zhangjiakou.

Socio-economic factor	AQI	$PM_{2.5}$	PM_{10}	SO_2	NO_2	CO – 95_{per}
Resident population (10,000 people)	0.916	0.903	0.904	0.762	0.935	0.879
Number of tourists (10,000 people)	0.64	0.634	0.649	0.596	0.649	0.63
Total output value (hundred million yuan)	0.849	0.838	0.861	0.726	0.876	0.821
Industrial added value (hundred million yuan)	0.905	0.905	0.884	0.765	0.896	0.884
Output value of tertiary industry (hundred million yuan)	0.766	0.758	0.779	0.678	0.785	0.746
Total output value of agriculture, forestry, animal husbandry and fishery (hundred million yuan)	0.89	0.877	0.901	0.748	0.92	0.859
Artificial forestation area (hectares)	0.606	0.6	0.608	0.567	0.618	0.597
Urban green area (hectares)	0.88	0.869	0.891	0.743	0.91	0.848

example, the following steps of grey relational analysis were carried out in calculation: the reference sequence is $Y_0 = (7.4, 6.8, 5.87, 5.35, 4.7)$, and the comparison sequence is:

$$Y_1 = (2170.5, 2172.9, 2170.7, 2154.2, 2153.6),$$
$$Y_2 = (27279, 28531.5, 29746.2, 31093.6, 32209.9),$$
$$Y_3 = (24779.1, 27041.2, 29883, 33106, 35371.3),$$
$$Y_4 = (1002.6, 1074.5, 1210.9, 1387.8, 1513.7),$$
$$Y_5 = (3458.9, 3635.5, 3885.9, 4139.9, 4241.1),$$
$$Y_6 = (20218.9, 22245.7, 24711.7, 27508.1, 29542.5),$$
$$Y_7 = (6802.8, 6916.7, 7088.3, 7269.8, 7360.3),$$
$$Y_8 = (440.3, 452.8, 467.2, 479, 497.4),$$
$$Y_9 = (8252, 12667, 11853, 17974, 18698),$$
$$Y_{10} = (29503, 30069, 31019, 32619, 35157).$$

From steps 1–4,

$$\gamma(Y_{(0)}, Y_{(1)}) = 0.831, \gamma(Y_{(0)}, Y_{(2)}) = 0.772$$
$$\gamma(Y_{(0)}, Y_{(3)}) = 0.713, \gamma(Y_{(0)}, Y_{(4)}) = 0.708$$
$$\gamma(Y_{(0)}, Y_{(5)}) = 0.757, \gamma(Y_{(0)}, Y_{(6)}) = 0.706$$
$$\gamma(Y_{(0)}, Y_{(7)}) = 0.801, \gamma(Y_{(0)}, Y_{(8)}) = 0.79$$
$$\gamma(Y_{(0)}, Y_{(9)}) = 0.564, \gamma(Y_{(0)}, Y_{(10)}) = 0.785$$

were obtained.

The grey relation of each reference sequence and its corresponding comparison sequence for Beijing and Zhangjiakou were calculated using the same steps with the results shown in **Tables 5** and **6**, respectively.

In order to further confirm the feasibility of the resolution coefficient of 0.5, this paper selected the resolution coefficient of 0.25, 0.5, and 0.75 for comparison. The results of the analysis of the socioeconomic factors and air pollutants in Beijing are shown in **Table 5**. When the resolution coefficient is 0.5, it can be obtained from step 5.

$$\bar{\gamma}_{AQI} = 0.743,\ \bar{\gamma}_{PM_{2.5}} = 0.727,\ \bar{\gamma}_{PM_{10}} = 0.747,\ \bar{\gamma}_{SO_2} = 0.682,$$
$$\bar{\gamma}_{NO_2} = 0.78,\ \bar{\gamma}_{CO-95_{per}} = 0.704,\ \bar{\gamma}_{O_3-8h-90_{per}} = 0.81$$

In the same way, the results when the resolution coefficients are 0.25 and 0.75 can also be calculated in steps

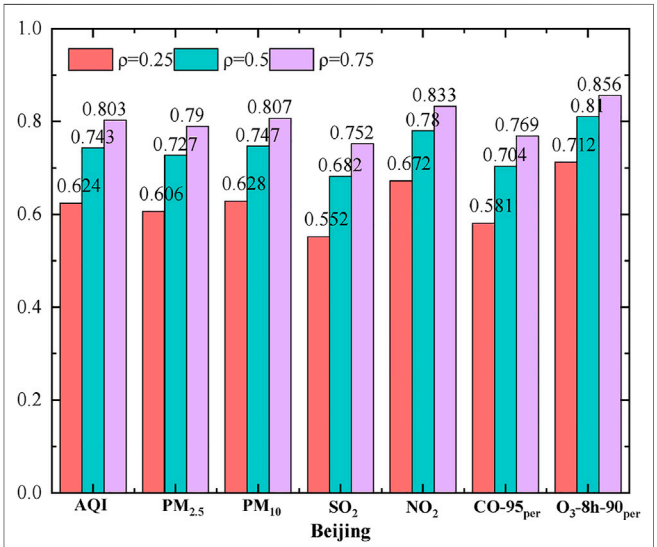


FIGURE 1 | Grey average correlation coefficient of air index under different resolution coefficients in Beijing.

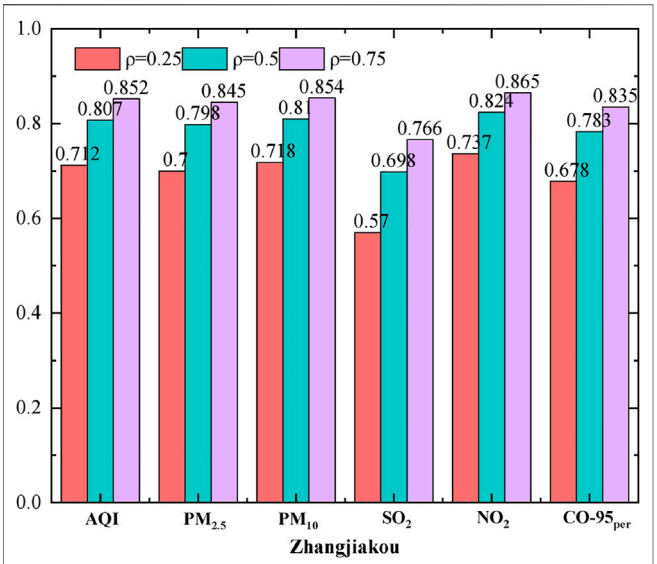


FIGURE 2 | Grey average correlation coefficient of air index under different resolution coefficients in Zhangjiakou.

1–5. The comparison of different resolution coefficient values in Beijing is shown in **Figure 1**, and the comparison of different resolution coefficient values in Zhangjiakou City is shown in **Figure 2**. From **Figures 1** and **2**, it can be clearly observed that when the resolution coefficient is 0.5, it is more reasonable and objective than when the value is 0.25 or 0.75.

DISCUSSION AND RESULTS

From the results in **Table 5**, it can be obtained that the three social and economic factors that are most relevant to the Beijing air quality index (AQI) are permanent population (0.831), energy consumption (0.801), and private car ownership (0.79). At the same time, these three socioeconomic indicators are also the main factors affecting $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO-95_{per}, and O3-8h-90_{per}. As the capital of China, Beijing is a mega city with a resident population of more than 20 million, and the urban resident population has reached more than 10 million. Faced with such a large population, just taking essential breathing every day will produce a lot of carbon dioxide emissions. What is more, the work and daily life of these permanent residents in Beijing will cause more exhaust gas to be discharged into the natural environment to affect the air condition. Besides, the overall situation of air quality will also be affected by the large concentration of the population in some urban areas and the population of many other provinces going to Beijing for job hunting and entrepreneurship. In addition to the permanent population factor, energy consumption is the second indicator that is closely related to all air quality indicators in Beijing. Although many high-emission and high-polluting factories and facilities have been moved out of Beijing in recent years, massive emissions of air pollutants in Beijing have been solved in a short period of time. However, because the government encourages the expansion and development of the tertiary industry, it has also promoted the accelerated development of machinery manufacturing, artificial intelligence technology industries, various processing companies, and social service-related industries. This makes the consumption of non-renewable energy and electric energy show a certain upward trend. Since the consumption of non-renewable resources will have a greater impact on the environment than the consumption of renewable resources, it is necessary to always strive to explore possible renewable and clean energy. On the other hand, due to the influence of the region, Beijing needs to provide a large area of heating guarantee every winter, and the consumption of energy such as coal and natural gas also increases. Many indicators that consider environmental quality are disturbed by the dual effects of weather factors and pollutant emissions in winter. Second, affected by the continuous improvement of the living standards of the current people, the private car ownership in Beijing has shown a trend of steadily increasing year by year within the time frame of the study. The increase in private car ownership will undoubtedly increase the consumption of oil, natural gas, and electric energy, resulting in the exhaust gas of vehicles using such energy as petroleum containing harmful substances such as CO, hydrocarbons, nitrogen oxides, and particulate matter of different sizes. In turn, it will have an extremely bad impact on the atmospheric environment. The research and development and vigorous promotion and use of pollution-free or low-polluting vehicles will play an indispensable role in improving the quality of the air environment in Beijing and even the whole of China.

The correlation between air pollutants and economic indicators in Zhangjiakou City is shown in **Table 6**. The relevant factors affecting the air indicators in Zhangjiakou City are different from those in Beijing. From **Table 6**, it can be seen that the most closely related to the air quality index (AQI) are the permanent population (0.916), the industrial added value (0.905), and the total output value of agriculture, forestry, and animal husbandry and fishery (0.89). These three factors are also the main factors affecting $PM_{2.5}$, SO_2, and CO-95_{per}. For the two air indicators PM_{10} and NO_2, the three most relevant factors are permanent population, total output value of agriculture, forestry, animal husbandry and fishery, and urban green area. To a certain extent, it shows that the urban green area plays a vital role in air purification in Zhangjiakou. From the perspective of population size, the number of permanent residents in Zhangjiakou is much less than that in Beijing. The Zhangjiakou area is affected by its geographical location, which has led to a relatively slow opening to the outside world. The overall quality of the population and other aspects of weak environmental protection awareness still need to be improved. For a long time, the rural residents in many counties of Zhangjiakou have used wood or straw to carry out their daily lives. The smoke produced by combustion contains air pollutants, such as hydrocarbons, and it will have a negative impact on the atmospheric environment. In recent years, Zhangjiakou City has worked hard to expand and open up, attracting many industrial enterprises to set up factories in Zhangjiakou to promote the economic development of the region and expand the popularity of the region. However,

industrial enterprises need to emit a large amount of waste gas into the atmosphere to increase industrial output. The emitted fine particles, SO_2, CO, and other pollutants will undoubtedly have a serious impact on the local excellent atmospheric environment. It is worth further discussing how the Zhangjiakou government should change its development focus and how industrial enterprises should reduce harmful substances in exhaust gas. The natural geographical location makes the development of agriculture, forestry, and animal husbandry and fishery an important support for the economy of Zhangjiakou, and agriculture, forestry, and animal husbandry and fishery need to emit a lot of carbon oxide and nitrogen hydride in the daily production or breeding process. Therefore, it is necessary to explore a circular sustainable agriculture, forestry, and animal husbandry and fishery development model that can reduce the emission of air pollutants to a large extent. The area of urban green space can be used as a criterion for evaluating the environmental quality of a city. From the results in **Table 6**, it can be seen that the urban green area of Zhangjiakou City has a high degree of correlation with PM_{10} and NO_2, indicating that the urban green area can have a greater impact on some air pollution indicators. The urban green area of Zhangjiakou City can have a greater positive effect on the absorption and decomposition of particles with larger diameters and nitrogen oxides. The government and social workers in Zhangjiakou also need to increase their energy investment and financial support in the green development of the city, and create a beautiful business card for Zhangjiakou by taking the good opportunity of taking the 2022 Winter Olympics express train.

SUGGESTIONS

Suggestions for Improving Air Quality in Beijing

1) A reasonable adjustment of the urban population layout and the reduction of the urban population density. Beijing is a mega city with a relatively high population density, and its population is the primary factor affecting its air quality. For its part, the government should consider setting up high-tech industrial parks and economic zones in Miyun District and Huairou District in order to ease the population pressure within Beijing, which could promote industrial development while also reducing the population density in central Beijing. The government should guide some Beijing administrative departments to move to Xiong'an New Area in batches, effectively dispersing local Beijing residents and other people coming to Beijing to do business in dense areas, giving full play to the complementary advantages of Beijing–Tianjin–Hebei integration. The government should increase the number of quality schools in the development zones of Beijing to reduce the population density of the schools in the city, and increase the construction of parks, shopping plazas, and fitness equipment around the city to attract people to live in good environmental conditions. In addition, the government should give certain preferential policies to workers and job seekers in Beijing to encourage them to move to Xiong'an New Area, which is the centralized bearer of the non-capital functions of Beijing, in order to further relieve the existing population pressure in Beijing.
2) Continuously optimize the energy structure and encourage the use of clean energy. The state and government should continue to promote clean energy, such as natural gas, as the main source of energy for daily life. In the industrial industry, the government should adopt certain policies to help enterprises gradually replace high pollution and high energy consumption equipment with low pollution and low energy consumption equipment; upgrade sewage equipment; strengthen energy consumption monitoring; recycle water, coal, and industrial raw materials as many times as possible; and take better optimization measures to further improve the efficiency of energy use. Air conditioners and refrigerators are most frequently used in summer. Encouraging the public to buy or switch to fluoride-free air conditioners and refrigerators will not only reduce energy consumption but also reduce the damage to the ozone. Regarding urban residential heating in the winter, the government should strive to popularize natural gas heating, electric heating, geothermal heating, and other clean heating methods as soon as possible. Where possible, streetlights in urban areas should be converted to solar power to reduce the energy consumption of public resources, while wind power stations can be set up in the northwestern part of Beijing as appropriate. The promotion of various clean energy sources will reduce the harmful emissions from the burning of non-renewable resources, which will better promote the low-carbon green development of Beijing and enter the 2022 Beijing Winter Olympic and Paralympic Games with improved environmental quality.
3) Promote green travel for all people and the use of new energy vehicles. The government should actively promote the public taking the subway and bus as much as possible when they go out, which is convenient and environmentally friendly. People who do need to purchase a car should be encouraged to respond to the national call for low carbon and environmental protection by purchasing hybrid or electric cars to meet their daily needs. The promotion of low-carbon travel should be increased on bus billboards and publicity boards in various public places. For new purchases and car exchanges, according to personal circumstances, the recommendation is to purchase electric or hybrid cars for daily city trips and gasoline dual-purpose cars for frequent long-distance business trips. In terms of traffic control, a limiting number system should be applied to fuel cars to reduce traffic flow, while the restrictions on new energy vehicles should be relaxed to motivate people to buy new energy vehicles. The number of refueling stations in the city should be increased along with the number of charging posts in city parking lots to ensure that new energy vehicles can be recharged in time for continuous use. The bus routes in Beijing should be optimized to make it easier for people to travel by avoiding congestion in the morning and evening rush hour bus lanes. The government should call on the

people of the city to create a favorable travel environment for the main venue of the 2022 Beijing Winter Olympic and Paralympic Games, and to fully demonstrate the low-carbon awareness of the people of China to the rest of the world at the Winter Olympics.

Suggestions for Improving Air Quality in Zhangjiakou

1) Actively guide urban and rural residents to enhance their awareness of air protection. In order to maintain the air quality in Zhangjiakou in the long term, it is necessary to further increase the awareness of the people to air protection. This can be done by holding a series of public meetings such as "Environmental protection tips" and "Hazards of air pollution" in neighborhood committees and village committees, calling on the general public to learn more about the importance of air for human survival. Township residents should be encouraged to promote the use of natural gas, solar energy, and other environmentally friendly energy sources instead of the traditional straw burning in order to reduce emissions. Television, radio, the internet, and other information dissemination channels should be used to promote the importance of protecting the air environment. Real and rewarding methods, such as reporting, will attract people to supervise each other, and eventually everyone will consciously restrain themselves. Taking these steps will create a good social atmosphere for the hosting of the Winter Olympics and fully highlight the high-quality self-cultivation level of the Chinese people.
2) Accelerate the optimization and upgrading of industrial industries and attract high-quality enterprises to set up factories. In the industrial industry, the government should continue to deepen structural reform on the supply side, provide increased assistance to industrial enterprises to accelerate the optimization and upgrading of industrial structures, and subsidize a certain percentage of the high-energy consumption and high-pollution equipment after elimination and replacement to encourage enterprises to reduce energy consumption and reduce the emission of harmful air pollutants. Zhangjiakou should also make great efforts to attract investment, focusing on attracting low-pollution processing enterprises such as processing manufacturing and the garment industry to set up factories in Zhangjiakou. This would not only promote the development of the economy of Zhangjiakou but also help to protect the air quality in Zhangjiakou to improve continuously so as to contribute to the smooth operation of the 2022 Winter Olympic Games and Winter Paralympic Games in Zhangjiakou.
3) Ensure stable production in agriculture, forestry, animal husbandry, and fisheries, and strengthen technical guidance and research in the exploration of recyclable models. In agriculture, farmers should be encouraged to use low pollution fertilizers. Drones should be used to spray pesticides instead of traditional manual spraying. Notices should be posted in prominent locations in rural streets to clearly prohibit straw burning. More trees should be planted on some unutilized hills, which would increase the forest cover and have a purifying effect on the air environment. The livestock industry should encourage technical experts to increase the number of field guidance trips in the countryside and actively help herders to carry out scientific breeding and improve the utilization rate of farmland and feed. Animal manure should be stored centrally and regularly transported to septic tanks for treatment. Farmers near grasslands should take full advantage of their location to build horseriding facilities, both to improve their income and reduce the amount of breeding. Construct a comprehensive industrial experimental zone integrating agriculture, forestry, and animal husbandry, give full play to the advantages of agriculture, and provide organic feed. The forestry industry stabilizes the soil and protects it from wind, absorbing exhaust gas, and replenishing sufficient oxygen at the same time. For animal husbandry, in order to reduce the use of chemical fertilizers, a variety of natural fertilizers can be explored to reduce exhaust gas emissions. Efforts should be made to explore a realistic model of multi-resource recyclable use. While these efforts will help ensure the stability of the daily life and livelihoods of the people, they will also help carry out the important initiative of "accelerating the reform of the ecological civilization system and building a beautiful China" emphasized by General Secretary Xi Jinping in the report of the 19th Party Congress and welcome the 2022 Winter Olympics and Winter Paralympics with a strong ecological outlook.

Research Limitations and Prospects

Research limitations

With the continuous acceleration of the social process, the influence of economic factors on the atmospheric environment has become more and more guiding. However, the current economic system is difficult to sort out in the short term due to the many elements involved. Therefore, this study selects as many air and socioeconomic indicators as possible instead of selecting all the indicators as the research object is one of the limitations of this paper. In addition, this paper does not carry out joint research on the air index and social economic index of the area around the two cities.

Prospects

Since the concept of regional coordinated development is valued by the government and society, it is also of great significance to combine the grey correlation analysis model with the new coupling coordination model to study some specific regions. This paper only discusses the relationship between air pollution indicators and social economy from the perspective of time, and it is also interesting to consider combining time and space for analysis in the future. Considering that the grey relational analysis model has unique advantages in factor analysis, it can also be widely used in more fields to give play to the practical value of the model.

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

WW: Conceptualization, Methodology, Software, Visualization, Writing - original draft. YC: Data curation, Investigation, Survey design, Writing - review and editing. YX: Conceptualization, Survey design, Survey translation. LW: Writing - review and editing.

FUNDING

The relevant researches are supported by the National Natural Science Foundation of China (71871084,U20A20316), Young talent support scheme of Hebei Province (360-0803-YBN-7U2C), the Excellent Young Scientist Foundation of Hebei Education Department (SLRC2019001) and the Natural Science Foundation of Hebei Province (E2020402074).

ACKNOWLEDGMENTS

The authors appreciate the valuable suggestions and objective comments made by the reviewers.

REFERENCES

Chen, H., Liu, L., Fang, J., Li, C., Wang, L., Quan, Q., et al. (2021). Spatio-Temporal Analysis of the Coupling Relationship Between Urbanization and Eco-Environment in Backward Regions of China. *Environ. Sci. Pollut. Res.* 28 (45), 1–18. doi:10.1007/S11356-021-16240-Z

Deng, J. (1989). *Grey System Theory*. Wuhan: Huazhong University of Science and Technology Press.

Feng, Q., and Sun, T. (2020). Comprehensive Evaluation of Benefits From Environmental Investment: Take China as an Example. *Environ. Sci. Pollut. Res.* 27 (8), 15292–15304. doi:10.1007/s11356-020-08033-7

Gao, M., and Wu, X. (2017). Analysis of Influencing Factors of Air Quality in Beijing Based on Entropy Weight Grey Correlation Method. *Ecol. Economy.* 33 (3), 142–147.

Gong, M., Zhang, H., Yang, B., and Sun, Q. (2021). Analysis of the Influencing Factors on the Anti-Aging Performance of a Hybrid-Modified Asphalt Mixture Using the Grey Relational Theory. *Int. J. Pavement Eng.* 22 (5), 597–612. doi:10.1080/10298436.2019.1630738

Khuman, A. S., Yang, Y., and John, R. (2019). The Quantification of Subjectivity: The R-Fuzzy Grey Analysis Framework. *Expert Syst. Appl.* 136, 201–216. doi:10.1016/j.eswa.2019.06.043

Li, H., and Wang, X. (2019). Analysis of Air Quality and Influencing Factors in Beijing-Tianjin-Hebei Region From 2013 to 2017. *Adm. Tech. Environ. Monit.* 31 (2), 21–25. doi:10.19501/j.cnki.1006-2009.20190313.012

Li, J., and Ye, S. (2020). Regional Policy Synergy and Haze Governance-Empirical Evidence From 281 Prefecture-Level Cities in China. *Environ. Sci. Pollut. Res.* 28 (9), 10763–10779. doi:10.1007/S11356-020-11251-8

Li, X., Zhang, Q., Zhang, Y., Zheng, B., Wang, K., Chen, Y., et al. (2015). Source Contributions of Urban PM2.5 in the Beijing-Tianjin-Hebei Region: Changes Between 2006 and 2013 and Relative Impacts of Emissions and Meteorology:Changes Between 2006 and 2013 and Relative Impacts of Emissions and Meteorology. *Atmos. Environ.* 123, 229–239. doi:10.1016/j.atmosenv.2015.10.048

Liu, R., Liu, J., Zhang, Z., Borthwick, A. G. L., Cai, Y., Dong, L., et al. (2018). Risks of Airborne Pollution Accidents in a Major Conurbation: Case Study of Zhangjiakou, a Host City for the 2022 Winter Olympics. *Stoch Environ. Res. Risk Assess.* 32 (11), 3257–3272. doi:10.1007/s00477-018-1590-5

Luo, X., Lu, L., Wang, Z., and Yang, L. (2019). Gray Correlation Analysis of Energy Consumption, Environmental Pollution, and Economic Growth in Subtropical Regions of China: Guangxi and Zhejiang as Examples. *Trop. Conservation Sci.* 12, 194008291984810. doi:10.1177/1940082919848101

Nan, G., and Sun, H. (2017). Analysis of Factors Affect-Ing O_3 Concentration in Shaanxi Province Based on Grey Relational Model. *Acta Scie-ntiae Circumstantiae.* 37 (12), 4519–4527. doi:10.13671/j.hjkxxb.2017.0313

Ni, H. (2013). *Assessment of Air Quality and Analysis of Influencing Factors in Wuhan Based on Grey Model*. Wuhan: Huazhong University of Science and Technology.

Qazi, T. F., Shaukat, M. Z., Niazi, A. A. K., and Basit, A. (2021). Evaluating the Immediate Response of Country-Wide Health Systems to the Covid-19 Pandemic: Applying the Gray Incidence Analysis Model. *Front. Public Health.* 9, 9. doi:10.3389/FPUBH.2021.635121

Tian, Y., Jiang, Y., Liu, Q., Xu, D., Zhao, S., He, L., et al. (2019). Temporal and Spatial Trends in Air Quality in Beijing. *Landscape Urban Plann.* 185, 35–43. doi:10.1016/j.landurbplan.2019.01.006

Wang, C., Wang, G., and Wu, L. (2019). Study on Influencing Factors of Air Quality in Ha-Ndan Based on Grey Relational Analysis Fr-Om the Perspective of Economic and Soci-Al Indicators. *Maths. Pract. Theor.* 49 (17), 151–155.

Wang, J., Qian, W., Du, J., and Liu, Y. (2020). Effective Allocation of Resources in Water Pollution Treatment Alternatives: a Multi-Stage gray Group Decision-Making Method Based on Hesitant Fuzzy Linguistic Term Sets. *Environ. Sci. Pollut. Res.* 27 (3), 3173–3186. doi:10.1007/s11356-019-07265-6

Wang, L., and Tian, D. (2019). Analysis of Influencing Factors of Air Quality in Yichang Based on Grey Relational Analysis Method. *Energ. Environ. Prot.* 33 (5), 60–64.

Wang, O., and He, B. (2018). Study on the Relationship Between PM 2.5 and Six Elements of Ambient Air Quality in Urumqi Based on Grey Relational Grade. *J. Arid Land Resour. Environ.* 32 (6), 176–181. doi:10.13448/j.cnki.jalre.2018.191

Wang, X. (2020). Analysis of Influencing Factors of Smog in Zhengzhou Based on Improve-D Grey Relational Grade. *Maths. Pr-actice Theor.* 50 (01), 143–153.

Wang, Y., Shi, M., Lv, Z., Liu, H., and He, K. (2021). Local and Regional Contributions to PM2.5 in the Beijing 2022 Winter Olympics Infrastructure Areas During Haze Episodes. *Front. Environ. Sci. Eng.* 15 (6), 140. doi:10.1007/S11783-021-1434-2

Wu, H., and Qu, Z. (2020). Gray Clustering Model Based on the Degree of Dynamic Weighted Incidence for Panel Data and its Application. *Grey Systems: Theory and Application.* 10 (4), 413–423. doi:10.1108/GS-09-2019-0040

Wu, L., Guo, X., and Chen, Y. (2021). Grey Relational Entropy Calculation and Fractional Prediction of Water and Economy in the Beijing-Tianjin-Hebei Region. *J. Maths.* 2021, 1–16. doi:10.1155/2021/4418260

Xiao, Q., Shan, M., Gao, M., Xiao, X., and Guo, H. (2021). Evalua-tion of the Coordination Between China's Technology and Economy Using a Grey Mul-Tivariate Coupling Model. *Technol. an-d Econ. Develop. Economy.* 27 (1), 24–44. doi:10.3846/tede.2020.13742

Xiao, Q., Shan, M., Xiao, X., and Rao, C. (2020). Evaluation Model of Industrial Operation Quality Under Multi-Source Heterogeneous Data Information. *Int. J. Fuzzy Syst.* 22 (2), 522–547. doi:10.1007/s40815-019-00776-x

Xiao, X. (1997). Theoretical Research and Comments on Grey Relational Metric Model. *Syst. Engineering-Theory Pract.* 6 (8), 76–81. doi:10.3321/j.issn:1000-6788.1997.08.014

Yan, D., Lei, Y., Shi, Y., Zhu, Q., Li, L., and Zhang, Z. (2018). Evolution of the Spatiotemporal Pattern of PM2.5 Concentrations in China - A Case Study From the Beijing-Tianjin-Hebei Region. *Atmos. Environ.* 183, 225–233. doi:10.1016/j.atmosenv.2018.03.041

Zhan, S., Kuang, Y., and Ruan, Z. (2018). Analysis of Influencing Factors of Air Quality in Gua-Ngdong-Hong Kong-Macao Greater Bay Ar-Ea Based on Grey

Correlation Degree. *Jo-urnal Tsinghua University (Science Technology)*. 58 (8), 761–767. doi:10.16511/j.cnki.qhdxxb.2018.26.031

Zhang, X., Shi, M., Pang, R., and Xiang, N. (2018). Correlating PM2.5 Concentrations With Air Pollutant Emissions: A Longitudinal Study of the Beijing-Tianjin-Hebei Region:A Longitudinal Study of the Beijing-Tianjin-Hebei Region. *J. Clean. Prod.* 179, 103–113. doi:10.1016/j.jclepro.2018.01.072

Zhu, L., Zhao, C., and Dai, J. (2021). Prediction of Compressive Strength of Recycled Aggregate concrete Based on Gray Correlation Analysis. *Construction Building Mater.* 273, 121750. doi:10.1016/J.CONBUILDMAT.2020.121750

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.

Copyright © 2021 Wang, Chen, Xie and Wu. 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 in Environmental Science

ORIGINAL RESEARCH
published: 15 December 2021
doi: 10.3389/fenvs.2021.761287

Hybrid Data Mining Forecasting System Based on Multi-Objective Optimization and Selection Model for Air pollutants

Yanwen Huang[1], Yuanchang Deng[1], Chen Wang[1]* and Tonglin Fu[2,3,4]*

[1]School of Intelligent System Engineering, Sun Yat-Sen University, Shenzhen, China, [2]School of Mathematics and Statistics LongDong University, Qingyang, China, [3]Shapotou Desert Research and Experiment Station, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou, China, [4]University of Chinese Academy of Sciences, Beijing, China

The air quality index (AQI) indicates the short-term air quality situation and changing trend of the city, which includes six air pollutants: $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2 and O_3. Due to the diversity of pollutants and the fluctuation of single pollutant time series, it is a challenging task to find out the main pollutants and establish an accurate forecasting system in a city. Previous studies primarily focused on enhancing either forecasting accuracy or stability and failed to analyze different air pollutants at length, leading to unsatisfactory results. In this study, a model selection forecasting system is proposed that consists of data mining, data analysis, model selection, and multi-objective optimized modules and effectively solves the problems of air pollutants monitoring. The proposed system employed fuzzy C-means cluster algorithm to analyze 13 original AQI series, and fuzzy comprehensive evaluation is used to find out the main air pollutants in each city. And then multiple artificial neural networks are used to forecast the main air pollutants for each category and find the optimal models. Finally, the modified multi-objective optimization algorithm is used to optimize the parameters of optimal models and model selection to obtain final forecasting values from optimal hybrid models. The experiment results of datasets from 13 cities in the Beijing–Tianjin–Hebei Urban Agglomeration demonstrated that the proposed system can simultaneously obtain efficient and reliable data for air quality monitoring.

Keywords: air quality index, data analysis, data mining, artificial neural networks, model selection

OPEN ACCESS

Edited by:
Yan Hao,
Shandong Normal University, China

Reviewed by:
Qingli Dong,
Dalian University of Technology, China
Xiaofei Luo,
University of Aizu, Japan

****Correspondence:***
Yuanchang Deng
dengych@mail.sysu.edu.cn
Chen Wang
wangch339@mail.sysu.edu.cn

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: *19 August 2021*
Accepted: *04 October 2021*
Published: *15 December 2021*

Citation:
Huang Y, Deng Y, Wang C and Fu T (2021) Hybrid Data Mining Forecasting System Based on Multi-Objective Optimization and Selection Model for Air pollutants. Front. Environ. Sci. 9:761287. doi: 10.3389/fenvs.2021.761287

INTRODUCTION

In recent years, air pollution has received increasing attention due to the negative effects, such as respiratory diseases, that it has on human health (Jiang et al., 2017). Simultaneously, air pollution is a growing environmental concern, responsible for approximately 2 million premature deaths per year worldwide (World Health Organization, 2008). A report issued by the World Health Organization (WHO) acknowledges that air pollution is one of the biggest health risks (Xu et al., 2016). Since the industrial revolution, many countries have focused on economic development while ignoring air quality, and incidents that cause harm are everywhere. In 1930, the Mas Valley event in Belgium caused nearly 60 deaths in a week. In the 1940s, the smog incident in Los Angeles caused many people to have red eyes, pharyngitis, respiratory disease deterioration, and even confusion and pulmonary edema. In 1948, the American Donora incident caused 5,911 people to become violent.

The most serious is the well-known London smog event of 1952—more than 4,000 deaths in 4 days and more than 8,000 deaths in 2 months. In addition, air pollution in China is also quite serious. The previous results in 2009 showed that the air quality index (AQI) in 107 cities of China did not meet the country's national air quality standards (NAAS) (Zhou et al., 2014). In addition, 7 of the 10 most polluted cities in the world are in China. According to the World Bank, China loses 10% of its gross domestic product each year due to air pollution. Air pollution is also associated with elevated rates of mortality, causing between 350,000 and 500,000 premature deaths each year in China (Shanshan et al., 2014). Air pollution has become the fourth leading health risk factor for China after smoking, diet, and obesity (Zhang et al., 2018). In order to reduce the losses caused by air pollution, several health and governmental institutions gather and publish data regarding what is known as AQI to inform people about the state of air pollution. For instance, the European Environment Agency (EEA) and the European Commission (EC) have launched, in 2017, an online platform that provides information about current air quality situation based on measurements from more than 2,000 air quality monitoring stations across Europe (Akyüz and Çabuk, 2009). In addition, China's environmental supervisors have also issued some plans and programs, including EIA (Environmental Influence Assessment) and Emergency Response for reducing air pollution. Since 2013, China has also begun to evaluate the quality of air through AQI values and graded the city's air quality by AQI values. AQI is an important evaluation indicator that comprehensively reflects the air pollution status related to human health. Through the use of the AQI it was possible to synthesize, in a single daily value, concentrations of major pollutants in urban areas (NO_2, O_3, CO, SO_2, $PM_{2.5}$, PM_{10}) for the entire period (Feng et al., 2015). The greater the AQI value, the more serious the air pollution. But real-time air quality monitoring can no longer meet people's needs. Like weather forecasts, people also long for air quality prediction to arrange their activities and take protective measures in advance (Hao et al., 2021).

Obviously, if we can provide early warning before the hazard occurs, based on a good air quality early-warning system, these losses might be avoided by taking effective corresponding protection measures. In order to establish an effective air warning system, observation and control of air quality is the key issue for authorities. The most significant point in any kind of air pollution control system is to be able to detect increasing (deterioration) or decreasing (improvement) trends (Hao and Tian, 2018). Unfortunately, because air quality data is obtained in limited time and space, its incompleteness and non-stationarity may result in low accuracy and poor stability of the forecasting results (Hao et al., 2019). Therefore, the prediction of AQI or other pollution indicators is a challenging task.

In recent years, many studies on air quality have focused on the prediction of atmospheric pollutant concentrations. From the angle of methodology, various quantitative prediction methods of the atmosphere pollutant concentrations can be classified into two categories, including deterministic models and empirical models (Steffens et al., 2017). The deterministic model is mainly the chemical transport model (CTM), which is based on the fundamental principles of simulating atmospheric physics and chemistry that involve transportation, emissions, and conversion processes in air pollution (Rivas et al., 2018). The forecasts are used to support flight planning by enabling the representation of important three-dimensional (3-D) atmospheric chemical structures (such as dust storm plumes, polluted air masses originated by large cities, and widespread biomass burning events) and their time evolution, which are often research targets to be detected and investigated through specific flight plans (Latif et al., 2018). Various models have been proposed to identify the interactions between various air pollutants and their emission sources (Yang and Wang, 2017). Nonetheless, due to the incomplete knowledge and understanding of the sources, dispersion and sinks of pollutants, transport processes, and atmospheric chemicals, there are some significant uncertainties in the models, resulting in air pollutant concentrations being among the most difficult to forecast accurately using CTMs (Liu et al., 2008). Therefore, CTM forecasts are less accurate than empirical air quality predictive models that are trained with local meteorological data and air quality.

A large number of empirical models include statistical models and machine learning models for the forecast of atmospheric pollutant concentrations. Common statistical models for air quality prediction include autoregressive (AR) models, moving average (MA) models, autoregressive integrated moving average (ARIMA) models, and multiple linear regression (MLR) models. For example, Zhang et al. (2018) applied the RIMA model to predict the concentration of $PM_{2.5}$ based on time series air quality data covering two warm periods and two cold periods and concludes that $PM_{2.5}$ concentration is higher in the cold period and lower in the warm period. MLR models are applied by Mehmet Akyüz et al. (Pereira et al., 2018) to forecast the concentration of individual pollutants. The study also considers the effects of contaminant concentrations and other meteorological parameters. Although Box–Jenkins Time Series (ARIMA) and MLR models have been applied to air quality forecasting in urban areas, they have limited accuracy owing to their inability to predict extreme events, and they are not applicable when performing long-term prediction and nonlinear sequence prediction.

On the contrary, artificial neural networks (ANNs) are more popular for their no-linear systems, especially when it is difficult to determine the theoretical models (Lanzafame et al., 2015). Díaz-Robles et al. (2008) combined a new hybrid model of ARIMA and ANN to improve the prediction accuracy of areas with limited air quality and meteorological data. Xiao Feng and Qi Li et al. (Feng et al., 2015) combined air mass trajectory analysis and wavelet transform and proposed that ANN predicts the daily average concentration of pollutants 2 days in advance, improves the accuracy of prediction, and is superior to other models. However, they also have certain shortcomings that may fall into local optimum or over-fitting, which may result in poor prediction.

Any model has its inevitable shortcomings, and due to the advent of the world's big data era, data mining techniques such as

decomposition methods (Güçlü et al., 2019), feature selection techniques (Pan et al., 2011), and optimization algorithms (Liu et al., 2019) combined with artificial intelligence technology are more operational. Therefore, with consideration of forecast accuracy, hybrid models which combine a new method with artificial intelligence are of great significance in air quality forecasting field (D'Allura et al., 2011). Although the construction of the combined model is usually based on actual problems to achieve the expected test objectives, there are still some problems that most of the past studies have focused on improving the prediction accuracy of the model while ignoring the stability of the model prediction. Many optimization algorithms inspired by nature including cuckoo algorithm (Urbancok et al., 2017), firefly algorithm (Bessagnet et al., 2019), bat algorithm (Liu et al., 2018), and particle swarm optimization algorithm (Kumar et al., 2019) have been developed to solve single-objective problems in recent years. However, real-world optimization problems always involve multiple objectives and so-called multi-objective optimization, which means, in this case, the solutions for a multi-objective problem, which is the main focus of the algorithm, represent the trade-offs between the objectives due to the nature of such problems (Shenfield and Rostami, 2015). The developed multi-objective optimization algorithm has been applied more and more widely in the fields of finance (Li et al., 2019) and mechanical engineering (Dhiman and Kumar, 2018). The atmosphere is a highly complex dynamic system. The air quality data sequence usually has characteristics such as non-stationarity and nonlinearity; thus, the multi-objective optimization algorithm is a suitable choice.

Furthermore, air quality assessment algorithms are developed to assess air quality and protect human health from air pollution and play a vital role in air quality warning systems. The early-warning system can increase the environmental consciousness of society and protect the public against hazardous air quality. It can also aid the relevant departments to better control air pollution and avoid negative social, economic, and environmental impacts. According to the aforementioned analysis, developing a novel and robust air quality early-warning system has become highly desirable for society. Therefore, a variety of models are employed in air quality assessment, including mobile monitoring (Li et al., 2018), CFD-RANS simulation (Lauriks et al., 2020), principal component regression (PCR), sensitivity analysis (Kim et al., 2018), Bayesian models (Han et al., 2021), support vector machines (SVM) (Leong et al., 2019), ANNs (Davood et al., 2021), and fuzzy techniques (Dass et al., 2021). However, although the air quality warning system has important practical significance to the public in other fields, China's research in this field is still relatively small.

Looking back at the previous literature on air quality forecasting research, the shortcomings of the traditional air quality forecasting models are summarized as follows: 1) the large amount of information required by the CTM model leads to uncertainty in the forecasting. 2) The single statistical models with low forecasting accuracy cannot meet the requirements of air quality forecasting. 3) In the past, many air quality studies focused on eliminating the effects of noise on data processing and less on the feature extraction of data. 4) It is easy for single-objective optimization algorithm commonly used to fall into local optimum and over-fitting, resulting in poor stability. 5) In addition, previous studies on air quality have focused on air quality forecasting, while the research on air quality assessment was relatively rare.

Based on the above analysis, it is necessary to overcome these deficiencies and develop a novel and robust air quality warning system. The evaluation–forecast system developed in this study consists of two parts: evaluation and forecasting. The evaluation part involves feature extraction and finding out the main air pollutants; in the forecasting part, a new metric is developed to find the optimal model in each category, and optimal forecasting models are optimized with modified gray wolf optimization (DEGWO) optimization algorithm and leave-one-out deciding weight strategy to improve the accuracy of forecasting results and provide support for early warning systems. The specific implementation steps of the hybrid forecasting system are as follows: First, the feature extraction of the original data is performed to find similar attributes of AQI time series according to the relevant theory of fuzzy C-mean cluster.

Moreover, air quality evaluation based on the forecasting results of air pollutant concentration plays a crucial role in the development of the air quality warning system.

In this paper, in view of the uncertainty and ambiguity of each air pollutant, the fuzzy comprehensive evaluation is applied in AQI. According to the implementation of fuzzy comprehensive evaluation results, finding out the main pollutants in each city is another important part of this work. Next, we use long short-term memory (LSTM), backpropagation neural network (BPNN), adaptive network-based fuzzy inference system (ANFIS), generalized regression neural network (GRNN), and SVM models to forecast the main air pollutants time series, and a developed new metric is used to select optimal forecasting model. Finally, all these individual forecasting models' predictors based on the leave-one-out deciding weight strategy are optimized by the DEGWO optimization algorithm, and the final forecasting results are obtained. Therefore, the combination of these methods will result in more accurate forecasts and assessments performance, providing significant advantages for the construction and implementation of early warning systems for detecting air quality. The main contributions of this paper are as follows:

1) The fuzzy comprehensive evaluation is established for six air pollutants, which calculates the fuzzy membership degree of each pollutant and determines the main pollutants of each city.
2) A model selection index is established to select the optimal forecasting model from different neural network models. Based on model selection, the established weighted information criterion can select the optimal forecasting model for $PM_{2.5}$, PM_{10}, and NO_2 forecasting.
3) The forecasting performance of the optimal single model is improved. In the forecasting process, an improved multi-objective optimization algorithm is used to optimize the parameters of the single forecasting model, which not only

improves the prediction accuracy but also improves the stability of the single model

4) The model selection index is used to select the optimal forecasting value from the optimal hybrid model.

METHODOLOGY

In this subsection, the relative methods are presented in detail, including the data mining technique, forecasting model, and the DEGWO) algorithm. Subsequently, the marching process of our developed combined model is demonstrated.

Forecasting Model

Five typical models, namely, the multilayer perceptron (MLP) (You et al., 2017), ANFIS (Jang, 1993), LSTM (Muzaffar and Afshari, 2019), SVM (Brereton and Lloyd, 2010), and GRNN (Land and Schaffer, 2020), have been widely used for air pollutants forecasting because of their robustness, efficiency, and accuracy.

Modified Gray Wolf Optimization (DE-GWO)

For the DE algorithm and gray wolf optimization (GWO) algorithm, the defects of prematurity, poor stability, and ease in falling into local optimum will occur when solving the optimization problem separately. Combining the advantages and disadvantages of the two algorithms, a more efficient hybrid optimization algorithm, DEGWO algorithm, is proposed to improve global search capabilities.

Firstly, in order to avoid the phenomenon in which the population is iteratively reduced to a certain area, the crossover and selection operations of the DE algorithm are used to maintain the diversity of the population. Then, as the initial population of the GWO algorithm, the objective function value of the individual is calculated. The optimal three individuals X_α, X_β, and X_δ are selected to update the positions of other gray wolf individuals. Then, the position of the gray wolf individual is updated by the intersection and selection operations of DE, and the iterative update is repeated until the optimal one is selected. The target function value is output.

The hybrid algorithm not only improves the global search ability but also effectively avoids the defects of early maturity stagnation and falling into local optimum. The specific implementation steps of the algorithm are as follows:

Step 1: Set the relevant parameters of the hybrid optimization algorithm, population size N, maximum iteration number t_{max}, crossover probability CR, search dimension D, search range ub, lb, and scaling factor range F.

Step 2: The parameters a, A, and C are initialized, and the DE variant operation is performed on the population individual according to **Eq. 1** to generate an intermediate; an initial population and the number of iterations is set to $t = 1$.

$$v_{ij}(g+1) = \begin{cases} h_{ij}(g), & rand(0,1) \le CR \quad or \quad j = rand(1,n) \\ x_{ij}(g), & rand(0,1) > CR \quad or \quad j \ne rand(1,n) \end{cases} \tag{1}$$

Then the competition selection operation is performed according to **Eq. 2** to generate.

$$x_i(g+1) = \begin{cases} v_i(g), & f[v_i(g+1)] < f[x_i(g+1)] \\ x_i(g), & f[v_i(g+1)] \ge f[x_i(g+1)] \end{cases} \tag{2}$$

Step 3: Calculate the objective function value of each gray wolf individual in the population, sort according to the size of the objective function value, and select the optimal first three individuals as X_α, X_β, and X_δ, respectively.

Step 4: Calculate the distance between other gray wolf individuals in the population and the optimal X_α, X_β, and X_δ according to **Eqs 3–5**.

$$D_\alpha = |C_1 X_\alpha(t) - X(t)| \tag{3}$$

$$D_\beta = |C_2 X_\beta(t) - X(t)| \tag{4}$$

$$D_\delta = |C_3 X_\delta(t) - X(t)| \tag{5}$$

Finally, update the current position of each gray wolf individual according to **Eqs 6–9**.

$$X_1(t+1) = X_\alpha(t) - A_1 D_\alpha \tag{6}$$

$$X_2(t+1) = X_\beta(t) - A_2 D_\beta \tag{7}$$

$$X_3(t+1) = X_\delta(t) - A_3 D_\delta \tag{8}$$

$$X_p(t+1) = \frac{X_1 + X_2 + X_3}{3} \tag{9}$$

Step 5: Update the values of a, A, and C in the algorithm, cross-operate the position of the individual population according to **Eq. 1**, retain the better components, then perform **Eq. 2** to select new individuals and calculate the objective function values of all gray wolf individuals.

Step 6: Update the positions of the top three gray wolf individuals X_α, X_β, and X_δ.

Step 7: Determine the count value. If the maximum iteration number t_{max} is reached, the algorithm exits and, based on **Eq. 10**, outputs the multi-objective function value of the global optimal X_α; otherwise, let $t = t + 1$, and then go to Step 3 to continue execution.

$$\min f = f_1(x) + f_2(x)$$
$$s.t \begin{cases} f_1(x) = \frac{1}{n}\sum_{i=1}^{n} (error_i)^2 \\ f_2(x) = \frac{1}{n}\sum_{i=1}^{n} (error_i - \overline{error})^2 \end{cases} \tag{10}$$

Fuzzy C-Means Clustering

Fuzzy C-means clustering (FCM), known as fuzzy ISODATA, is a clustering algorithm that uses membership degrees to determine the extent to which each data point belongs to a certain cluster. In 1973, Bezdek proposed the algorithm as an improvement to the early hard C-means clustering (HCM) method (Gayen and Biswas, 2021). The clustering steps are as follows:

Step 1: Initialize the membership matrix U with a random number whose value is between 0 and 1, so that it satisfies the constraint in **Eq. 11**.

$$\sum_{i=1}^{c} u_{ij} = 1, \forall_j = 1, \ldots, n \tag{11}$$

Step 2: Calculate c cluster centers c_i ($i = 1, c$) using **Eq. 12**.

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{12}$$

Step 3: Calculate the value function according to **Eq. 13**. If it is less than a certain threshold, or if the amount of change from the value of the last value function is less than a certain threshold, the algorithm stops.

$$\mathrm{J}(\mathrm{U}c_1, ..., c_c) = \sum_{\mathrm{i}=1}^{\mathrm{c}} J_i = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^{m} d_{ij}^{2} \tag{13}$$

Step 4: Calculate the new U matrix with **Eq. 14**. Go back to Step 2.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{(m-1)}}} \tag{14}$$

Fuzzy Synthetic Evaluation Theory

The process of establishing a fuzzy synthetic evaluation (FSE) system is as follows (Lu et al., 2011).

Step 1. The set of factors for the evaluation object is determined.

The selected factors should possess the traits of representativeness, feasibility, and system. Air quality evaluation relies on the concentration levels of the main air pollutants. Therefore, in this study, the indicators were chosen according to China's ambient air quality standards (AAQS: GB3095-2012). Moreover, different geographical areas have different topographic and economic characteristics, and consequently, the different key pollutants in the study areas should be also considered.

Step 2. The evaluation rank standard is determined.

The evaluation rank set is described as $\mathrm{V} = \{v_1, v_2, \cdots, v_n\}$. In our study, the air pollution degrees were divided into five levels. The pollutants grading standard according to AAQS is shown in **Supplementary Appendix S1**.

Step 3. Index fuzzification.

In this step, the membership functions (MFs) corresponding to each index are obtained. The process of fuzzification constitutes the process of membership calculation by using MFs. In this study, we used the trapezoidal membership to calculate the membership value.

Step 4. The factor weight is calculated.

Weight reflects the importance of each factor in synthetic evaluation and directly affects the outcome of the evaluation. Many methods exist for determining the weight, such as weighted statistics, coefficient of variation method, the Delphi method, and entropy methods. In our study, the weight was calculated by fuzzy weighting method.

Step 5. The evaluation results are output.

The objective function of the DEGWO algorithm is based on stability and accuracy, in which MSE is the standard to measure accuracy and the variance of error is the standard to measure stability. **Algorithm 1** briefly outlines the process of the MODEGWO.

Algorithm 1. *MMODA*

Input: Objective function Min $fitness(x) = f_1 + f_2$

$$min \begin{cases} f_1(x) = MSE = \dfrac{1}{N}\sum_{i=1}^{N} (\hat{x}_i - x_i),\ i = 1, 2, ..., N \\ f_2(E) = Var(E) = \dfrac{1}{N}\sum_{i=1}^{N} \left(E_i - \bar{E}\right)^2,\ i = 1, 2, ..., N \end{cases}$$

Note: E_i is the test error; the calculation equation is $E_i = \hat{x}_i - x_i$

$\hat{x}_i$ and x_i are the actual data and output data by each model

Parameters of DEGWO

CR is crossover probability: 0.2;

MaxGen is the maximum number of the iteration: 500;

F is the scaling factor: [0.2, 0.8];

$\mathrm{p_{size}}$ is population size: 50.

Output: The optimal solution and the best objective function value.

Initialize a parent population, mutant population, and child population of gray wolf with a random position in a feasible region using equation;

$$x_p^k = x_p^k(low) + \left(x_p^k(up) - x_p^k(low)\right) \times rand(0, 1)$$

Note: $x_p^k(low)$ is the lower bound of the pth component of the kth individual.

$x_p^k(up)$ is the upper bound of the pth component of the kth individual.

rand(0, 1) represents a random number in [0, 1].

$p = 1, 2\cdots$, d. $k = 1, 2\cdots, \mathrm{p_{size}}$

Initialize crossover probability Pc and scaling factor F;

Initialize a, A, and C;

Evaluate f for all individuals in the parent population;

Sort the parent population in a non-decreasing order, according to the objective function value;

X_α is the best individual in the parent population of gray wolves;

X_β is the second individual in the parent population of gray wolves;

X_δ is the third individual in the parent population of gray wolves;

While ($t < MaxGen$)

for each individual in the parent population of gray wolves

Update the position using the following equation;

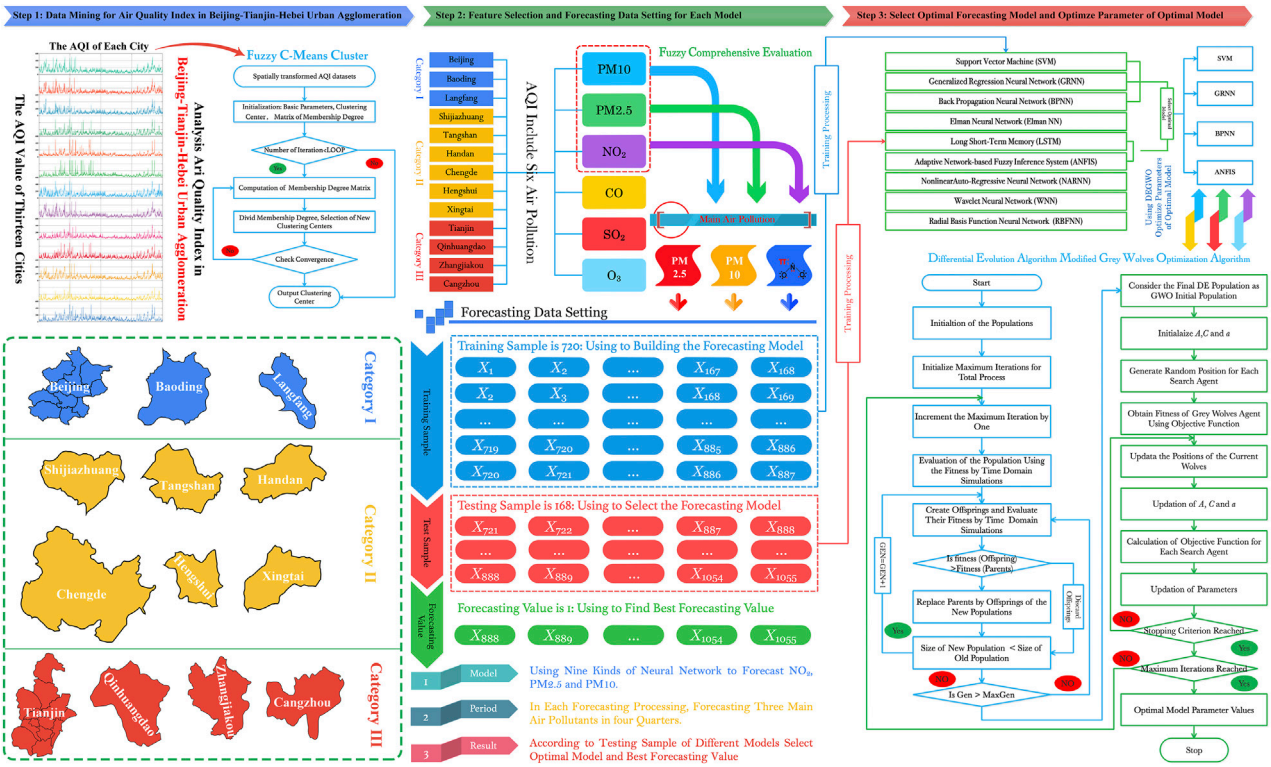


FIGURE 1 | Flowchart of air quality index forecasting system for Beijing–Tianjin–Hebei Urban Agglomeration.

$$X_p(t+1) = \frac{X_1 + X_2 + X_3}{3}$$

end for

Obtain a mutant population of gray wolves using the following equation;

$$V^i(g) = X^{r1} + F \cdot \left(X^{r2}(g) + X^{r3}(g)\right)$$

Note: g is the generation number,

F is the scaling factor, $g = 0, 1, 2, \cdots$, $MaxGen$,

$MaxGen$ is the maximum number of the iteration

Obtain a child population of gray wolves using the following equation;

$$v_{ij}(g+1) = \begin{cases} h_{ij}(g), rand(0,1) \le CR & j = rand(1,n) \\ x_{ij}(g), rand(0,1) > CR & j = rand(1,n) \end{cases}$$

CR represents the crossover probability

for each individual $Parent_i$ in a parent population of gray wolves

If $f(Child_i) < f(Parent_i)$

Replace $Parent_i$ with $Child_i$

end if

Update A, C, and a;

Sort the parent population of gray wolves in a nondecreasing order;

Update X_α, X_β, X_δ;

$t = t + 1$;

End while

Return $Parent\alpha$ and $f(Parent\alpha)$.

Formulation of the Hybrid Model

The hybrid AQI forecasting system in this paper is composed of the above three parts. A flow chart of the hybrid model is presented in **Figure 1**.

From the above, we can see that the AQI forecasting step using the hybrid forecasting system proceeds as follows:

Step 1: Data Mining

1. Collect the original data in the proposed hybrid forecasting model. Specifically, the average hourly AQI and six air pollutants are utilized as experiment data in this work.
2. Using data mining technology, 13 cities in Beijing–Tianjin–Hebei Urban Agglomeration (BJ-TJ-HE) are clustered, the characteristics of each category are summarized, and each category is further analyzed.

The AQI and six air pollution time series with missing points is filling processed by shape-preserving piecewise cubic spline interpolation, which maintains the continuity of each time series.

Step 2: Feature selection and data setting for each model

1. Feature selection: According to the result of cluster, establish fuzzy comprehensive evaluation for six pollutants and find out the main air pollutants of each in the same category.
2. Data setting: Each main air pollutants time series can be divided into three parts: training sample and testing samples for the forecasting values. The training sample is used to construct and train the ANNs, which in this work consist of a BPNN, SVM, GRNN, LSTM, and ANFIS. In addition, the testing sample is used to select the optimal model. For this, the WIC values of the ANNs are calculated, and the best model in terms of the WIC is selected. The input data are used to train the ANNs before calculating the forecasting value, with 1–6 input nodes and 1–30 hidden nodes. According to the value of WIC, the best forecasting model and best structure are chosen.

Step 3: Optimize the parameters of the best forecasting model.

To ensure the forecasting performance, a modified optimization algorithm is used to further optimize the parameters of the best forecasting model (expect LSTM). Finally, the main air pollutants forecasting results are obtained and compared with those of different hybrid forecasting models.

EXPERIMENT DESIGN AND ANALYSIS

In this section, the specific information of experiment datasets in BJ-TJ-HE are described in detail. Eight performance metrics are applied to assess the performance of the proposed model. The experiments conducted in this study were implemented on Matlab 2018a, and the specifications of the hardware were as follows: Intel Core i9-7920X 2.90 GHz CPU and 32 GB RAM.

Data Description

The BJ-TJ-HE is the national capital region of the People's Republic of China. It is the biggest urbanized megalopolis region in Northern China, where Beijing, Tianjin, Baoding, and Langfang are the central core areas of BJ-TJ-HE. In this paper, the 13 cities of BJ-TJ-HE are evaluated to develop an early warning indicator for air quality. The datasets of hourly concentrations of the six major air pollutants used in this study are retrieved from the website of the China National Environment Monitoring Centre (http://www.cnemc.cn/sssj/). The first dataset includes AQI hourly concentrations collected from January 1, 2017, to December 31, 2018, in BJ-TJ-HE. **Figure 2** shows the result of fuzzy C-mean cluster, which displays the construction of a fuzzy matrix based on the attributes of AQI in 13 cities and objectively and accurately cluster (Category I: Beijing, Baoding, Langfang; Category II: Shijiazhuang, Tangshan, Handan, Chengde, Hengshui, Xingtai; Category III: Tianjin, Qinhuangdao, Zhangjiakou, Cangzhou). The result of fuzzy comprehensive evaluation is shown in **Table 1**, which found that the main air pollutants are PM_{10}, PM_2, and NO_2 in 13 cities.

According to the analysis in **Table 1**, the main air pollutants from statistical analysis of BJ-TJ-HE are NO_2, $PM_{2.5}$, and PM_{10} shown in **Table 2**, in which the average value of the main air pollutants shows obvious differences among the 13 cities. The average value of NO_2 in the different cities is between 22.2525 and 49.4348 $\mu g/m^3$, in which the average value in Xingtai is higher than in the other cities. At the same time, the $PM_{2.5}$ and PM_{10} average values in Xingtai are 69.6938 and 135.8368 $\mu g/m^3$, which are also higher than in the other cities. The maximum values of NO_2, $PM_{2.5}$, and PM_{10} were in Hengshui, Baoding, and Zhangjiakou, with values of 215, 402, and 1581 $\mu g/m^3$, and the minimum values of the three main air pollutants were in Zhangjiakou, Beijing, and Zhangjiakou, with values of 1, 3, and 12 $\mu g/m^3$.

In terms of skewness, all data sets are rightward, with values of skewness are greater than 0. For the values of kurtosis, only three data sets of NO_2 were less than 3, which meant that these three sets (Qinhuangdao, Shijiazhuang, and Xingtai) had a fat tail. At the same time, the other data sets had a thin tail.

Forecasting Metric

This section focuses on the efficiency of the different forecasting model with respect to computational performance. Eight evaluation criteria are applied to estimate the forecasting performance, namely, mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), Theil U statistic 1 (U1), and Theil U statistic 2 (U2) were calculated for all the fits; the goodness of forecasting fit (R^2) and the standard of forecasting error (STDE) indicates the stability of the forecasting models; and the direction accuracy (DA) evaluates the optimal decision-making, often relying on correct forecasting directions or turning points between the actual and forecasting values. These performance metrics are defined in **Table 3**.

Experiment Preparation: Model Selection

In the forecasting processing, there is no model that can be applied to all time series in the process of forecasting. Therefore, in this paper we developed a new metric, which measures accuracy of each hybrid model testing set and determines whether the model can provide the optimal forecasting value. The process of model selection is as follows:

Each model data is divided into 840 training samples, 168 testing samples, and one forecasting value. The accuracy of the testing sample is calculated by using the WIC. In order to eliminate the difference of the order of magnitude of forecasting metric, the MAE, MAE RMSE, MAPE, STDE, U1, and U2 are normalized. The calculation formula is as follows:

$$WIC = NMAE + NRMSE + NMAPE + NSTDE + U1 + U2 + R2 + (1 - \mathrm{DA})$$

For the first forecasting, the 1st to 840th samples are the training samples, the 841st to 1008th samples are the testing samples, and the 1009th sample is the forecasting value. At the end of the forecasting, the WIC value of the testing sample is calculated. If the WIC of the *i*th model is the smallest, the forecasting value of the *i*th model provides the optimal forecasting value.

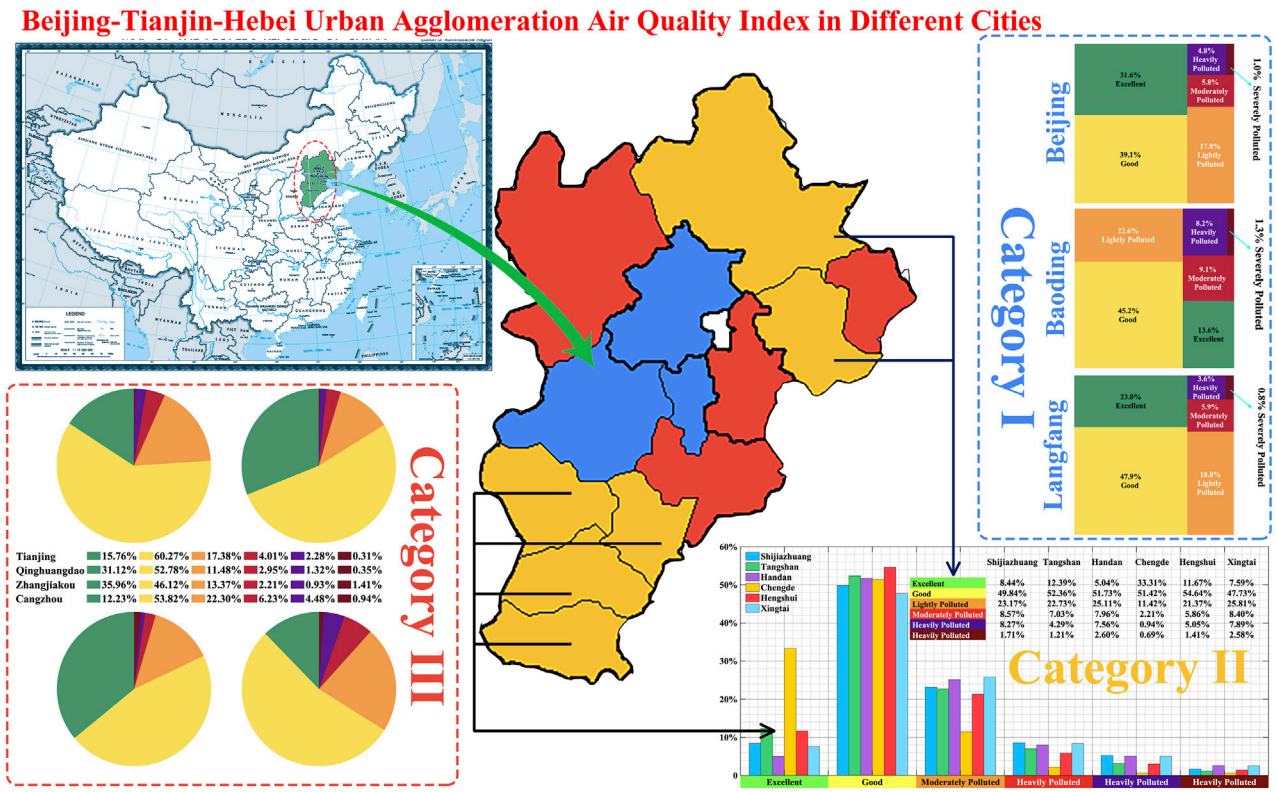


FIGURE 2 | Result of data mining for Beijing–Tianjin–Hebei Urban Agglomeration.

TABLE 1 | The result of fuzzy comprehensive evaluation

City	Air pollution					
	PM_{10}	$PM_{2.5}$	NO_2	CO	SO_2	O_3
Beijing	**0.91237**	**0.98072**	**0.63530**	0.43071	0.39424	0.38603
Tianjin	**0.86953**	**0.90659**	**0.64191**	0.30671	0.37710	0.48547
Shijiazhuang	**0.75937**	**0.70576**	**0.67125**	0.38993	0.43133	0.41205
Tangshan	**0.91044**	**0.74470**	**0.64615**	0.37757	0.35546	0.43152
Qinhuangdao	**0.87371**	**0.89213**	**0.72628**	0.41083	0.30361	0.41777
Handan	**0.82031**	**0.98267**	**0.72983**	0.48603	0.31468	0.49503
Baoding	**0.87530**	**0.95838**	**0.75981**	0.34638	0.41152	0.39322
Zhangjiakou	**0.90838**	**0.99560**	**0.63953**	0.30755	0.42579	0.31045
Chengde	**0.81195**	**0.73604**	**0.82316**	0.40926	0.48029	0.31550
Langfang	**0.75893**	**0.81218**	**0.87910**	0.47526	0.42354	0.30550
Cangzhou	**0.84830**	**0.82908**	**0.86664**	0.49499	0.31973	0.41617
Hengshui	**0.75544**	**0.93064**	**0.97252**	0.35830	0.40872	0.33812
Xingtai	**0.90721**	**0.93224**	**0.70024**	0.47492	0.44532	0.31803

Note: if the result of fuzzy comprehensive evaluation is greater than 0.5, the pollutant is the main pollutant. The bold values are main Air Pollution, which the fuzzy comprehensive evaluation results are greater than 0.5.

For the second forecasting, the 2nd to 841st samples are the training samples, the 842nd to 1009th samples are the testing samples, and the 1010th sample is the forecasting value. At the end of the forecasting, the WIC value of the testing sample is calculated. If the WIC of the *i*th model is the smallest, the forecasting value of the *i*th model provides the optimal forecasting value.

For the *k*th forecasting, the *k*th to $(840 + k - 1)$th samples are the training samples, the $(840 + k)$th to $(1008 + k - 1)$th samples are the testing samples, and the $(1008 + k)$th sample is the forecasting value. At the end of the forecasting, the WIC value of the testing sample is calculated. If the WIC of the *i*th model is the smallest, the forecasting value of the *i*th model provides the optimal forecasting value.

TABLE 2 | Statistic of each main air pollutant in different cities.

City	Pollutant	Min	Max	Mean	Skewness	Kurtosis	Mode
Beijing	NO_2	16	158	39.1947	0.7894	3.2759	18
	$PM_{2.5}$	3	273	50.0068	1.7017	6.5652	5
	PM_{10}	56	1,058	83.0744	5.7125	63.4085	27
Tianjin	NO_2	17	195	49.4348	0.8042	3.5754	36
	$PM_{2.5}$	11	237	49.9759	2.0621	9.2214	40
	PM_{10}	23	321	85.0560	1.7136	7.1605	68
Shijiazhuang	NO_2	4	145	44.0150	0.6088	2.7786	16
	$PM_{2.5}$	12	343	69.6938	1.7761	6.5778	43
	PM_{10}	29	461	132.2141	1.3934	4.9017	80
Tangshan	NO_2	5	200	54.2055	0.6325	3.5206	36
	$PM_{2.5}$	12	350	60.4803	2.4753	12.7626	44
	PM_{10}	26	484	114.9052	1.8083	8.0477	76
Qinhuangdao	NO_2	4	130	43.2894	0.6124	2.7408	17
	$PM_{2.5}$	6	250	38.2879	2.1722	9.8110	21
	PM_{10}	15	392	78.6504	1.7126	7.8222	54
Handan	NO_2	4	153	40.4909	0.9210	3.4780	17
	$PM_{2.5}$	8	321	69.4114	1.9836	7.2348	45
	PM_{10}	24	513	135.8368	1.6375	5.8557	85
Baoding	NO_2	4	159	44.7402	0.8947	3.3395	15
	$PM_{2.5}$	11	402	66.6175	1.9971	8.9785	33
	PM_{10}	24	527	115.9434	1.5658	6.3524	65
Zhangjiakou	NO_2	1	136	22.2525	1.9035	8.0002	19
	$PM_{2.5}$	7	184	31.1680	2.2524	10.1861	16
	PM_{10}	12	1,581	88.7293	6.2068	64.1126	31
Chengde	NO_2	2	126	30.6221	0.7458	3.5114	6
	$PM_{2.5}$	7	189	31.7188	2.2798	10.0002	13
	PM_{10}	14	561	81.6457	2.6593	14.4003	36
Langfang	NO_2	7	206	45.0315	1.0029	3.9963	20
	$PM_{2.5}$	9	282	51.8249	1.9651	7.9869	15
	PM_{10}	21	428	100.0498	1.6177	6.3231	69
Cangzhou	NO_2	5	170	41.2205	1.0111	4.0626	22
	$PM_{2.5}$	10	381	58.8355	2.5108	12.7637	35
	PM_{10}	16	509	104.4355	1.8495	8.5249	75
Hengshui	NO_2	2	215	32.2611	1.1356	4.8037	15
	$PM_{2.5}$	14	310	61.7583	2.3912	10.0981	40
	PM_{10}	19	391	103.5030	1.7636	6.3377	68
Xingtai	NO_2	3	155	47.4996	0.5950	2.9794	18
	$PM_{2.5}$	11	369	69.1419	1.9618	7.4487	35
	PM_{10}	17	524	135.3799	1.3911	5.2029	88

TABLE 3 | Definition of the performance metrics.

Metric	Definition	Equation
MAE	The average absolute forecasting results error of *n*	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
RMSE	The mean absolute percentage error of *n* forecasting results	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
MAPE	The root mean square error of *n* forecasting results	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\lvert\frac{y_i - \hat{y}_i}{y_i}\rvert \times 100\%$
STDE	The standard of error of *n* forecasting results	$STDE = \sqrt{Var(y_n - \hat{y}_n)}$
U1	Theil U statistic 1	$U1 = \frac{\sqrt{\frac{1}{T}\sum_{T=1}^{T}(y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{T}\sum_{T=1}^{T}(y_i)^2} + \sqrt{\frac{1}{T}\sum_{T=1}^{T}(\hat{y}_i)^2}}$
U2	Theil U statistic 2	$U2 = \frac{\sqrt{\frac{1}{T}\sum_{T=1}^{T}(\frac{\hat{y}_{i+1} - y_{i+1}}{y_i})^2}}{\sqrt{\frac{1}{T}\sum_{T=1}^{T}(\frac{y_{i+1} - y_i}{y_i})^2}}$
R^2	The goodness of forecasting fit	$\frac{R^2 = \sum_{n=1}^{N}(y_n - \bar{y}) - \sum_{n=1}^{N}(y_n - \hat{y}_n)}{\sum_{n=1}^{N}(y_n - \bar{y})}$
DA	Directions or turning points between actual and forecasting values	$DA = \frac{1}{T}\sum_{i=1}^{T} a_i,\ a_i = \begin{cases} 1 & if\ (y_{i+1} - \hat{y}_i)(\hat{y}_{i+1} - y_i) > 0 \\ 0 & otherwise. \end{cases}$

Note: y_i is the actual value, $\hat{y}_i$ is the forecasted value, and T *is the total number of data items.*

TABLE 4 | The result of model selection for main air pollutants in different seasons.

Category	Air pollutants	Selection	First season				Second season				
I	NO_2	Model	SVM	GRNN	BPNN	ANFIS	SVM	GRNN	BPNN	LSTM	
		Percentage	67.86%	21.43%	7.14%	3.57%	39.29%	46.43%	10.71%	3.57%	
II		Model	SVM	GRNN	BPNN	LSTM	SVM	GRNN	BPNN	ANFIS	LSTM
		Percentage	53.57%	35.71%	3.57%	7.14%	64.29%	17.86%	7.14%	3.57%	7.14%
III		Model	SVM	GRNN	BPNN		SVM	GRNN	BPNN	LSTM	
		Percentage	53.57%	35.71%	10.71%		32.14%	50.00%	10.71%	7.14%	
I	$PM_{2.5}$	Model	SVM	BPNN			SVM	BPNN			
		Percentage	42.86%	57.14%			57.14%	42.86%			
II		Model	SVM	BPNN	ANFIS	LSTM	SVM	BPNN	ANFIS	LSTM	
		Percentage	64.29%	21.43%	3.57%	10.71%	64.29%	17.86%	4.29%	3.57%	
III		Model	SVM	BPNN	ANFIS	LSTM	SVM	BPNN	ANFIS	LSTM	
		Percentage	71.43%	17.86%	3.57%	7.14%	78.57%	14.29%	3.57%	3.57%	
I	PM_{10}	Model	SVM	BPNN	ANFIS		SVM	BPNN			
		Percentage	82.14%	14.29%	3.57%		78.57%	17.86%			
II		Model	SVM	BPNN			SVM	BPNN			
		Percentage	76.79%	16.07%			78.57%	21.43%			
III		Model	SVM	BPNN			SVM	GRNN	BPNN		
		Percentage	82.14%	17.86%			75.00%	14.29%	10.71%		

Category	Air pollutants	Selection	Third season					Fourth season			
I	NO_2	Model	SVM	GRNN	BPNN			SVM	GRNN	BPNN	LSTM
		Percentage	57.14%	25.00%	17.86%			60.71%	28.57%	5.36%	5.36%
II		Model	SVM	GRNN	BPNN			SVM	GRNN	BPNN	LSTM
		Percentage	60.71%	32.14%	7.14%			51.79%	32.14%	8.93%	7.14%
III		Model	SVM	GRNN	BPNN	ANFIS	LSTM	SVM	GRNN	BPNN	
		Percentage	46.43%	25.00%	14.29%	7.14%	7.14%	46.43%	39.29%	10.71%	
I	$PM_{2.5}$	Model	SVM	BPNN	LSTM			SVM	BPNN	ANFIS	
		Percentage	60.71%	17.86%	21.43%			50.00%	33.93%	14.29%	
II		Model	SVM	BPNN	ANFIS	LSTM		SVM	BPNN	ANFIS	
		Percentage	71.43%	17.86%	7.14%	3.57%		71.43%	21.43%	7.14%	
III		Model	SVM	BPNN	ANFIS	LSTM	GRNN	SVM	BPNN	ANFIS	LSTM
		Percentage	53.57%	28.57%	7.14%	7.14%	3.57%	64.29%	21.43%	10.71%	3.57%
I	PM_{10}	Model	SVM	BPNN				SVM	BPNN	ANFIS	
		Percentage	82.14%	17.86%				71.43%	21.43%	7.14%	
II		Model	SVM	BPNN	ANFIS			SVM	BPNN		
		Percentage	71.43%	17.86%	10.71%			67.86%	21.43%		
III		Model	SVM	BPNN	ANFIS			SVM	BPNN	ANFIS	
		Percentage	78.57%	17.86%	3.57%			76.79%	17.86%	5.36%	

In the whole forecasting process, 168 optimal forecasting values are generated, and 168*5 WIC values are generated at the same time. **Table 4** only shows the optimal model and the percentage of optimal forecasting value for three main air pollutants.

From **Table 4**, it can be seen that SVM provides more optimal forecasting value for the three main pollutants at different times, especially in the PM_{10} forecasting process; the optimal forecasting value for the first quarter and the third quarter is 82.14% (138 optimal forecasting value), and the other four models also provide corresponding optimal forecasting value.

Experiment I: Forecasting Processing for Three Categories of NO_2 by Each Model in the First Season

In this portion, the hourly NO_2 time series for 13 cities in three categories were utilized as the testing data for the five hybrid models with one-step-ahead forecasting. Beyond that, with the purpose of comprehensively comparing the precision of the modeling forecasting, this experiment consisted of two parts: the multi-step forecasts demonstrated in **Table 4** and, for the local analysis horizon, the local forecasts presented in **Table ′5** and **Figure 3**, which focus on first season. **Table 5** and **Figure 3** demonstrate the following:

1) Focusing on Category I, the new proposed model based on model selection realizes excellent results on the eight evaluation indices in the first season forecasting. On the contrary, DEGWO-ANFIS has the lowest effectiveness. The maximum reduction of MAPE for the proposed model compared with the other hybrid models is approximately 71.18% in Beijing's NO_2 forecasting, 53.93% in Baoding's NO_2 forecasting, and 61.61% in Langfang's NO_2 forecasting, respectively. The reduction was about 62.39% and 76.49% for one-step forecasting and 2.79%, 6.10%, and 19.33% for the three cities at the hourly interval

TABLE 5 | The forecasting results of each model for NO_2 in three categories.

Metric	Model	MAE	RMSE	STDE	DA	U1	U2	MAPE	R^2
Category I									
Beijing	MODEGWO-SVM	2.1754	3.1919	3.2010	80.84%	0.0464	2.0148	9.95%	0.9868
	MODEGWO-GRNN	3.0082	3.9179	3.9242	50.90%	0.0569	2.6113	15.23%	0.9800
	MODEGWO-BPNN	2.6589	3.7849	3.7857	69.46%	0.0550	2.4488	13.07%	0.9814
	MODEGWO-ANFIS	3.4548	6.0636	6.0514	64.07%	0.0889	4.1096	16.57%	0.9524
	Model select	**2.1300**	**3.1765**	**3.1847**	**77.25%**	**0.0463**	**1.9446**	**9.68%**	**0.9869**
Baoding	MODEGWO-SVM	4.7162	7.4952	7.5055	77.25%	0.0665	2.1441	12.69%	0.9620
	MODEGWO-GRNN	5.1070	7.9556	7.8716	60.48%	0.0706	3.6465	17.57%	0.9569
	MODEGWO-BPNN	6.9580	9.8870	9.5789	59.28%	0.0906	2.3774	17.26%	0.9355
	MODEGWO-ANFIS	7.0574	10.5964	10.1816	56.89%	0.2401	2.4591	18.41%	0.9223
	Model select	**4.4448**	**7.3418**	**7.3599**	**76.65%**	**0.0654**	**2.1115**	**11.96%**	**0.9628**
Langfang	MODEGWO-SVM	4.5141	7.5578	7.5731	76.05%	0.0880	2.2411	13.46%	0.9372
	MODEGWO-GRNN	3.8320	5.2949	5.3098	62.28%	0.0625	1.7523	12.52%	0.9682
	MODEGWO-BPNN	5.3965	8.4998	8.3895	62.87%	0.0980	2.9703	17.22%	0.9214
	MODEGWO-ANFIS	6.3336	11.4957	11.3948	60.48%	0.1305	3.2108	18.23%	0.8745
	Model select	**3.6594**	**5.3249**	**5.3380**	**68.86%**	**0.0627**	**1.7819**	**11.28%**	**0.9676**
Category II									
Shijiazhuang	MODEGWO-SVM	2.8162	4.8606	4.8733	82.04%	0.0498	2.3785	8.21%	0.9695
	MODEGWO-GRNN	3.5480	5.0067	5.0163	50.90%	0.0517	2.2642	10.46%	0.9674
	MODEGWO-BPNN	4.3533	6.7126	6.6382	58.68%	0.0694	2.9803	11.62%	0.9450
	Adam-LSTM	5.8291	8.3652	8.3630	55.09%	0.0859	3.5319	14.95%	0.9167
	Model select	**2.8001**	**4.0036**	**4.0095**	**74.25%**	**0.0412**	**2.0289**	**8.02%**	**0.9792**
Tangshan	MODEGWO-SVM	4.2928	6.4751	6.4876	82.63%	0.0584	2.3028	12.41%	0.9643
	MODEGWO-GRNN	3.5396	5.2423	5.2563	73.65%	0.0474	1.5259	9.14%	0.9763
	MODEGWO-BPNN	5.9179	8.2307	8.2550	68.86%	0.0743	2.3721	16.20%	0.9423
	Adam-LSTM	5.0169	6.9705	6.9845	74.25%	0.0630	2.3430	13.86%	0.9579
	Model select	**3.5604**	**5.3107**	**5.3266**	**77.84%**	**0.0481**	**1.5419**	**9.10%**	**0.9757**
Handan	MODEGWO-SVM	3.6331	5.9196	5.9315	72.46%	0.0570	2.2986	9.69%	0.9552
	MODEGWO-GRNN	4.0699	6.2263	6.2395	60.48%	0.0604	2.3158	11.35%	0.9494
	MODEGWO-BPNN	5.7352	8.4969	8.4366	55.69%	0.0832	2.3412	13.92%	0.9058
	Adam-LSTM	5.3202	7.0150	6.9787	60.48%	0.0682	2.2208	12.31%	0.9378
	Model select	**3.3618**	**5.4628**	**5.4791**	**76.05%**	**0.0527**	**2.1130**	**8.79%**	**0.9620**
Chengde	MODEGWO-SVM	2.6775	4.0365	4.0326	71.26%	0.0653	1.8783	15.35%	0.9735
	MODEGWO-GRNN	2.6787	4.0350	4.0461	62.87%	0.0658	2.0476	17.48%	0.9735
	MODEGWO-BPNN	3.4332	4.9117	4.9230	60.48%	0.0804	2.2906	20.15%	0.9605
	Adam-LSTM	4.5095	6.9772	6.9648	67.07%	0.1123	2.6262	26.28%	0.9201
	Model select	**2.4481**	**3.7646**	**3.7746**	**73.05%**	**0.0612**	**1.7081**	**13.90%**	**0.9769**
Hengshui	MODEGWO-SVM	4.7952	6.5453	6.5494	73.65%	0.0764	2.1369	15.41%	0.9405
	MODEGWO-GRNN	6.0417	8.0467	8.0648	47.90%	0.0950	3.2945	23.59%	0.9078
	MODEGWO-BPNN	10.7085	17.2255	16.7668	55.69%	0.2122	3.6978	27.02%	0.5924
	Adam-LSTM	7.5643	10.4959	9.9896	49.70%	0.1296	2.3976	22.00%	0.8544
	Model select	**4.3346**	**6.1328**	**6.1423**	**74.85%**	**0.0717**	**2.0600**	**14.27%**	**0.9477**
Xingtai	MODEGWO-SVM	3.7973	5.6732	5.6880	70.66%	0.0494	2.1280	9.22%	0.9688
	MODEGWO-GRNN	4.0825	6.1326	6.1358	65.87%	0.0535	2.3674	10.61%	0.9633
	MODEGWO-BPNN	5.5707	8.1697	8.0221	61.68%	0.0727	2.6425	12.39%	0.9361
	Adam-LSTM	7.1767	11.0393	11.0367	65.87%	0.0967	2.9015	15.29%	0.8884
	Model select	**3.5614**	**5.4338**	**5.4465**	**67.07%**	**0.0474**	**1.9570**	**8.58%**	**0.9710**
Category III									
Tianjin	MODEGWO-SVM	2.5155	3.5288	3.5317	80.84%	0.0355	1.8109	5.98%	0.9830
	MODEGWO-GRNN	2.9357	4.2368	4.2449	70.06%	0.0427	1.9437	7.51%	0.9750
	MODEGWO-BPNN	3.1418	4.3375	4.3501	77.84%	0.0436	2.0924	7.57%	0.9744
	Model select	**2.3407**	**3.3033**	**3.3130**	**87.24%**	**0.0333**	**1.7699**	**5.66%**	**0.9850**
Qinhuangdao	MODEGWO-SVM	3.8108	5.4949	5.4950	81.44%	0.0554	1.7553	10.87%	0.9744
	MODEGWO-GRNN	3.7633	5.3865	5.3965	65.87%	0.0549	2.1029	12.88%	0.9752
	MODEGWO-BPNN	5.4520	7.6113	7.3974	67.66%	0.0791	2.0834	14.56%	0.9531
	Model select	**3.4285**	**4.8590**	**4.8733**	**74.85%**	**0.0494**	**1.7261**	**10.28%**	**0.9797**
Zhangjiakou	MODEGWO-SVM	2.3096	3.5862	3.5967	75.45%	0.0945	1.8076	13.87%	0.9058
	MODEGWO-GRNN	3.1474	4.1463	4.1587	44.91%	0.1106	2.3875	20.99%	0.8669
	MODEGWO-BPNN	2.7829	4.2140	4.1925	61.08%	0.1137	2.0222	16.66%	0.8640
	Model select	2.0741	2.8440	2.8524	71.86%	0.0753	1.7182	13.10%	0.9395
Cangzhou	MODEGWO-SVM	6.1489	9.0417	9.0626	78.44%	0.0803	2.0706	15.17%	0.9373
	MODEGWO-GRNN	5.4532	7.6253	7.6448	58.68%	0.0682	2.0220	14.67%	0.9535
	MODEGWO-BPNN	8.0522	11.4542	11.4009	55.69%	0.1031	2.8158	19.38%	0.9018
	Model select	**5.3105**	**7.9582**	**7.9179**	**70.66%**	**0.0716**	**1.9162**	**12.87%**	**0.9509**

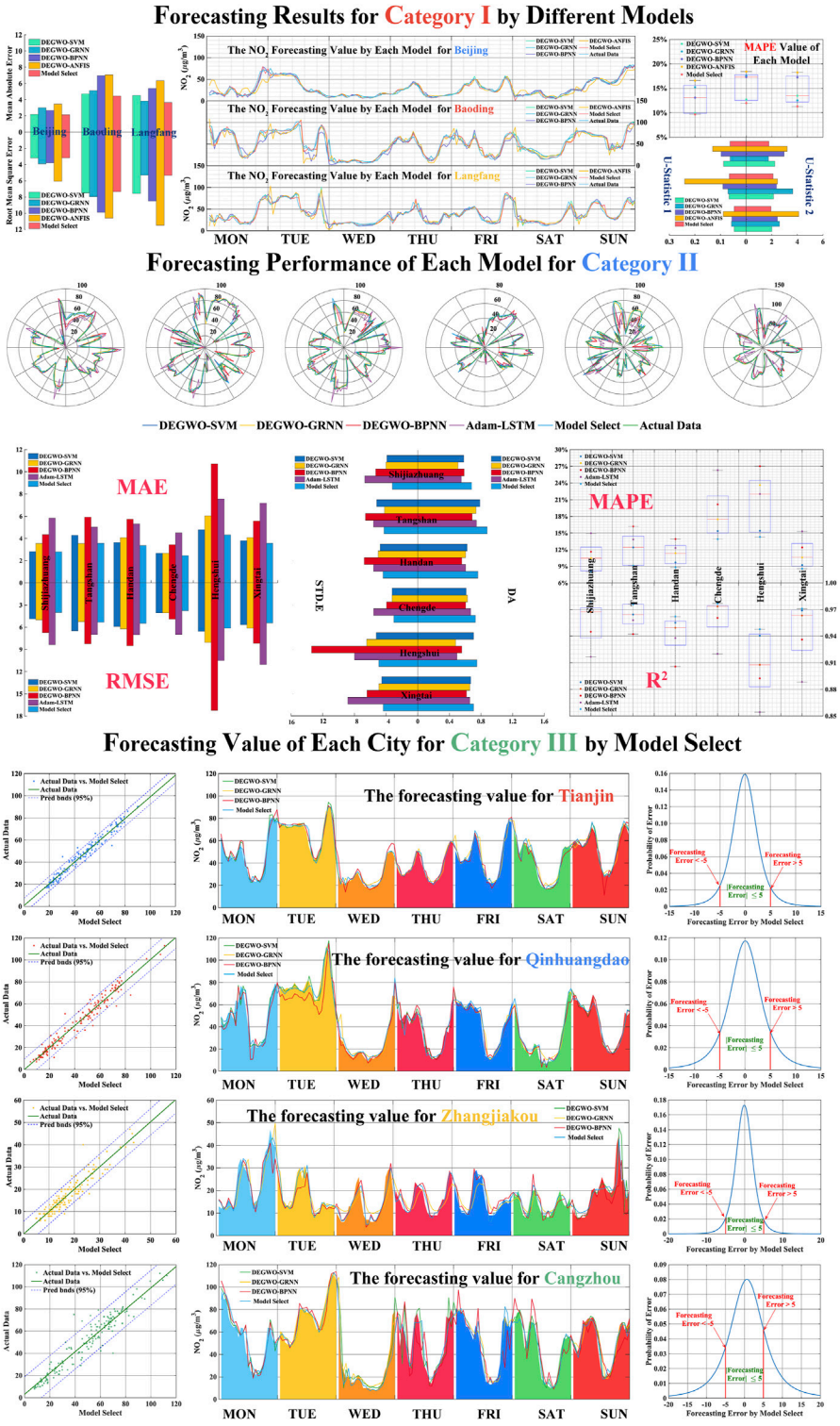


FIGURE 3 | Forecasting result of NO_2 for three categories in the first season.

NO_2 forecasting in Category I. As for the R^2, the proposed model has the best performance among the four single-hybrid models for hourly interval NO_2 time series.

2) Focusing on Category II, it is clear that proposed model based on model selection exhibits the best performance among the single fourth hybrid models implemented for all eight criteria

TABLE 6 | The $PM_{2.5}$ forecasting result for each city in Category II by different models.

Metric	Shijiazhuang				
	MODEGWO-SVM	**MODEGWO-BPNN**	**MODEGWO-ANFIS**	**Adam-LSTM**	**Model select**
MAE	0.4632	0.4450	0.6554	1.2945	0.3518
RMSE	0.6106	0.5877	1.2148	2.0978	0.4497
STDE	0.6124	0.5809	1.2182	2.1038	0.4509
DA	75.45%	73.05%	70.66%	42.51%	81.44%
U1	0.0067	0.0064	0.0133	0.0221	0.0049
U2	1.4464	1.3998	3.4525	4.3147	1.2087
MAPE	1.31%	1.08%	1.85%	2.88%	0.88%
R^2	0.9991	0.9992	0.9964	0.9889	0.9995
Metric	**Handan**				
	MODEGWO-SVM	**MODEGWO-BPNN**	**MODEGWO-ANFIS**	**Adam-LSTM**	**Model select**
MAE	0.5522	0.5250	1.3025	2.1956	0.4694
RMSE	0.7024	0.7009	2.5927	3.1495	0.6125
STDE	0.6835	0.7007	2.5352	3.1450	0.6092
DA	82.63%	83.83%	71.26%	43.71%	88.02%
U1	0.0051	0.0051	0.0191	0.0232	0.0045
U2	1.3944	1.4399	2.7102	4.7186	1.2395
MAPE	0.90%	0.85%	2.08%	3.46%	0.76%
R^2	0.9993	0.9992	0.9896	0.9831	0.9994
Metric	**Hengshui**				
	MODEGWO-SVM	**MODEGWO-BPNN**	**MODEGWO-ANFIS**	**Adam-LSTM**	**Model select**
MAE	0.4649	0.5305	1.5424	2.6723	0.4392
RMSE	0.6107	0.7545	3.3746	3.8908	0.6260
STDE	0.6044	0.7515	3.3556	3.8552	0.6279
DA	82.63%	74.85%	65.87%	38.92%	83.44%
U1	0.0056	0.0070	0.0312	0.0356	0.0058
U2	1.3886	1.4988	6.8777	6.1294	1.2831
MAPE	0.92%	1.07%	3.34%	5.02%	0.88%
R^2	0.9985	0.9984	0.9744	0.9724	0.9989
Metric	**Tangshan**				
	MODEGWO-SVM	**MODEGWO-BPNN**	**MODEGWO-ANFIS**	**Adam-LSTM**	**Model select**
MAE	0.3165	0.4449	1.1059	1.5655	0.3107
RMSE	0.3984	0.5718	2.9201	2.4642	0.3948
STDE	0.3961	0.5714	2.8491	2.4622	0.3947
DA	80.24%	69.46%	62.28%	43.11%	82.04%
U1	0.0052	0.0074	0.0381	0.0321	0.0051
U2	1.0656	1.3226	11.8381	5.4107	1.0363
MAPE	0.94%	1.36%	3.85%	4.63%	0.91%
R^2	0.9996	0.9992	0.9818	0.9845	0.9996
Metric	**Chengde**				
	MODEGWO-SVM	**MODEGWO-BPNN**	**MODEGWO-ANFIS**	**Ada-LSTM**	**Model select**
MAE	0.2413	0.3283	0.6149	0.9537	0.2388
RMSE	0.3103	0.4255	1.1569	1.7033	0.3017
STDE	0.3101	0.4241	1.1353	1.7052	0.3007
DA	70.06%	61.08%	53.89%	37.13%	72.06%
U1	0.0096	0.0104	0.0281	0.0414	0.0076
U2	1.0296	0.9017	2.7485	5.1292	1.0022
MAPE	1.46%	1.79%	3.20%	4.77%	1.34%
R^2	0.9987	0.9976	0.9827	0.9663	0.9987
Metric	**Xingtai**				
	MODEGWO-SVM	**MODEGWO-BPNN**	**MODEGWO-ANFIS**	**Ada-LSTM**	**Model select**
MAE	0.4611	0.5550	0.9154	1.4376	0.4354
RMSE	0.5936	0.7151	1.4925	1.9675	0.5685
STDE	0.5905	0.7119	1.4969	1.9238	0.5691
DA	74.85%	70.06%	56.89%	33.53%	77.25%
U1	0.0049	0.0059	0.0123	0.0161	0.0047
U2	1.3588	1.4830	3.4545	4.4867	1.2932
MAPE	0.79%	0.98%	1.67%	2.47%	0.75%
R^2	0.9988	0.9983	0.9923	0.9874	0.9989

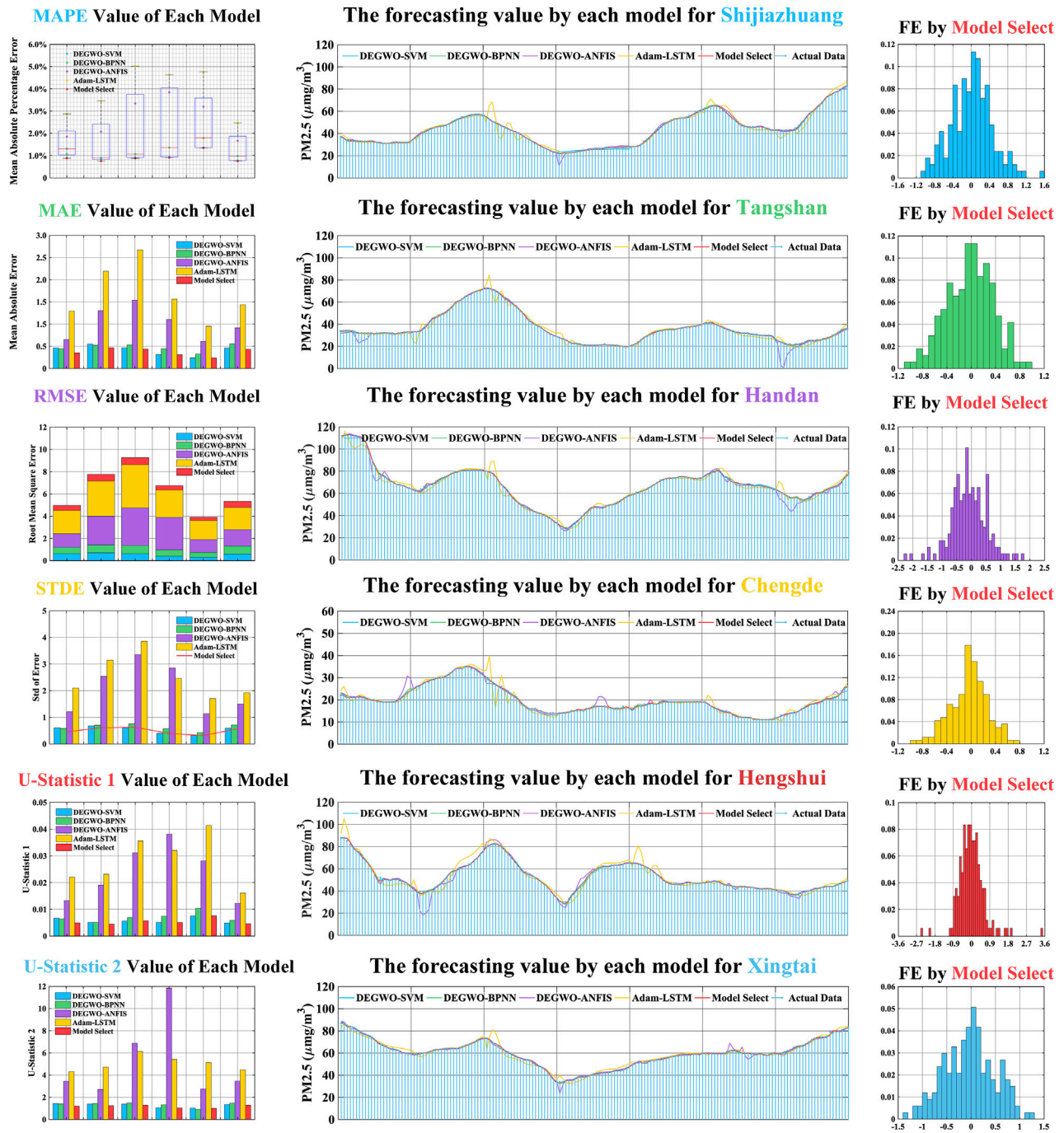


FIGURE 4 | Forecasting performance of each model for Category II in the first season.

involved. For the MAPE, there are average reductions between the proposed model and single hybrid model, by approximately 12.48%, 29.12%, 60.00%, and 66.70% in six cities for the hourly NO_2 time series forecasting, respectively. Comparing the four single hybrid models for the hourly NO_2 forecasting, the forecasting accuracy of DEGWO-SVM is higher than that of the other three hybrid models. The average reduction of MAPE among the MODEGWO-SVM and the other three hybrid models is 16.70%, 42.79%, and 50.10%, respectively. In addition, all the R^2 values of the proposed model are over 90%, which underlines the higher fitting effect on Category II.

3) The forecasting metric of the single hybrid models and the proposed model in **Table 4** indicates that the proposed model based on model selection performs better than the single hybrid model in Category III. As an example, with respect to Tianjin, the DA values of the individual hybrid models are 80.84% (MODEGWO-SVM), 70.06% (MODEGWO-GRNN), and 77.84% (MODEGWO-BPNN), while the DA values of the proposed models is 87.24%, respectively. The comparative analysis between the proposed model and the single model confirms the advantages of the hybrid forecasting model.
4) Moreover, **Table 5** displays each metric of NO_2 forecasting among the developed hybrid forecasting system and the single hybrid models. According to **Table 5**, it is obvious that the values of MAE, RMSE, MAPE, U1, and U2 of the proposed hybrid model are all smaller than the other considered models, and the values of DA and R^2 of the developed hybrid forecasting system are all greater than that of the single hybrid model, which further confirms the superiority of the presented hybrid forecasting system in terms of forecasting ability.

In summary, from the analyses above, it can be concluded that the model selection forecasting system realizes the best forecasting results compared to the single hybrid model. Model selection also gives better forecasting performance in the other season with the results shown in **Supplementary Appendices S2–S4** indicating better robustness of the model selection forecasting system.

Experiment II: The Forecasting Results of Category II $PM_{2.5}$ Forecasting in the First Season

This experiment mainly focused on the forecasting performance of each model for $PM_{2.5}$ of Category II in the first season, with the forecasting results of four different hybrid models (MODEGWO-SVM, MODEGWO-BPNN, MODEGWO-ANFIS, Adam-LSTM) and model selection represented in **Table 6** and **Figure 4**

1) For first season $PM_{2.5}$ forecasting accuracy, the final forecast results of $PM_{2.5}$ for six cities in Category II are composed of four hybrid models, which include MODEGWO-SVM, MODEGWO-BPNN, MODEGWO-ANFIS, and Adam-LSTM. Among the four models, MODEGWO-SVM and MODEGWO-BPNN have better forecasting performance, with the MODEGWO-SVM obtaining 64.29% optimal forecasting points and the MODEGWO-BPNN obtaining 21.43% optimal points for six cities in Category II. The smallest MAPE values of MODEGWO-SVM are 0.92%, 0.94%, 1.36%, and 0.79% for Hengshui, Tangshan, Chengde, and Xingtai $PM_{2.5}$ forecasting, and the MODEGWO-BPNN obtains the best MAPE (1.08% and 0.85%) value for Shijiazhuang and Handan.
2) For the goodness of fit, the R^2 values of four different hybrid models are over 0.95 for six cities in the first season, which indicates that the forecasting values obtained by these models is close to the actual value. The forecasting result of Shijiazhuang shows that the R^2 value of best hybrid modes (MODEGWO-BPNN) is 0.9993, very close to 1, which indicates that there is less difference between forecasting data and actual data, and the forecasting value is basically consistent with the actual value. Meanwhile, The DA values of the best hybrid model are over 70%, which proves the best models can effectively capture the changing trend of the actual data.
3) For the forecasting results of model selection, **Table 6** and **Figure 4** clearly show that the forecasting performance of model selection is better than the hybrid model. The MAPE value of model selection is 0.88%; compared with the optimal hybrid model the MAPE improved 22.73%, and the MAE and RMSE reduced 26.49% and 30.69%, respectively. Although the model selection can improve the forecasting accuracy, in the $PM_{2.5}$ forecasting of some cities, due to the better forecasting performance of the single hybrid model, the forecasting accuracy of the model selection is not significantly improved. For example, in the $PM_{2.5}$ forecasting of Xingtai, the MAPE value of optimal forecasting model (MODEGWO-SVM) is 0.79%, and the MAPE of the model selection is 0.76%, which shows that forecasting accuracy has no significant improvement.
4) Similar to the first season, the $PM_{2.5}$ forecasting results of Category II in the second to fourth seasons are listed in **Supplementary Appendix Table 7** in which the best forecasting performances of the hybrid model are shown by DEGWO-SVM, DEGWO-BPNN, DEGWO-ANFIS, and Adam-LSTM for $PM_{2.5}$ forecasting in each city. Compared with the optimal hybrid model, the final forecasting results obtained by the model selection is more accurate than single hybrid model, which indicates that the optimal hybrid model has good forecasting performance but cannot be applied to all the forecasting data.

In summary, for the Category I and Category III $PM_{2.5}$ forecasting list in **Supplementary Appendices S5–S7**, the model selection forecasting system exhibits the best forecasting accuracy among the different hybrid models for four seasons. It is evident that the forecasting capacity of the model selection is robust when considering each forecasting metrics. The accuracy of model selection depends on the hybrid model, so it is necessary to increase the types of models in the modeling process which ensures that more forecasting results can be obtained, and the optimal forecasting value can be selected in the model selection process.

Experiment III: PM_{10} Forecasting Analysis for Category III in Four Seasons

For the hourly PM_{10} time series for three categories, it can be observed that the model selection forecasting system attains satisfactory results. Specifically, the lowest values of MAE are 0.4643, 0.4600, and 0.3869 and of RMSE are 0.7302, 0.7906, and 0.5561, corresponding to PM_{10} forecasting in Category I in three cities, successively. The results of Category I indicate that the smaller the MAE and MSE, the smaller the deviation between the

TABLE 7 | The forecasting result of each model in different seasons for Category III.

Location	Metric	First season			Second season					Third season				Fourth season			
		DEGWO-SVM	DEGWO-BPNN	Model selection	DEGWO-SVM	DEGWO-GRNN	DEGWO-BPNN	DEGWO-ANFIS	Model selection	DEGWO-SVM	DEGWO-BPNN	DEGWO-ANFIS	Model selection	DEGWO-SVM	DEGWO-BPNN	DEGWO-ANFIS	Model selection
Tianjin	MAE	0.4119	0.4687	0.3576	2.3288	3.1388	2.6986	6.2681	1.8811	0.7567	0.8285	1.1091	0.5870	0.6775	0.8685	1.0346	0.5747
	RMSE	0.5572	0.6236	0.4552	8.1840	7.7183	8.6510	25.5466	7.1780	0.9382	1.1333	1.5570	0.7880	0.8649	1.1016	1.4746	0.7649
	STDE	0.5472	0.6019	0.4535	8.2080	7.7255	8.6768	25.5589	7.1987	0.9259	1.1090	1.5400	0.7841	0.8531	1.0938	1.4463	0.7531
	DA	81.44%	74.25%	76.05%	76.05%	31.74%	64.67%	65.87%	76.05%	55.09%	55.09%	47.90%	68.26%	71.26%	50.30%	50.90%	71.26%
	U1	0.0037	0.0051	0.0037	0.0435	0.0412	0.0460	0.1343	0.0382	0.0064	0.0078	0.0107	0.0054	0.0063	0.0076	0.0101	0.0053
	U2	1.3126	1.4969	1.2192	1.9855	1.9275	2.0215	13.1964	1.6470	2.8029	1.8298	2.7405	1.7659	2.2110	2.2721	2.9704	2.1810
	MAPE	0.71%	0.83%	0.65%	2.65%	3.69%	3.09%	7.55%	2.12%	1.09%	1.13%	1.53%	0.84%	0.93%	1.19%	1.41%	0.83%
	R^2	0.9998	0.9996	0.9998	0.9601	0.9629	0.9534	0.7662	0.9681	0.9974	0.9965	0.9930	0.9982	0.9973	0.9969	0.9938	0.9983
Qinhuangdao	MAE	0.4140	0.5075	0.4034	0.7588	2.5923	1.0097	1.9170	0.7441	0.4645	0.6423	1.1009	0.4303	0.7377	0.8356	1.1088	0.6459
	RMSE	0.5527	0.6725	0.5440	1.2183	3.4536	1.6718	3.9976	1.2117	0.6622	0.8380	2.4277	0.5914	1.0205	1.2012	1.5364	0.9614
	STDE	0.5529	0.6736	0.5452	1.2219	3.4639	1.6274	3.9538	1.2146	0.6639	0.8317	2.3892	0.5932	1.0232	1.2029	1.5351	0.9642
	DA	82.04%	68.86%	83.23%	85.63%	32.93%	79.64%	72.46%	86.23%	81.44%	70.66%	72.46%	82.63%	79.04%	80.84%	64.67%	85.63%
	U1	0.0044	0.0054	0.0044	0.0061	0.0173	0.0083	0.0199	0.0061	0.0053	0.0067	0.0194	0.0047	0.0054	0.0063	0.0081	0.0050
	U2	1.1881	1.2933	1.1573	1.3515	4.7857	1.5998	4.0494	1.2920	1.6354	1.6295	5.0927	1.4787	1.4708	1.4926	1.9014	1.3827
	MAPE	0.81%	0.96%	0.75%	0.86%	3.22%	1.09%	2.09%	0.84%	0.96%	1.19%	2.18%	0.82%	1.00%	1.07%	1.49%	0.83%
	R^2	0.9997	0.9996	0.9997	0.9995	0.9964	0.9993	0.9959	0.9996	0.9994	0.9991	0.9925	0.9995	0.9996	0.9994	0.9991	0.9997
Zhangjiakou	MAE	0.6366	0.7970	0.4310	4.1069	7.4836	4.7017	4.8835	3.7104	0.4875	0.5652	0.8045	0.4782	0.7864	1.0681	1.7907	0.7297
	RMSE	0.7642	1.1575	0.5662	11.3281	12.9323	11.1572	12.2400	9.2301	0.6852	0.7945	1.2237	0.6836	1.0777	1.4880	2.8905	1.0128
	STDE	0.7655	1.1117	0.5666	11.3585	12.8990	11.1833	12.2350	9.2556	0.6797	0.7559	1.1982	0.6802	1.0638	1.4884	2.8853	0.9772
	DA	67.25%	62.87%	77.25%	83.83%	33.53%	74.25%	67.07%	82.04%	72.46%	70.06%	65.87%	72.46%	74.25%	66.47%	56.89%	75.45%
	U1	0.0099	0.0101	0.0049	0.0219	0.0251	0.0217	0.0237	0.0179	0.0042	0.0049	0.0075	0.0042	0.0066	0.0091	0.0177	0.0062
	U2	1.7496	2.5892	1.3923	1.6411	3.0821	1.5223	1.8703	1.3734	1.4487	1.7344	1.5603	1.4729	1.9104	2.1059	2.8750	1.8712
	MAPE	1.09%	1.93%	0.97%	2.27%	6.19%	2.51%	3.28%	2.06%	0.61%	0.72%	1.02%	0.59%	1.24%	1.73%	2.78%	1.18%
	R^2	0.9993	0.9991	0.9998	0.9979	0.9973	0.9980	0.9976	0.9986	0.9988	0.9985	0.9963	0.9988	0.9993	0.9986	0.9945	0.9994
Cangzhou	MAE	0.5447	0.5872	0.5178	4.8605	4.9067	4.8833	4.8570	2.0438	0.4579	0.5585	0.7233	0.4391	0.6380	0.8810	1.8394	0.6300
	RMSE	0.7422	0.7952	0.7083	7.2624	7.2642	7.2682	7.2725	3.8192	0.5983	0.7079	1.0983	0.5802	0.8898	1.2163	3.6287	0.8931
	STDE	0.7422	0.7963	0.7061	5.9796	5.9749	5.9578	5.9428	3.8275	0.5990	0.6926	1.0950	0.5813	0.8872	1.2080	3.6191	0.8812
	DA	81.44%	82.63%	83.23%	80.84%	73.89%	77.84%	80.84%	87.31%	80.24%	68.86%	64.07%	82.63%	85.03%	78.44%	68.86%	86.83%
	U1	0.0045	0.0048	0.0043	0.0293	0.0293	0.0294	0.0294	0.0127	0.0040	0.0048	0.0074	0.0039	0.0043	0.0059	0.0176	0.0043
	U2	1.3594	1.3911	1.3065	5.5390	5.6602	5.5440	5.5906	1.8692	1.4269	1.4604	1.5698	1.4152	1.4389	1.4227	2.7705	1.3837
	MAPE	0.72%	0.76%	0.67%	2.83%	2.90%	2.84%	2.81%	1.67%	0.65%	0.80%	0.99%	0.62%	0.66%	0.93%	1.83%	0.65%
	R^2	0.9996	0.9995	0.9996	0.9251	0.9288	0.9386	0.9472	0.9984	0.9992	0.9990	0.9975	0.9993	0.9994	0.9989	0.9904	0.9994

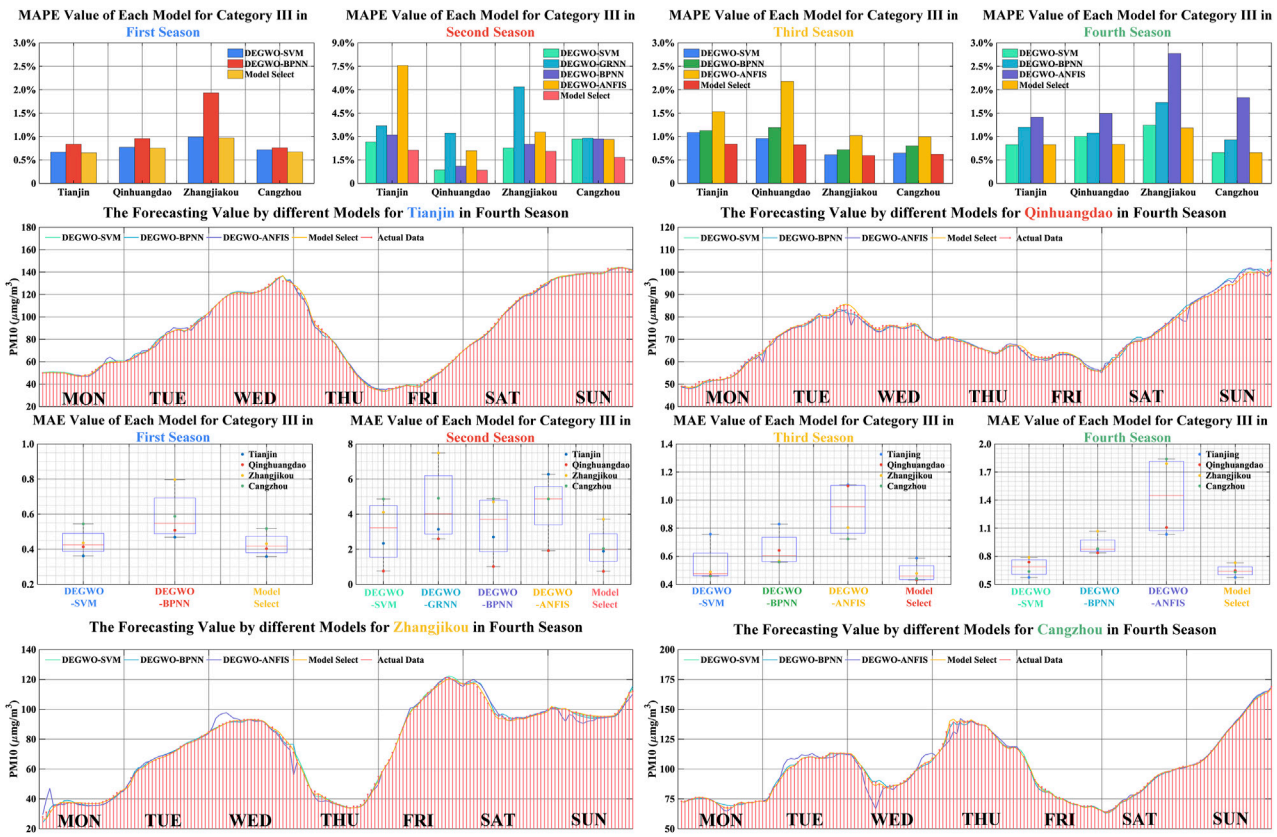


FIGURE 5 | The forecasting result of PM_{10} for Category III in different seasons.

observations and forecasting, which verifies the forecasting performance. In addition, the average MAPE values of model selection for six cities of Category II in four seasons are lower than 1%. Compared with the optimal hybrid model, the model selection is approximately reduced by 10%. The analyses reveal the forecasting superiority of the model selection system. By parity of reasoning, a similar conclusion can be reached through the analyses of the hourly PM_{10} forecasting results for Category III (the forecasting results are shown in **Table 7** and **Figure 5**).

1) For PM_{10} forecasting in the first season, the optimal hybrid models are DEGWO-SVM and DEGWO-BPNN, with which the MAPE values of the best hybrid model (DEGWO-SVM) for four cities of Category III are 0.71%, 0.81%, 1.09%, and 0.72%. The final forecasting values are obtained by model selection; based on the results of DEGWO-SVM and DEGWO-BPNN the MAPE values of model selection are 0.65%, 0.75%, 0.97%, and 0.67%. Additionally, the values of other forecasting metrics are at their best under the model selection.
2) In the second season of PM_{10} forecasting for four cities, the final forecasting results consist of four different hybrid models (DEGWO-SVM, DEGWO-GRNN, DEGWO-BPNN, and DEGWO-ANFIS). From the forecasting performance of four hybrid models, the forecasting performance of DEGWO-SVM is better than the other three models, in which the values of MAPE are 2.65%, 0.86%, 2.27%, and 2.83% for different cities' PM_{10} forecasting in Category III. Additionally, the DA value of MODEGWO-SVM is over 75%, which indicates that the hybrid model can capture future changing trends of PM_{10}. The final forecasting results are obtained by model selection, in which the forecasting accuracy is better than those of the four hybrid models. Compared with the best hybrid model, the MAPE values of model selection are reduced by 25.00%, 2.38%, 10.19%, and 69.46%, respectively.
3) According to forecasting results in **Table 7** and **Figure 5** for PM_{10} of the third season, the three kinds of hybrid models (DEGWO-SVM, DEGWO-BPNN, and DEGWO-ANFIS) are employed to forecast hourly PM_{10}; the DEGWO-SVM has the best forecasting performance among the three hybrid models in Zhangjiakou, and the MAPE is 0.61%. DEGWO-SVM has better forecasting accuracy, and model selection has little improvement on the forecasting accuracy in the final prediction results, but the MAPE has maximum improvement of 29.76% for Tianjin PM_{10} forecasting.
4) According to the results in **Table 6**, the three kinds of hybrid models are used to forecast PM_{10} for four cities of Category III in the fourth season, and the R^2 value of each model was greater than 0.99, which shows that these models have a good forecasting performance for the PM_{10}. Meanwhile, model

selection uses the predicted values of each model to form the final forecasting results, and the corresponding MAE values are 0.5747, 0.6459, 0.7297, and 0.6300 for four cities.

In summary, whether for Category III or the other categories (the results are shown in **Supplementary Appendix S8** and **Supplementary Appendix S9**) PM_{10} forecasting, the model selection system attained the best performance for 13 cities. In the comparison of various hybrid models, the forecasting performance of MODEGWO-SVM is better than other hybrid models. Additionally, it can be observed that some models have low accuracy, which can still provide some optimal forecasting values for the model selection system for hourly PM_{10} forecasting. Based on the above analysis, it can be seen that none of the models has been playing the best forecasting performance in the forecasting process, and various hybrid models are needed to make up for the shortcomings of the single hybrid model.

DISCUSSIONS

This section mainly discusses the hyperparameter related to the SVM and ANN model that would influence the forecasting performance. A large variety of machine learning models and ANNs are available for air pollution time series including three different type pollutants. Finally, compare computing in different model.

Support Vector Machine

According to the results of each experiment, SVM provided the more optimal forecast values for the three main pollutants in the four quarters of 13 cities. The reason for the favorable score produced by SVM is that SVM provides a way to avoid the complexity of high-dimensional space by directly using the inner product function of the space (which is the kernel function) and then directly solving the corresponding decision-making problem in high-dimensional space by using the solution method under the condition of linear separability. When the kernel function is known, it can simplify the difficulty of solving the problem in high-dimensional space. Meanwhile, SVM is based on the small sample statistical theory, which conforms to machine learning. In addition, support vector machine has better generalization ability than neural network. Although the time series of the three main air pollutants are neither regular nor seasonal, SVM can also effectively capture future changes of the three main air pollutants.

SVM has two very important parameters: c and g. c is the penalty coefficient, that is, tolerance of errors. The higher the c, the more intolerable the errors and easy to over-fit. The smaller the c, the less easy fitting is. If c is too large or too small, the generalization ability becomes worse. g is a parameter that comes with RBF function when it is selected as a kernel. Implicitly it determines the distribution of data after mapping to a new feature space. The bigger the g, the less support vectors; it will only act near the support vector samples. For the unknown samples, the classification effect is very poor. There is a possibility that the training accuracy can be very high, and the test accuracy is not high, that is, over-fitting. The smaller the g value is, the more support vectors there are, and the greater the smoothing effect will be; the higher accuracy of the training processing cannot be obtained, and the accuracy of the testing processing will also be affected. This paper used DEGWO algorithm and GWO algorithm to optimize the parameters of SVM (c and g). The results of two types of hybrid SVMs are shown in **Table 8**, which displays that the optimum penalty coefficients of SVM corresponding to pollutant forecasting in different cities vary widely. For example, in the forecasting processing of NO_2, the variation range of parameters is [2, 99]. However, the fluctuation range of g is small, with most variations ranging from 0 to 1. In general, the air pollutants forecasting performances of support vector machine are very dependent on the penalty coefficient. In the whole experiment it can be observed that the support vector machine has good forecasting accuracy for three main air pollutants forecasting, but it cannot provide the best forecasting value in each point. It indicates that the support vector machine is suitable for hourly air pollutants forecasting.

Artificial Neural Network: Number of Input Layer and Number of Hidden Layers

ANN as a nonlinear mapping model is used to solve the problem of time series forecasting, because the ANN model can find the optimal solution of a complex problem with the help of high-speed computing ability of the computer. In order to ensure the forecasting accuracy of the ANN model, parameters of ANN need to be elaborately configured. However, there is no effective rule for establishing the values of these parameters on air pollutants forecasting. Although there are many studies on the tuning of the parameters of the neural network, it is obvious that the selection of the whole parameter space is beyond the scope of this study. Therefore, the parameters of the neural network are set by means of simulation experiment and optimization algorithm: the experimental design is as follows:

This processing configured various input layers and a number of hidden layers to find out the influence of the usage of recent data on the performance of different ANN models. The number of input layers from 1 to 10 increases for three main air pollutants, which means there are 1,008 pieces of sample data on NO_2, $PM_{2.5}$, and PM_{10}; the train-to-verify ratio 5:1 means that 840 pieces of sample data were used as training data for building the ANN model, while 168 pieces of sample data were used as testing data for finding the training-to-testing ratio and parameter of each ANN model (the optimal number of input layers of each model and the number of hidden layers of LSTM and BPNN). **Figure 6** shows the forecasting performance with the different configurations of the optimal number of input layers and the number of hidden layers of LSTM and BPNN, in which it is difficult to find a regular correlation between the forecasting performances and the parameters. Consequently, it is difficult to find an optimal combination of ANN's parameters that brings the model to the best performance in the practical air pollutant forecasting where MAPE and R^2 are unknown.

During the experiment on configurations of ANN's parameter, the optimal parameter setting trained the networks models for each ANN model in 13 cities. The forecasting values with the best

TABLE 8 | The result of two types of SVM for three main air pollutants in different cities.

Air Pollution	Parameter	Beijing	Tianjin	Shijiazhuang	Tangshan	Qinhuangdao	Handan	Baoding	Zhangjiakou	Chengde	Langfang	Cangzhou	Hengshui	Xingtai
DEGWO-SVM (NO_2)	best.*C*	46.9568	99.5518	99.309	99.161	99.8544	9.6662	83.7412	3.5907	16.9145	2.6853	49.9809	99.1334	4.3575
	best.*g*	0.9694	0.0763	0.1945	0.3663	0.0393	0.5403	0.0101	4.0931	0.0105	1.4979	0.0105	0.4098	0.4622
	SSE	$1.74*10^3$	$1.92*10^3$	$3.82*10^3$	$7.76*10^3$	$5.30*10^3$	$7.12*10^3$	$8.78*10^3$	$2.09*10^3$	$3.06*10^3$	$8.96*10^3$	$1.27*10^4$	$7.37*10^3$	$5.62*10^3$
	min.E	0.0338	0.0186	0.0061	0.0336	0.0002	0.0446	0.0216	0.0127	0.0355	0.0077	0.0494	0.0305	0.0259
	max.E	10.7395	13.5376	36.7462	27.3748	20.6095	41.3333	38.0069	13.0612	18.5541	44.2883	34.1065	27.4761	24.1837
	Var.E	10.3228	11.1665	22.8619	46.4519	31.2395	42.479	52.0663	12.3554	18.01	53.4352	75.7013	43.7583	33.6634
	Time	467.0666	493.4147	487.456	503.9043	510.1607	497.2697	474.2788	495.2658	499.3622	474.5837	490.1128	478.4362	492.8458
GWO-SVM (NO_2)	best.*C*	78.5905	324.3692	210.5643	31.6254	212.0154	10.2694	200.0915	33.5771	17.2807	278.5659	50.1119	169.9275	401.8597
	best.*g*	0.9669	0.0765	0.1961	0.3913	0.0395	0.5486	0.01	3.9411	0.1216	1.4618	0.0105	0.4118	0.2077
	SSE	$1.86*10^3$	$1.94*10^3$	$4.12*10^3$	$7.96*10^3$	$5.70*10^3$	$7.67*10^3$	$9.29*10^3$	$2.10*10^3$	$3.33*10^3$	$9.74*10^3$	$1.32*10^4$	$7.90*10^3$	$5.76*10^3$
	min.E	0.0368	0.0187	0.0062	0.0367	0.0003	0.049	0.023	0.0134	0.0383	0.0083	0.0507	0.0321	0.0283
	max.E	11.3172	13.5905	38.0581	29.5744	20.7387	44.7556	38.7649	13.1132	19.3	45.5339	35.1327	28.8798	25.4748
	Var.E	10.612	11.2675	23.2602	50.7895	32.7046	46.1177	53.6559	13.2875	18.7424	56.4917	78.4229	46.1941	34.9107
	Time	414.2995	409.8461	397.1843	439.9873	449.712	434.443	409.6929	405.8517	405.4657	426.6385	438.0274	395.0127	436.3719
DEGWO-SVM ($PM_{2.5}$)	best.*C*	77.7686	0.079	0.1832	32.0261	0.697	4.7476	47.0616	88.2993	99.8909	2.7945	1.4898	0.0972	5.0422
	best.*g*	5.0885	2.6319	7.6937	0.01	3.736	0.01	0.01	0.016	0.0134	4.7102	3.5942	1.9463	0.01
	SSE	$8.67*10^2$	$4.91**10^2$	$7.68*10^2$	$2.42*10^2$	$1.10*10^2$	$1.61*10^3$	$1.71*10^3$	50.63	$1.63*10^2$	$7.55*10^2$	$1.69*10^3$	$1.31*10^3$	$5.85*10^2$
	min.E	0.024	0.0079	0.0308	0.0114	0.0008	0.0034	0.0016	0.0011	0.0039	0.0182	0.0711	0.0363	0.0053
	max.E	4.3326	4.8468	7.4561	3.3869	2.1798	9.1996	9.9057	1.5044	2.5344	5.0337	8.2179	8.8812	6.0958
	Var.E	3.6799	2.544	4.0871	1.4022	0.6552	7.9582	10.6013	0.2743	0.6602	3.8421	9.8701	6.7322	3.4501
	Time	78.5201	79.9691	79.8954	80.0222	81.0965	85.7327	85.5043	80.7387	85.0065	80.3694	84.4306	83.4787	83.4082
GWO-SVM ($PM_{2.5}$)	best.*C*	40.8136	93.0752	8.4678	31.86	60.0461	4.7539	52.6286	87.9288	100	92.1335	17.8947	41.9715	59.3705
	best.*g*	5.0366	1.397	8.7715	0.01	3.7298	0.01	0.01	0.0165	0.0134	4.7102	3.6199	3.306	0.01
	SSE	$8.83*10^2$	$8.08*10^2$	$8.20*10^2$	$2.43*10^2$	$1.11*10^2$	$1.63*10^3$	$1.77*10^3$	50.61	$1.57*10^2$	$7.55*10^2$	$1.68*10^3$	$2.63*10^3$	$6.07*10^2$
	min.E	0.0086	0.0136	0.0147	0.0106	0.0039	0.0263	0.0038	0.0038	0.0013	0.0182	0.0474	0.0024	0.0018
	max.E	4.2931	6.2686	8.4904	3.422	2.1569	9.2235	9.9212	1.5028	2.4958	5.0337	8.1801	7.6894	6.0805
	Var.E	3.7291	4.6545	4.282	1.4157	0.6563	7.9943	10.5457	0.2739	0.6459	3.8421	9.753	9.6812	2.0456
	Time	63.6136	65.0956	67.7701	71.8366	67.2112	70.9237	68.4242	68.8076	68.1498	69.0608	74.1707	71.644	68.8802
DEGWO-SVM (PM_{10})	best.*C*	3.5114	18.3746	4.8387	61.0409	56.5602	0.847	83.9649	51.5595	5.1477	19.6408	12.4613	1.2529	0.553
	best.*g*	4.7786	7.0919	0.01	0.0366	0.0187	0.3343	0.0103	0.01	0.0563	5.9592	3.3231	0.4991	4.3841
	SSE	$2.58*10^3$	$9.35*10^2$	$8.17*10^3$	$1.49*10^3$	$8.61*10^2$	$2.98*10^3$	$6.80*10^3$	$8.81*10^2$	$1.69*10^3$	$1.71*10^3$	$1.72*10^3$	$1.67*10^3$	$1.46*10^3$
	min.E	0.0176	0.0076	0.0036	0.0026	0.0005	0.0778	0.1005	0.0003	0.1025	0.0015	0.0039	0.03	0.0134
	max.E	10.9659	8.4861	14.0939	8.4406	5.9929	10.8543	15.9743	6.9388	7.4138	10.0266	10.3846	7.3109	8.9293
	Var.E	10.2272	5.1871	28.3304	8.9646	5.1568	17.5229	33.6806	3.9992	7.9292	5.211	14.8334	9.6962	8.9122
	Time	81.873	85.2247	85.6066	84.1662	79.4368	80.1318	80.0802	86.2987	82.036	83.7506	82.7005	81.4053	85.2475
GWO-SVM (PM_{10})	best.*C*	5.9521	18.4212	0.6008	56.4077	82.7556	0.8624	88.0444	51.7353	5.124	67.6243	5.4252	1.2881	9.1351
	best.*g*	4.7723	7.0158	0.0101	0.0367	0.0187	0.3284	0.0101	0.0101	0.0565	5.9744	0.01	0.4842	0.01
	SSE	$2.61*10^3$	$9.43*10^2$	$8.30*10^3$	$1.50*10^3$	$8.65*10^2$	$3.00*10^3$	$6.88*10^3$	$9.74*10^2$	$1.69*10^3$	$1.72*10^3$	$3.11*10^3$	$1.68*10^3$	$1.51*10^3$
	min.E	0.0249	0.0083	0.0207	0.0192	0.0528	0.081	0.1022	0.079	0.1026	0.0071	0.0256	0.066	0.0899
	max.E	10.9762	8.5423	14.7419	8.4423	6.0089	10.9071	16.0619	6.9653	7.416	10.0514	11.5914	7.4373	12.3556
	Var.E	10.2477	5.2718	38.3341	8.8981	5.1728	17.469	33.8092	4.1143	7.9363	5.2235	10.2618	9.5788	8.6932
	Time	73.6691	70.3273	73.1108	72.0992	69.9234	64.3441	65.0347	75.2683	71.9299	72.8245	69.424	65.4326	71.6009

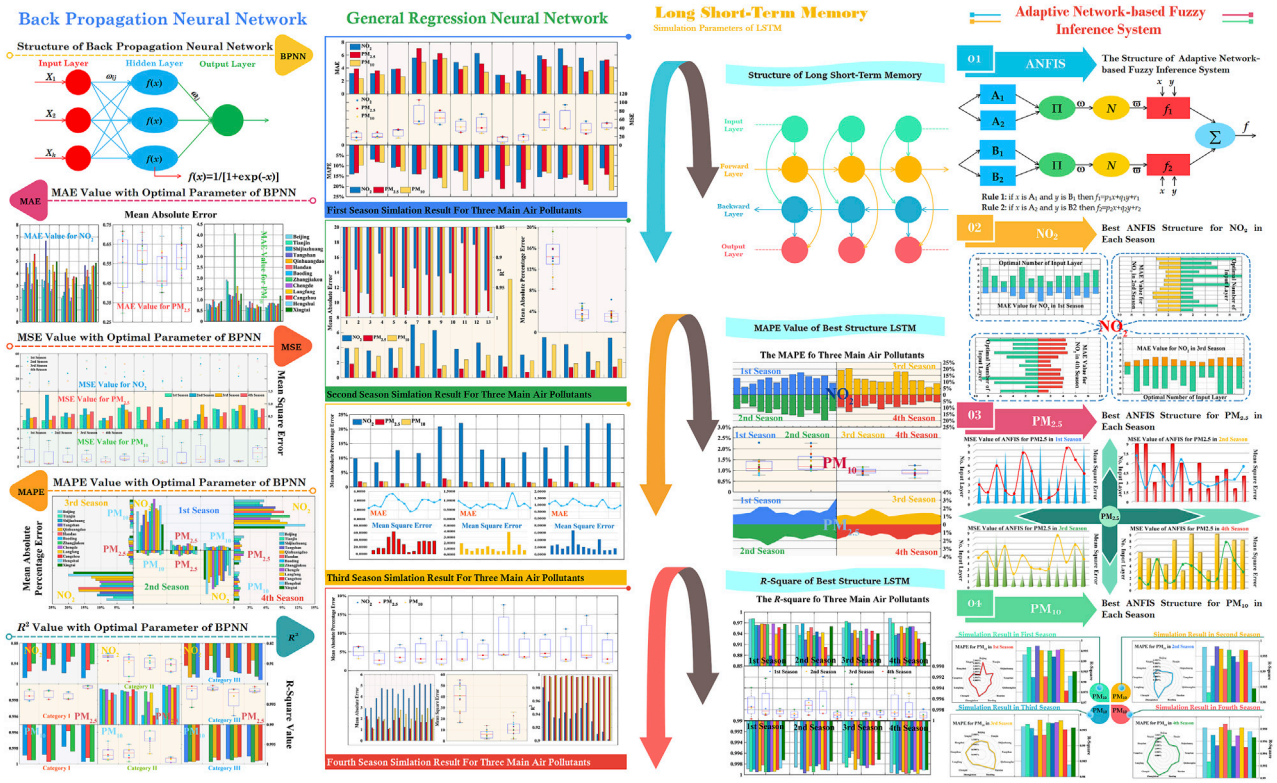


FIGURE 6 | The simulation result of each ANN model for three main air pollutants in different cities.

performance of each ANN (the best forecasting metric in **Table 9**) were selected to forecast three main air pollutants in each experiment. However, it can be found that there were no giant differences on forecasting performance among the networks trained with the same configuration, even if the neural network has the randomness and probability mechanism inside the training processing. A large sample of the times series is another reason that the training stability of the neural network can be ensured. For example, the best MAPE of BPNN for the forecasting of NO_2 in Beijing shown in **Figure 6** is 5.82%, and the worst one is 8.94%. Most of the MAPEs are between 6% and 7%. It is practical to use ANN in real air pollutants forecasting application where forecasting the changing air pollutant time series is suitable.

In summary, with the rapid development of ANN, it has become a powerful tool to solve prediction problems. Neural network is used in the field of air pollution to solve the problem of non-linear forecasting which cannot be solved by statistical models. Its non-linear mapping is especially suitable for the application of air pollutant forecasting. The main reason is that the ANN has the following advantages:

1) Non-linear mapping ability: ANN realizes a mapping function from input to output in essence. Mathematical theory proves that three-layer neural network can approximate any non-linear continuous function with arbitrary precision. This makes it especially suitable for solving complex internal mechanism problems, that is, ANN has strong non-linear mapping ability.
2) Self-learning and self-adapting ability: ANN can automatically extract the "reasonable rules" between output and output data by learning and self-adaptively memorizing the learning content in the weights of the network. ANN has a high ability of self-learning and self-adaptation.
3) Generalization ability: When designing pattern classifiers, the so-called generalization ability refers to whether the network can forecast the unknown time series correctly after training, while ensuring that the classified objects are correctly classified. ANN can apply learning results to new knowledge.
4) Fault-tolerant ability: ANN will not have a great impact on the global training results after its local or partial neurons are damaged; the system can work normally even when it is damaged locally. ANN has certain fault-tolerant ability.

Computing Time for Each Model

In order to improve the computing efficiency and save the computing time, training and forecasting processing of all the models for the main air pollutants time series with parallel computing by central processing unit (CPU) and graphics processing unit (GPU). The computing times of every independent hybrid model in each experiment are shown in **Table 10**, from which we can further research the computational efficiency of the developed model selection forecasting system for the main air pollutants. Specifically, the

TABLE 9 | The simulation result of each ANN model.

Model	BPNN						ANFIS				
Metric	**NO. Input**	**NO. Hidden**	**MAE**	**MSE**	**MAPE**	R^2	**NO. Input**	**MAE**	**MSE**	**MAPE**	R^2
Beijing	7	29	2.7515	13.4095	12.82%	0.9825	7	2.2208	14.1852	11.61%	0.9816
Tianjin	4	21	2.6124	12.3784	6.07%	0.9830	4	1.8908	18.6934	5.00%	0.9747
Shijiazhuang	7	23	3.2795	28.5628	8.32%	0.9628	2	1.7853	9.7964	5.20%	0.9876
Tangshan	6	15	4.8308	49.0277	11.76%	0.9583	4	3.2171	25.9431	8.18%	0.9779
Qinhuangdao	8	24	4.4735	38.3830	13.34%	0.9672	7	4.2183	33.3884	14.23%	0.9722
Handan	10	21	3.9385	30.7437	9.47%	0.9602	3	1.9940	16.6067	5.23%	0.9788
Baoding	9	26	4.8845	54.5972	12.89%	0.9626	4	5.1107	52.7317	16.64%	0.9637
Zhangjiakou	2	21	2.6965	14.1237	15.52%	0.8956	2	3.0306	16.3035	19.68%	0.8805
Chengde	5	21	3.1457	21.7384	17.12%	0.9642	4	2.0878	12.3017	13.23%	0.9799
Langfang	4	21	5.0108	53.2976	13.89%	0.9393	4	4.2474	37.6082	13.43%	0.9602
Cangzhou	3	24	5.5917	58.3720	13.31%	0.9547	4	3.0234	31.2348	9.23%	0.9759
Hengshui	7	23	4.5165	37.0716	14.85%	0.9482	5	3.7816	33.4263	11.99%	0.9550
Xingtai	9	21	3.4928	30.4963	8.57%	0.9702	6	2.4056	28.1033	5.40%	0.9725
Model	**LSTM**						**GRNN**				
Metric	**NO. Input**	**NO. Hidden**	**MAE**	**MSE**	**MAPE**	R^2	**NO. Input**	**MAE**	**MSE**	**MAPE**	R^2
Beijing	8	27	2.8379	15.0298	14.37%	0.9805	2	3.1653	17.7058	14.09%	0.9772
Tianjin	9	9	2.6240	13.6968	6.72%	0.9812	8	3.1290	20.3228	6.97%	0.9720
Shijiazhuang	3	30	3.1771	28.2116	8.22%	0.9634	9	3.8065	34.9429	10.99%	0.9577
Tangshan	5	21	4.7107	43.2069	10.36%	0.9630	10	5.5704	55.6046	14.00%	0.9522
Qinhuangdao	4	30	4.5057	37.7889	13.13%	0.9673	7	5.2814	47.7487	16.22%	0.9590
Handan	6	29	3.6311	25.9993	8.01%	0.9663	2	4.8767	60.1796	12.32%	0.9271
Baoding	8	9	4.7872	50.7279	11.98%	0.9651	5	6.2932	72.5647	16.31%	0.9538
Zhangjiakou	7	29	2.7874	17.0202	18.70%	0.8783	8	2.9293	17.9413	16.74%	0.8711
Chengde	1	29	3.1558	20.0247	15.77%	0.9673	3	3.5448	25.0984	18.05%	0.9588
Langfang	10	30	5.0278	53.3276	13.90%	0.9383	10	5.9260	76.8659	15.51%	0.9170
Cangzhou	8	30	5.3683	56.0168	12.77%	0.9570	7	6.9785	94.2087	16.45%	0.9265
Hengshui	10	27	4.3802	36.2420	14.52%	0.9523	6	5.5657	56.0395	16.84%	0.9239
Xingtai	4	27	3.8483	33.5689	9.43%	0.9670	9	5.1267	52.2457	11.13%	0.9551

TABLE 10 | Computing time by each model.

Category	Model	Min	Max	Average	Model	Min	Max	Average
Category I (NO_2)	MODEGWO-SVM	39.4238	48.1795	43.9663	**First season for Category III (PM_{10})**			
	MODEGWO-GRNN	65.7006	80.2975	72.8432	DEGWO-SVM	32.49657	39.71566	36.03983
	MODEGWO-BPNN	65.7133	80.2837	72.7285	DEGWO-BPNN	54.17679	66.18566	60.13886
	MODEGWO-ANFIS	39.4207	48.1762	43.7417	Model selection	0.108317	0.132383	0.120239
	Model select	0.1314	0.1605	0.1460	Final time	87.15716	105.6679	96.29893
	Final time	214.8663	251.5786	233.4256	**Second season for Category III (PM_{10})**			
Category II (NO_2)	MODEGWO-SVM	44.5487	54.4188	49.4898	DEGWO-SVM	43.09741	52.62296	47.78303
	MODEGWO-GRNN	74.2109	90.6824	82.7233	DEGWO-GRNN	71.80499	87.73968	79.55603
	MODEGWO-BPNN	74.2103	90.7002	82.6234	DEGWO-BPNN	43.08113	52.63649	48.01184
	Adam-LSTM	133.5967	163.2337	148.2152	DEGWO-ANFIS	71.80864	87.74184	79.77636
	Model select	0.1484	0.1813	0.1650	Model selection	0.143618	0.175492	0.159656
	Final time	330.6037	393.0909	363.2167	Final time	235.9748	276.3926	255.2869
Category III (NO_2)	MODEGWO-SVM	41.5062	50.7174	46.0609	**Third season for Category III (PM_{10})**			
	MODEGWO-GRNN	69.1727	84.5312	76.9191	DEGWO-SVM	37.79408	46.17638	42.03216
	MODEGWO-BPNN	69.1691	84.5052	76.7223	DEGWO-BPNN	62.98502	76.9619	70.11255
	Model select	0.1383	0.1689	0.1537	DEGWO-ANFIS	75.59966	92.35553	83.84003
	Final time	183.1891	218.0664	199.8560	Model selection	0.125959	0.15388	0.139823

(Continued on following page)

TABLE 10 | (*Continued*) Computing time by each model.

Category	Model	Min	Max	Average	Model	Min	Max	Average
Category	Model	Min	Max	Average	Final time	178.6782	212.5414	196.1246
Category II ($PM_{2.5}$)	MODEGWO-SVM	42.00122	51.30858	46.52758	**Fourth season for Category III (PM_{10})**			
	MODEGWO-BPNN	69.98488	85.5151	77.81983	DEGWO-SVM	37.68701	46.0525	41.96226
	MODEGWO-ANFIS	83.97059	102.6216	93.34656	DEGWO-BPNN	62.81456	76.76278	69.93992
	Adam-LSTM	125.9622	153.9263	139.7002	DEGWO-ANFIS	75.37328	92.07837	83.83009
	Model select	0.139947	0.171026	0.155573	Model selection	0.125629	0.153373	0.139332
	Final time	326.5897	387.7865	357.5498	Final time	176.9772	213.5771	195.8716

average computation time of the model selection forecasting system ranges from 330.6037 to 363.2167 s for NO_2 forecasting in Category II, with the longest computing time appearing in the different categories. Notably, Adam-LSTM with complex model structure has longer computing time than the other hybrid models, taking more time in the iterative optimization process. Moreover, this paper establishes multiple hybrid models and uses the model selection method to find the best forecasting value, in which the final forecasting accuracy is improved but needs more computing time.

CONCLUSION

In this study, a novel model selection forecast system was proposed that overcomes the shortcomings of the single hybrid model, which cannot give the optimal results for the forecasting process. First, the FSE theory is employed to analyze the major pollutant for each city in BJ-TJ-HE, and the fuzzy c-means algorithm is used to analyze the feature of the 13 cities. Then, to further improve modeling accuracy and rationality of modeling, a modified optimization algorithm (DEGWO) was used to optimize the premasters of different models. Finally, the model selection forecasting system obtains forecasting results at each time point from different hybrid models.

The developed model selection forecasting system was evaluated on hourly NO_2, $PM_{2.5}$, and PM_{10} from 13 cities, and several performance metrics were calculated, with experimental results indicating that the model selection forecasting system is superior to single hybrid models with the smallest MAPE in the different cities pollutant forecasting, indicating its strong forecasting performance. Overall, the proposed model selection forecast system exhibits outstanding performance in data analysis and time series forecasting for air pollutants. Specifically, it can not only deeply analyze major pollutants of AQI for BJ-TJ-HE but also approximate the actual values with high accuracy and stability.

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 authors.

AUTHOR CONTRIBUTIONS

The design experiment, data analysis, and paper writing were conducted by YH; the forecasting experiment and data analysis were completed by YH and CW; supervision, paper writing, and editing were conducted by CW and YD; validation, methodology, paper editing, and supervision were handled by QL and GZ. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by Western Project of the National Social Science Foundation of China (Grant No. 18XTJ003).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenvs.2021.761287/full#supplementary-material

REFERENCES

Akyüz, M., and Çabuk, H. (2009). Meteorological Variations of PM2.5/PM10 Concentrations and Particle-Associated Polycyclic Aromatic Hydrocarbons in the Atmospheric Environment of Zonguldak, Turkey. *J. Hazard. Mater.* 170, 13–21. doi:10.1016/j.jhazmat.2009.05.029

Bessagnet, B., Couvidat, F., and Lemaire, V. (2019). A Statistical Physics Approach to Perform Fast Highly-Resolved Air Quality Simulations – A New Step towards the Meta-Modelling of Chemistry Transport Models. *Environ. Model. Softw.* 116, 100–109. doi:10.1016/j.envsoft.2019.02.017

Brereton, R. G., and Lloyd, G. R. (2010). Support Vector Machines for Classification and Regression. *Analyst* 135, 230–267. doi:10.1039/b918972f

D'Allura, A., Kulkarni, S., Carmichael, G., Finardi, S., Adhikary, Wei, et al. (2011). Meteorological and Air Quality Forecasting Using the WRF–STEM Model during the 2008 ARCTAS Field Campaign. *Atmos. Environ.* 45. doi:10.1016/j.atmosenv.2011.02.073

Dass, A., Srivastava, S., and Chaudhary, G. (2021). Air Pollution: A Review and Analysis Using Fuzzy Techniques in Indian Scenario. *Environ. Technology Innovation* 22, 101441. doi:10.1016/j.eti.2021.101441

Davood, N.-K., Goudarzi, G., Taghizadeh, R., Asumadu-Sakyi, A., and Fehresti-Sani, M. (2021). Long-term Effects of Outdoor Air Pollution on Mortality and

Morbidity–Prediction Using Nonlinear Autoregressive and Artificial Neural Networks Models. *Atmos. Pollut. Res.* 12, 46–56. doi:10.1016/j.apr.2020.10.007

Dhiman, G., and Kumar, V. (2018). Multi-objective Spotted Hyena Optimizer: A Multi-Objective Optimization Algorithm for Engineering Problems. *Knowledge-Based Syst.* 150, 175–197. doi:10.1016/j.knosys.2018.03.011

Díaz-Robles, L. A., Ortega, J. C., Fu, J. S., Reed, G. D., Chow, J. C., Watson, J. G., et al. (2008). A Hybrid ARIMA and Artificial Neural Networks Model to Forecast Particulate Matter in Urban Areas: The Case of Temuco, Chile. *Atmos. Environ.* 42, 8331–8340. doi:10.1016/j.atmosenv.2008.07.020

Feng, X., Li, Q., Zhu, Y., Hou, J., Lingyan, J., and Wang, J. (2015). Artificial Neural Network Forecasting of PM2.5 Pollution Using Air Mass Trajectory Based Geographic Model and Wavelet Transformation. *Atmos. Environ.* 107, 118–128. doi:10.1016/j.atmosenv.2015.02.030

Gayen, S., and Biswas, A., (2021). Pythagorean Fuzzy C-Means Clustering Algorithm. In *Computational Intelligence in Communications and Business Analytics Cham*. Editors J. K. Mandal and S. Mukhopadhyay.

Güçlü, Y. S., Dabanlı, İ., Şişman, E., and Şen, Z. (2019). Air Quality (AQ) Identification by Innovative Trend Diagram and AQ index Combinations in Istanbul Megacity. *Atmos. Pollut. Res.* 10, 88–96. doi:10.1016/j.apr.2018.06.011

Han, Y., Lam, J. C., Li, V. O., and Reiner, D. (2021). A Bayesian LSTM Model to Evaluate the Effects of Air Pollution Control Regulations in Beijing, China. *Environ. Sci. Pol.* 115, 26–34. doi:10.1016/j.envsci.2020.10.004

Hao, Y., Niu, X., and Wang, J. (2021). Impacts of Haze Pollution on China's Tourism Industry: A System of Economic Loss Analysis. *J. Environ. Manag.* 295, 113051. doi:10.1016/j.jenvman.2021.113051

Hao, Y., and Tian, C. (2018). The Study and Application of a Novel Hybrid System for Air Quality Early-Warning. *Appl. Soft Comput.*, 74.

Hao, Y., Tian, C., and Wu, C. (2019). Modelling of Carbon price in Two Real Carbon Trading Markets. *J. Clean. Prod.* 244, 118556.

Jang, J.-S. R. (1993). Anfis: Adaptive-Network-Based Fuzzy Inference System. *IEEE Trans. Syst. Man. Cybern.* 23, 665–685. doi:10.1109/21.256541

Jiang, P., Dong, Q., and Li, P. (2017). A Novel Hybrid Strategy for PM 2.5 Concentration Analysis and Prediction. *J. Environ. Manage.* 196, 443–457. doi:10.1016/j.jenvman.2017.03.046

Kim, J., Kim, D., Choi, M., and Choi, M. (2018). Impact of Air Pollution on Cause-specific Mortality in Korea: Results from Bayesian Model Averaging and Principle Component Regression Approaches. *Sci. total Environ.* 636, 1020–1031.

Kumar, K., Zindani, D., and Davim, J. P. (2019). Particle Swarm Optimization Algorithm. *Optimizing Engineering Problems through Heuristic Techniques*. 23–32. doi:10.1201/9781351049580-3

Land, W. H., and Schaffer, J. D. (2020). The Generalized Regression Neural Network Oracle. *The Art and Science of Machine Intelligence*, 77–105. doi:10.1007/978-3-030-18496-4_3

Lanzafame, R., Monforte, P., Patanè, G., and Strano, S. (2015). Trend Analysis of Air Quality Index in Catania from 2010 to 2014. *Energ. Proced.* 82, 708–715. doi:10.1016/j.egypro.2015.11.796

Latif, M. T., Ahamad, F., Abdul Halim, N. D., Khan, M., Juneng, L., Chung, J. X., et al. (2018). The Long-Term Assessment of Air Quality on an Island in Malaysia. *Heliyon* 4, 1–33.

Lauriks, T., Longo, R., Baetens, D., Derudi, M., Parente, A., Bellemans, A., et al. (2020). Application of Improved CFD Modeling for Prediction and Mitigation of Traffic-Related Air Pollution Hotspots in a Realistic Urban Street. *Atmos. Environ.* 246, 118127. doi:10.1016/j.atmosenv.2020.118127

Leong, W. C., Kelani, R. O., and Ahmad, Z. (2019). Prediction of Air Pollution index (API) Using Support Vector Machine (SVM). *J. Environ. Chem. Eng.* 8, 103208.

Li, W. k., Wang, W., Wang, Z., and Li, L. (2019). Opposition-based Multi-Objective Whale Optimization Algorithm with Global Grid Ranking. *Neurocomputing* 5. doi:10.1016/j.neucom.2019.02.054

Li, Z., Fung, J. C. H., and Lau, A. K. H. (2018). High Spatiotemporal Characterization of On-Road PM2.5 Concentrations in High-Density Urban Areas Using mobile Monitoring. *Building Environ.* 143, 196–205. doi:10.1016/j.buildenv.2018.07.014

Liu, H., Wu, H., Lv, X., Ren, Z., Liu, M., Li, Y., et al. (2019). An Intelligent Hybrid Model for Air Pollutant Concentrations Forecasting: Case of Beijing in China. *Sustainable Cities Soc.* 47, 101471. doi:10.1016/j.scs.2019.101471

Liu, Q., Wu, L., Xiao, W., Wang, F., and Zhang, L. (2018). A Novel Hybrid Bat Algorithm for Solving Continuous Optimization Problems. *Appl. Soft Comput.* 73. doi:10.1016/j.asoc.2018.08.012

Liu, Y., Guo, H., Mao, G., and Yang, P. (2008). A Bayesian Hierarchical Model for Urban Air Quality Prediction under Uncertainty. *Atmos. Environ.* 42, 8464–8469. doi:10.1016/j.atmosenv.2008.08.018

Lu, D.-X., Weng, W.-Y., Su, J., Wang, Z.-B., and Yang, X.-J. (2011). An Improved Fuzzy Synthetically Evaluation 163, 214–222. doi:10.1007/978-3-642-25002-6_30

Muzaffar, S., and Afshari, A. (2019). Short-Term Load Forecasts Using LSTM Networks. *Energ. Proced.* 158, 2922–2927. doi:10.1016/j.egypro.2019.01.952

Pan, L., Sun, B., and Wang, W. (2011). City Air Quality Forecasting and Impact Factors Analysis Based on Grey Model. *Proced. Eng.* 12, 74–79. doi:10.1016/j.proeng.2011.05.013

Pereira, A., Patrício, B., Fonte, F., Marques, S., Reis, C. I., and Maximiano, M. (2018). Collecting Information about Air Quality Using Smartphones. *Proced. Computer Sci.* 138, 33–40. doi:10.1016/j.procs.2018.10.006

Rivas, E., Santiago, J., Lechon, Y., Martin, F., Ariño, A., Pons Izquierdo, J., et al. (2018). CFD Modelling of Air Quality in Pamplona City (Spain): Assessment, Stations Spatial Representativeness and Health Impacts Valuation. *Sci. Total Environ.* 649, 1362–1380. doi:10.1016/j.scitotenv.2018.08.315

Shenfield, A., and Rostami, S. (2015). *A Multi Objective Approach to Evolving Artificial Neural Networks for Coronary Heart Disease Classification*. 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Auguat 12–15, 2015, 1–8. doi:10.1109/CIBCB.2015.7300294

Shanshan, Q., Liu, F., Wang, J., and Sun, B. (2014). Analysis and Forecasting of the Particulate Matter (PM) Concentration Levels over Four Major Cities of China Using Hybrid Models. *Atmos. Environ.* 98, 665–675. doi:10.1016/j.atmosenv.2014.09.046

Steffens, J., Kimbrough, S., Baldauf, R., Isakov, V., Brown, R., Powell, A., et al. (2017). Near-port Air Quality Assessment Utilizing a mobile Measurement Approach. *Atmos. Pollut. Res.* 8, 1023–1030. doi:10.1016/j.apr.2017.04.003

Urbancok, D., Payne, A., and Webster, R. (2017). Regional Transport, Source Apportionment and Health Impact of PM 10 Bound Polycyclic Aromatic Hydrocarbons in Singapore's Atmosphere. *Environ. Pollut.* 229, 984–993.

Xu, Y., Yang, W., and Wang, J. (2016). Air Quality Early-Warning System for Cities in China. *Atmos. Environ.* 148, 239–257. doi:10.1016/j.atmosenv.2016.10.046

Yang, Z., and Wang, J. (2017). A New Air Quality Monitoring and Early Warning System: Air Quality Assessment and Air Pollutant Concentration Prediction. *Environ. Res.* 158, 105–117. doi:10.1016/j.envres.2017.06.002

You, H., Ma, Z., Tang, Y., Wang, Y., Yan, J., Ni, M., et al. (2017). Comparison of ANN (MLP), ANFIS, SVM, and RF Models for the Online Classification of Heating Value of Burning Municipal Solid Waste in Circulating Fluidized Bed Incinerators. *Waste Management* 68, 186–197. doi:10.1016/j.wasman.2017.03.044

Zhang, L., Lin, J., Qiu, R., Hu, X., Zhang, H., Chen, Q., et al. (2018). Trend Analysis and Forecast of PM2.5 in Fuzhou, China Using the ARIMA Model. *Ecol. Indicators* 95, 702–710. doi:10.1016/j.ecolind.2018.08.032

Zhou, Q., Jiang, H., Wang, J., and Zhou, J. (2014). A Hybrid Model for PM 2.5 Forecasting Based on Ensemble Empirical Mode Decomposition and a General Regression Neural Network. *Sci. Total Environ.* 496, 264–274. doi:10.1016/j.scitotenv.2014.07.051

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.

Copyright © 2021 Huang, Deng, Wang and Fu. 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 in Marine Science

ORIGINAL RESEARCH
published: 25 January 2022
doi: 10.3389/fmars.2021.804541

Modeling of Direct Economic Losses of Storm Surge Disasters Based on a Novel Hybrid Forecasting System

Hongbo Guo [1,2], *Kedong Yin* [1,2] *and Chong Huang* [1,2*]

[1] School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan, China, [2] Institute of Marine Economy and Management, Shandong University of Finance and Economics, Jinan, China

Modeling the direct economic losses of storm surge disasters can assess the disaster situation in a timely manner and improve the efficiency of post-disaster management in practice, which is acknowledged as one of the most significant issues in clean production. However, improving the forecasting accuracy of direct economic losses caused by storm surge disasters remains challenging, which is also a major concern in the field of disaster risk management. In particular, most of the previous studies have mainly focused on individual models, which ignored the significance of reduction and optimization. Therefore, a novel direct economic loss forecasting system for storm surge disasters is proposed in this study, which includes reduction, forecasting, and evaluation modules. In this system, a forecasting module based on an improved machine learning technique is proposed, which improves the generalization ability and robustness of the system. In addition, the key attributes and samples are selected by the proposed reduction module to further improve the forecasting performance from the two innovative perspectives. Moreover, an evaluation module is incorporated to comprehensively evaluate the superiority of the developed forecasting system. Data on the storm surge disasters from three typical provinces are utilized to conduct a case study, and the performance of the proposed forecasting system is analyzed and compared with eight comparison models. The experimental results show that the mean absolute percentage error (MAPE) predicted by the Extreme Learning Machine (ELM) model was 16.5293%, and the MAPE predicted by the proposed system was 1.0313%. Overall, the results show that the performance of the proposed forecasting system is superior compared to other models, and it is suitable for the forecasting of direct economic losses resulting from storm surge disasters.

Keywords: storm surge, hybrid forecasting, forecasting, optimization algorithm, economic losses

OPEN ACCESS

Edited by:
Yiliao Song,
University of Technology Sydney, Australia

Reviewed by:
Hongmin Li,
Northeast Forestry University, China
Yurui Fan,
Brunel University London, United Kingdom

***Correspondence:**
Chong Huang
huangchong@sdufe.edu.cn

Specialty section:
This article was submitted to Global Change and the Future Ocean, a section of the journal Frontiers in Marine Science

Received: *29 October 2021*
Accepted: *16 December 2021*
Published: *25 January 2022*

Citation:
Guo H, Yin K and Huang C (2022) Modeling of Direct Economic Losses of Storm Surge Disasters Based on a Novel Hybrid Forecasting System. Front. Mar. Sci. 8:804541. doi: 10.3389/fmars.2021.804541

INTRODUCTION

Four portions are introduced in the section. The main issues are introduced in the first part, the second reviews the related literature and works, the third presents the main contributions, whereas the fourth provides the structure of this article.

Main Issues

Marine disasters caused by abnormal marine environments or extreme climate change have a significant impact on marine security, economic and social development, and clean production.

However, storm surge disasters have resulted in a considerable threat to human life and production, causing the most serious impact. The China Marine Disaster Bulletin (China Marine Disaster Bulletin, 2020–2021) shows that the total loss due to marine disasters was about 11.64 billion in 2019 and 832 million in 2020, of which the direct economic losses caused by storm surge disasters account for 99.44 and 97.36%, respectively. The importance of research on the economic impact of storm surge disasters has recently become clear (Neumann et al., 2015; Guo and Li, 2020; Zhou et al., 2020). The rapid and accurate forecasting of losses caused by storm surge disasters can provide strategic decision support for the prevention and reduction of disasters, disaster situation warnings, and rescue management. However, due to the particularity and irregularity of natural disasters, it is difficult to rapidly and accurately forecast marine disaster losses. Therefore, the development of an appropriate method to model the economic losses caused by storm surge disasters is required, as this is widely considered a key concern in clean production (Yang et al., 2019a).

Literature Review

In disaster loss assessment and forecasting, scholars have mainly focused on earthquakes (Jena et al., 2020; Kim et al., 2020; Pulinets et al., 2021), tropical cyclones (Qi and Du, 2018; Sawant et al., 2019; Giffard-Roisin et al., 2020; Zeng et al., 2021), and floods (Zhi et al., 2020; Soltani et al., 2021), but few studies have been conducted on storm surge. With the background and impact of global climate change (Hao et al., 2021), the problem of marine disasters (Fang et al., 2017; Yan et al., 2020) is becoming increasingly obvious. Among the losses caused by marine disasters, storm surge disaster losses account for a large proportion, so it is necessary to pay attention to the research on storm surge disasters (Arns et al., 2015). Most studies have focused on storm surge and its forecasting methods (Sahoo and Bhaskaran, 2018; Ohz et al., 2020), but only few on direct economic loss assessment. Three methods are commonly adopted in this forecasting field. (1) In previous studies, physical methods were mainly used in data monitoring, early warning, and forecasting. Meanwhile, a large number of data were outputted to assist the research and the forecasting (Lakshmi et al., 2017; Nahornyi et al., 2021). Physical models employ meteorological information or related physical information as input, and require a great deal of historical information to judge the specific relationship between disaster sequences and physical information to perform forecasting. However, the amount of data comprising this information is large and difficult to obtain, and consequently forecasting is difficult. (2) Statistical methods such as autoregressive integrated moving average (ARIMA) and generalized autoregressive conditional heteroskedasticity (GARCH) have exhibited good results in solving low-dimensional weakly non-linear problems (Yi et al., 2021). However, disasters themselves are complex problems, whereas disaster loss assessment and forecasting are multi-dimensional and non-linear problems. The traditional statistical methods have obvious disadvantages. (3) In machine learning methods, machine learning algorithms have shown stronger ability to fit complex non-linear data, and, thus, have been highly regarded. Machine learning methods had been widely applied in some forecasting fields such as electricity price forecasting (Yang et al., 2022). Support-vector machine (SVM), extreme learning machine (ELM), and backpropagation neural networks (BPNN) are commonly used forecasting methods. Xiong et al. (2018) successfully improved ELM to predict the seasonal price of vegetables. Liu et al. (2020) proposed a combined model using self-organizing map (SOM), kernel principal components analysis (KPCA), and an SVM to classify and to forecast the patent quality in the biomedical industry. Li et al. (2021) proposed the gray-BPNN model to predict the grain output of Henan Province. These three methods have been proposed to form different models for forecasting, such as single model, hybrid model (Sahin, 2019; Wang et al., 2019b; Yang et al., 2019b), combined model (Niu and Wang, 2019; Wang et al., 2019a), and ensemble model (Hao and Tian, 2019).

In the field of marine disaster forecasting, single models have been widely proposed. Wang et al. (2021) proposed GIS and open data to quantitatively evaluate the storm surge and to estimate the direct physical losses. Yin et al. (2017) established a gray relational model based on the panel data dispersion, which was applied to the study of storm surge disaster losses in the coastal areas of China. Hybrid models have been widely proposed in the field of disaster forecasting. Young et al. (2017) combined the traditional physical hydrological simulation method with the SVR to form a hybrid forecasting model in predicting the hourly runoff of Chishan River Basin in southern Taiwan. Compared with the physical hydrological model, ANN, SVR, and two hybrid models (HEC-HMS-ANN, HEC-HMS-SVR) which were based on a hydrologic modeling system, the novel model exhibited some advantages, especially a higher accuracy in long-term forecasting. However, the individual model research methods on the economic impact of disasters are relatively inaccurate. The combination model was developed to improve the forecasting accuracy. Chen et al. (2018) introduced a new combined model method to forecast disaster losses caused by tropical cyclones, hence, using the model combination method, GA-Elman neural network, SVR, and GRNN were combined into a comprehensive evaluation model. The results showed that their proposed model performed better than the single model. Feng and Liu (2017) established an index system with gray correlation analysis. Compared with the single model, the combined model of BP and SVM was proposed to better forecast the direct economic losses and the number of populations affected by storm surge. Zhao et al. (2020) combined the results of ENN and GRNN with a definite integral to achieve interval forecasting, and obtained a large number of high-precision results of the annual storm surge disaster economic losses. At present, ensemble models are widely proposed to reduce bias and to improve forecasting accuracy (Liu and Xu, 2020; Bravo and Ayuso, 2021), but the application of this model in the field of marine disaster forecasting remains rare (Ding et al., 2020). Zhao et al. (2019) proposed an ensemble learning model called Adaboost-BPNN which is designed to forecast direct economic losses of marine disasters. Considering the interaction between pressure, topographic constraints, and the resonant

characteristics of the basin, Žust et al. (2020) proposed an integrated sea level forecasting method called HIDRA. In the field of direct economic losses of storm surge disasters forecasting, those methods have been developed gradually, but they ignore the significance of attributes and of sample reduction, as well as the model optimization.

Primary Work and Contributions

Therefore, with the goal of exceeding the limitations of the above-mentioned models and methods, a novel hybrid forecasting system is proposed herein to forecast the direct economic losses of storm surge disasters. In contrast to most of the previous studies, it improves the forecasting performance by considering the complexity of the loss factors and the similarity of data. More specifically, the system consists of three modules, including a reduction module for data processing, a forecasting module, and an evaluation module. Specifically, the key attributes in the economic loss assessment attribute set of storm surge disasters are selected by the reduction module to obtain the optimal input of each sample. Subsequently, the forecasting module is designed to obtain the forecasting loss results for each storm surge disaster based on the optimal input. Afterward, the evaluation module comprehensively evaluates the performance of the developed forecasting system. To test the feasibility and the superiority of the forecasting system, real data samples of direct economic losses of storm surge disasters from 1989 to 2019 were collected for numerical experiments. The experimental results showed that the proposed approach exhibited good performance, which was superior compared to that of other models, and it is suitable for the forecasting of direct economic losses resulting from storm surge disasters in practical applications.

The main contributions of this study are summarized as follows:

(1) ***Modeling of direct economic losses of storm surge disasters is achieved by a novel hybrid forecasting system.*** Previous studies have mainly focused on improving the forecasting performance by introducing different individual models, which ignored the potential forecasting power of a hybrid modeling. Therefore, a novel hybrid forecasting system is proposed, which can bridge the research gap in current studies.

(2) ***An advanced reduction module is proposed to simultaneously obtain the key attributes and samples.*** Reduction is an effective method to improve forecasting performance. However, it is ignored by majority of the previous studies. Therefore, in this study, we combine rough sets with a SOM, reduce the samples after attribute reduction, and horizontally and vertically process the data to improve the forecasting accuracy of the system.

(3) ***A Forecasting module based on an improved machine learning technique is proposed to improve the generalization ability and robustness of the system.*** The advanced machine learning technique, named as ELM architectures, has, thus, long been ignored in modeling of direct economic losses from storm surge disasters, despite its superiority in other forecasting fields. Hence, the forecasting module is developed based on an improved ELM to further improve the performance of the forecasting systems.

Organization

The remainder of this article is organized as follows. Section *Detailed Process of the Developed Hybrid Forecasting System* introduces the reduction module and forecasting module, respectively. In section *The Direct Economic Loss Forecasting System*, the direct economic loss forecasting system is presented. The experimental setup and a summary of the results are shown in *sections Experiments and Summary*. Section *Conclusions and Future* work generalizes the conclusions and suggests some possible directions for future research.

DETAILED PROCESS OF THE DEVELOPED HYBRID FORECASTING SYSTEM

The hybrid forecasting system based on intelligent algorithm is successfully developed, which includes three modules: reduction module, forecasting module, and evaluation module.

Module 1: Reduction Module

In the process of forecasting the direct economic losses caused by regional storm surge disasters, many indicators affect the data. Varying time intervals of storm surge have different contributions to each other, and various factors have different effects on the loss results. Therefore, the processing of sample data and the selection of storm surge samples affect the accuracy of the forecasting results. The reduction module is applied to simultaneously reduce the indices and samples, so as to identify the key indices affecting the loss values and to select the samples with a strong forecasting correlation. Accuracy is also improved by an effective dataset.

Construction of Attribute set

To maintain the integrity and validity of the index system, the availability and accuracy of storm surge disaster loss data should be considered. By combining frequency statistics with theoretical analysis of disaster economic loss assessment (Yin and Sun, 2011; Yin et al., 2011), an economic assessment of the index system of storm surge disasters was constructed, and the initial attribute set of the module was formed. Attribute classification and spatiotemporal clustering were carried out for factors under the four dimensions of the risk of disaster-caused factors, economic loss index, vulnerability to hazards, and the adaptability of storm surge disasters. The initial attribute set is composed of 19 factors that affect the economic losses caused by storm surge disasters, as shown in **Figure 1.**

Stage I: Attribute Reduction

The choice of rough set theory as a tool for feature selection is based on two considerations. First, rough set theory does not require *a priori* knowledge and the mathematical technology is mature. Second, a rough set theory can directly analyze and infer data so as to discover the hidden knowledge between the data

Classification	Attribute	Variable
Risk of disaster- caused factors	Maximum storm surge (cm)	C_1
	Maximum wind speed at landing(m/s)	C_2
	Central air pressure at landing(hPa)	C_3
Economic loss index	Damaged area of crops(1000 HA)	C_4
	Affected area of aquaculture (1000 HA)	C_5
	The length of marine engineering damage(km)	C_6
	The number of damaged vessels	C_7
	The disaster-affected population(10,000)	C_8
Vulnerability for hazard bearing body	GDP per capita (CNY)	C_9
	Population density (people / km^2)	C_{10}
	Proportion of primary industry in GDP (%)	C_{11}
	Per capita disposable income of urban households(CNY)	C_{12}
	Per capita disposable income of rural households(CNY)	C_{13}
	Mariculture area (1000 HA)	C_{14}
	The sown area of crops(10,000 mu)	C_{15}
Adaptability of storm surge disasters	The number of doctors per 10,000 people	C_{16}
	The number of beds per 10,000 people	C_{17}
	The number of medical institutions	C_{18}
	The local fiscal revenue(CNY100M)	C_{19}

FIGURE 1 | Attribute set for economic loss assessment of storm surge disasters.

and to reveal potential patterns. Thus, it is an effective method of knowledge discovery.

Definition 1. In Pawlak's rough set (Pawlak, 1982) attribute reduction theory, an equivalence relation is central and primitive. The theory begins with the notion of an information system, which is considered as $IS = (U, AT, g, V)$, where is a non-empty set of finite objects, $U = \{x_1, \ldots, x_i, \ldots, x_n\}$, it is called the universe. AT is a non-nullable collection of attributes. The information function is$g : U \times AT \rightarrow V$, which expresses the value of x_i under a, that is $g(x_i, a) \in V_a (a \in AT, x_i \in U)$, V_a is the domain that attribute what a may take.

Definition 2. Given any$A \subseteq AT$, there is an indiscernible relation $IND(A)$ on U.

$$IND(A) = \{(x_i, x_j) \in U \times U | \forall a \in AT, g(x_i, a) = g(x_j, a)\} \quad (1)$$

If$(x_i, x_j) \in IND(A)$, then and cannot be discernible by attributes from A. Generated by $IND(A)$, the partition of is denoted as

$$U/IND(A) = \{[x_i]_A : x_i \in U\} \quad (2)$$

With regard to attribute set AT, $[X_i]_A$ are equivalent classes of indiscernible relation, which can describe arbitrary subsets of U. The equivalence classes of $IND(A)$ and the empty set $\varnothing$ are the base sets in the information system IS.

Definition 3. Let $X \subseteq U$, it may not be represented exactly and clearly in U. One can describe an arbitrary subset X by a pair of lower and upper approximations. They are defined, respectively, as

$$\underline{A}(X) = \{X_i [X_i]_A \subseteq X\} \text{ and}$$
$$\overline{A}(X) = \{X_i [X_i]_A \cap X \neq \varnothing\} \quad (3)$$

The pair $[\underline{A}(X), \overline{A}(X)]$ is termed as the rough set of X in regard to the set of attributesAT.

Definition 4. Let A and B be the two equivalence relations over U. Then the regions of B: positive region, negative region and boundary region can be defined as, respectively,

$$POS_A(B) = \bigcup_{X \in U/B} \underline{A}X \quad (4)$$

$$NEG_A(B) = U - \bigcup_{X \in U/B} \overline{A}X \quad (5)$$

$$BND_A(B) = \bigcup_{X \in U/B} \overline{A}X - \bigcup_{X \in U/B} \underline{A}X \quad (6)$$

TABLE 1 | Example dataset.

$x \in U$	a	b	c	d	$\Rightarrow$	e
1	2	1	3	3		1
2	1	2	2	2		3
3	3	1	1	2		2
4	2	2	1	3		3
5	2	1	3	1		2
6	3	3	1	2		2
7	3	2	2	2		3
8	1	2	2	1		2

Definition 5. An information system is called a decision system if the collection of attributes AT can be divided into condition attribute set A and decision attribute set B. In order to clearly illustrate these mathematical definitions, an example will be employed to explain the principle. In this case, four conditional attributes (a, b, c, d), one decision attribute e, and eight objects are represented in **Table 1**.

In the system, one of the significant matters in data analysis is to find the dependency relationship between attributes. The dependency between A and B can be defined in the following way:

$$\gamma_A(B) = \frac{|POS_A(B)|}{|U|} \tag{7}$$

If $\gamma_A(B) = 1$, B depends totally on A, if $0 < \gamma_A(B) < 1$, B depends partially on A, and if $\gamma_A(B) = 0$, then does not depend on A.

For example, if $A = \{b, c\}$ then objects 2, 7, and 8 are indiscernible, 1 and 5 are indiscernible. The partition of U can be shown as

$$\begin{aligned} U/IND(A) &= U/IND(b) \otimes U/IND(c) \\ &= \{\{1,3,5\}\{2,4,7,8\}\{6\}\} \otimes \{\{3,4,6\}\{2,7,8\}\{1,5\}\} \\ &= \{\{3\}\{1,5\}\{4\},\{2,7,8\},\{6\}\}. \end{aligned}$$

If$A = \{b, c\}$and $B = \{e\}$, then

$$\begin{aligned} POS_A(B) &= \bigcup \{\varnothing, \{3,6\}, \{4\}\} = \{3,4,6\} \\ NEG_A(B) &= U - \bigcup \{\{1,5\}, \{3,1,5,2,7,8,6\}, \{4,2,7,8\}\} = \varnothing \\ BND_A(B) &= U - \{3,4,6\} = \{1,2,5,7,8\}. \end{aligned}$$

It follows that in attribute set A, objects 3, 4, and 6 can definitely be classified as one class for attribute e. The rest of the objects, however, cannot be classified.

Hence, The dependency of $\{e\}$ from $\{b, c\}$ is

$$\gamma_{\{b,c\}}(\{e\}) = \frac{|POS_{\{b,c\}}(\{e\})|}{|U|} = \frac{|\{3,4,6\}|}{|\{1,2,3,4,5,6,7,8\}|} = \frac{3}{8}.$$

Definition 6. Given A, and an attribute $a \in A$, the significance of the attribute a is defined as

$$\sigma_A(B, a) = \gamma_A(B) - \gamma_{A-[a]}(B) \tag{8}$$

The more the dependency changes, the more important the attribute becomes. If the significance is 0, that means $POS_A(B) = POS_{A-[a]}(B)$, then the attribute is dispensable. Otherwise, a is said to be relatively indispensable in A. In E, if every attribute is relatively indispensable and $POS_E(B) = POS_A(B)$, So is called a relative reduction in condition attribute set A.

For example, if $A = \{a, b, c\}$ and $B = \{e\}$, then

$$\begin{aligned} \gamma_{\{a,b,c\}}(\{e\}) &= |\{3,4,6,7\}|/8 = 4/8 \\ \gamma_{\{a,b\}}(\{e\}) &= |\{3,4,6,7\}|/8 = 4/8 \\ \gamma_{\{b,c\}}(\{e\}) &= |\{3,4,6\}|/8 = 3/8 \\ \gamma_{\{a,c\}}(\{e\}) &= |\{3,4,6,7\}|/8 = 4/8 \end{aligned}$$

so the significance of a, b, c is calculated as follows:

$$\begin{aligned} \sigma_A(B,a) &= \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{b,c\}}(\{e\}) = 1/8 \\ \sigma_A(B,b) &= \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,c\}}(\{e\}) = 0 \\ \sigma_A(B,c) &= \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,b\}}(\{e\}) = 0 \end{aligned}$$

Hence, attribute a is relatively indispensable, while, attribute b and c can be dispensable. By experimenting, the final relative reduction can be built.

Stage II: Sample Reduction

On the basis of the data, the features of storm surge disaster direct economic losses. The SOM is put forward to classify the samples automatically. The samples with high correlation degree are gathered and retained, while the samples with low correlation degree are separated and removed to achieve a sample reduction.

Definition 1. Self-Organizing Map (Kohonen, 1990) is an unsupervised learning method, which can reduce the dimension of an n-dimensional input space X to a two-dimensional output plane, thus, forming a topology of M neurons. The weight vector $W_i (i \in 1, \ldots, M)$ represents each neuron, which is an n-dimensional vector related to the input samples. The SOM system schematic is described in **Figure 2**.

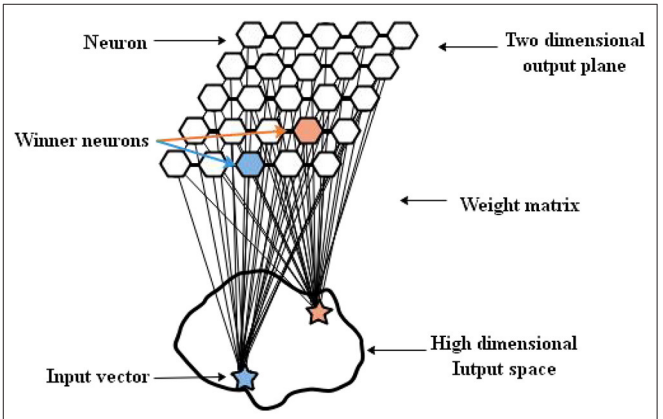


FIGURE 2 | The Self-Organizing Map (SOM) system schematic.

Definition 2. In the initial stage, the weight of neurons is varied to initialize the network with different topologies. When the input vectors are sent through the SOM neural network, each neuron varies its position by calculating the distance between the weight and the input vector. At time step t, a new input sample $X(t)$ is presented to SOM, and a winner neuron is stated:

$$WX(t) = \arg j \in \{\overset{\min}{1}, \ldots, M\} \left\| X(t) - W_j(t) \right\| \tag{9}$$

Then, the weight vectors are updated,

$$W_i(t+1) = W_i(t) + \eta(t)\Psi(i, WX(t))(X(t) - W_i(t)) \tag{10}$$

where $\eta(t)$ is a decaying learning rate and $\Psi(i, WX(t))$ is called the neighborhood function,

$$\Psi(i, WX(t)) = \exp(-\frac{\left\| WX(t) - W_i \right\|^2}{2\sigma^2(t)}) \tag{11}$$

$\sigma^2(t)$ is the neighborhood radius, $\left\| WX(t) - W_i \right\|$ is the Euclidean distance between the neurons. After repeated iterative training, the similar weight vectors are close, can even be clustered, while the dissimilar weight vectors are separated.

Module 2: Forecasting Module

The prediction module mainly adopts the Marine Predators Algorithm (MPA) to optimize the ELM. The MPA (Faramarzi et al., 2020) was developed as a novel meta-heuristic algorithm, which is primarily modeled on ocean predators to select an optimal foraging strategy according to the prey location. In the forecasting module, the position of the predator is represented by the multi-dimensional vector composed of input weights and hidden layer thresholds, which can determine the performance of the ELM. Then, the optimal parameters are obtained based on the optimization criteria. These two theoretical principles will briefly introduce the following.

Extreme Learning Machine

Definition 1. Extreme learning machine (Huang et al., 2004) was proposed as an easy-to-use and effective single hidden-layer feedforward neural network method. Instead of iteratively adjusting the weights and biases of the network, the hidden layer parameters are assigned randomly, and the least square method is proposed to generate the unique optimal solution. Therefore, it has the advantages of fast training speed, avoiding over fitting, and local optimization to a certain degree.

Definition 2. Given the original training dataset ($\mathbf{x_j}$,$\mathbf{t_j}$), where $\mathbf{x_j}$ is the input variables, $\mathbf{x_j} = [x_{j1}, x_{j2}, \ldots, x_{jn}]^T \in R^n$, and $\mathbf{t_j}$ is the output variables, $\mathbf{t_i} = [t_{i1}, t_{i2}, \ldots, t_{im}]^T \in R^m$. D is the number of hidden neurons. The output matrix of ELM is shown as follow:

$$T = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{mj} \end{bmatrix}_{m\times N} = \begin{bmatrix} \sum_{i=1}^{D} \beta_{i1} g(a_i x_j + b_i) \\ \sum_{i=1}^{D} \beta_{i2} g(a_i x_j + b_i) \\ \vdots \\ \sum_{i=1}^{D} \beta_{im} g(a_i x_j + b_i) \end{bmatrix}_{m\times N} (j \mp 1, 2, \ldots, N) \tag{12}$$

where $\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{im})^T$ is the output weights matrix between the ith hidden neuron and the output layer nodes, $g(\cdot)$ is the activation function of the hidden layer, $a_i = (a_{i1}, a_{i2}, \ldots, a_{iN})^T$ is the input weight matrix that connects the ith hidden layer node and input layer nodes, and b_i is the bias of the ith hidden layer node.

The above-mentioned matrix can be indicated as below:

$$H\beta = T \tag{13}$$

where H is the hidden layer output matrix:

$$H = \begin{bmatrix} g(a_1 x_1 + b_1) & \cdots & g(a_D x_1 + b_D) \\ \cdots & \cdots & \cdots \\ g(a_1 x_N + b_1) & \cdots & g(a_D x_N + b_D) \end{bmatrix}_{N\times D} \tag{14}$$

Definition 3. In the training dataset, the inputs samples and its corresponding targets are already given. The input weight matrix **a** and bias **b** can be given randomly, then the output weight matrix β can be calculated by Moore-Penrose generalized inverse to get its least square solution:

$$\boldsymbol{\beta} = H^+ T \tag{15}$$

where is called the Moore-Penrose generalized inverse of matrix. Then the orthogonal projection approach is taken to calculate H^+that is $H^+ = H^T(HH^T)^{-1}$.

The network architecture of ELM is displayed as (**Figure 3**).

As mentioned above, the input weight matrix **a** and bias **b** are the two crucial parameters that are presented in ELM.

Marine Predators Algorithm

According to the theory of "survival of the fittest," the predator determines the optimal strategy ensuring a reasonable contact rate with the prey. The MPA starts from the initial stage of the population, then goes through the three optimization stages considering different speed ratio and simultaneously simulating the whole life cycle of the predator and prey.

Definition 1. At initialization stage, the initial populations for both the prey and the predator can be randomly located in the search space *via* the following mathematical expression:

$$U = lb + rand^*(ub - lb) \tag{16}$$

where *lb* and *ub* are the lower and upper boundaries for variables, and *rand**is a random number in the range of 0–1. According to the formula 14, the initial location matrix of prey can be established, as below:

$$prey = \begin{bmatrix} U_{11} U_{12} \ldots U_{1d} \\ U_{21} U_{22} \ldots U_{2d} \\ \ldots \ldots \ldots \\ U_{n1} U_{n1} \ldots U_{nd} \end{bmatrix}_{n\times d} \tag{17}$$

In Equation (17), n is the number of search agents, while d is the number of dimensions.

Definition 2. Inspired by the concept of survival of the fittest, the first-class predators have the best foraging techniques. Therefore,

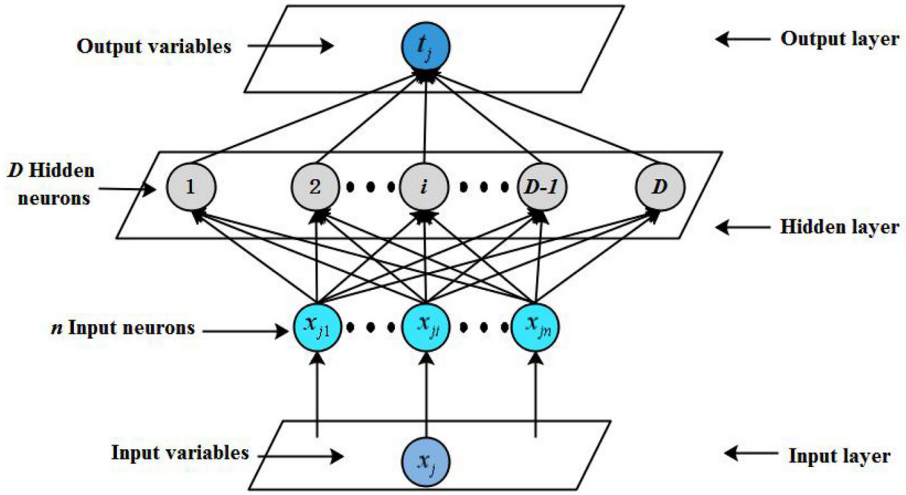


FIGURE 3 | Network structure of extreme learning machine (ELM).

when establishing the Elite matrix, the fittest population is selected as the first-class predator. On the basis of the location information of the prey, the digit group of the elite matrix is updated by searching and finding the quarry. At the end of each iteration, if the first-class predator was replaced by a better one, the Elite matrix would be altered.

$$Elite = \begin{bmatrix} U_{11}^I U_{12}^I \ldots U_{1d}^I \\ U_{21}^I U_{22}^I \ldots U_{2d}^I \\ \ldots \ldots \ldots \\ U_{n1}^I U_{n2}^I \ldots U_{nd}^I \end{bmatrix}_{n \times d} \quad (18)$$

Definition 3. In the initial iteration stage of optimization, $Iter < \frac{1}{3}Max_Iter$, ($Iter$ is the current iteration, Max_Iter is the maximum one), the prey searches for food in the exploration field, while the predator chooses the optimal strategy of immobility. Therefore, the location of the prey is determined by the following equations.

$$S_i = R_B \otimes (Elite_i - R_B \otimes U_i), i = 1, 2, \ldots, n \quad (19)$$

$$U_i = U_i + P.R \otimes S_i \quad (20)$$

where R_B indicates a vector including random numbers (based on the normal distribution of Brownian motion), $\otimes$ represents entry-wise multiplications. The $P = 0.5$. is a constant number and $R \in [0, 1]$ is a random number coming from a uniform distribution. The multiplication of R_B and prey imitates its movement.

Definition 4. In the intermediate iteration stage of optimization, $\frac{1}{3}Max_Iter < Iter < \frac{2}{3}Max_Iter$, not only do the prey and predator change their positions to seek food, but they also move at the equal velocity. The population is divided into two parts. The first part (i.e., prey) of the agents is allocated for exploitation, and the second half (i.e., predator) is in charge of exploration. Equation (21) imitates the movement of the first half of the population.

$$S_i = R_L \otimes (Elite_i - R_L \otimes U_i), i = 1, 2, \ldots, n/2 \quad (21)$$

$$U_i = U_i + P.R \otimes S_i \quad (22)$$

where R_L represents a random number vector based on the *Lévy* distribution. The multiplication of R_L by prey emulates the motion of prey, while the second half of the agents perform the following equations. The second half of the population is represented by the following mathematical formula:

$$S_i = R_B \otimes (R_B \otimes Elite_i - U_i), i = 1, 2, \ldots, n/2 \quad (23)$$

$$U_i = Elite_i + P.CF \otimes S_i,$$

$$CF = (1 - \frac{Iter}{Max_Iter})^{(2\frac{Iter}{Max_Iter})} \quad (24)$$

where CF is the adaptive parameter that controls the step size for the predator motion.

Definition 5. In the final iteration stage of optimization, $Iter > \frac{2}{3}Max_Iter$, the speed of prey is slower than that of predator. The predator adopts exploitation strategy based on *Lévy* migration and its location is updated as follows:

$$S_i = R_L \otimes (R_L \otimes Elite_i - U_i), i = 1, 2, \ldots, n \quad (25)$$

$$U_i = Elite_i + P.CF \otimes S_i,$$

$$CF = (1 - \frac{Iter}{Max_Iter})^{(2\frac{Iter}{Max_Iter})} \quad (26)$$

Definition 6. Faramarzi et al. believed that the external environmental factors make the behavior of the population change more or less, such as the eddy formation or Fish Aggregating Devices (FADs) effects. In order to avoid the local optimization, longer jumps should be considered in

the simulation process. Hence, Equation (27) shows the mathematical model of the FADs effects,

$$U_i = \begin{cases} U_i + CF[U_{\min} + R \otimes (U_{\max} - U_{\min}) \otimes M & r_5 < FADs \\ U_i + [FADs(1-r) + r](U_{r1} - U_{r2}) & r_5 > FADs \end{cases} \tag{27}$$

In Equation (27), where *FADs* represent the probability of affecting the search process and is set equivalent to 0.2, and M is the binary solution (0 or 1) corresponding to the stochastic solution. If the array is <0.2, the array is altered to zero. If the array is >0.2, the array is converted to one. The notation $r \in [0, 1]$ defines a random number. The r_1 and r_2 are the random indices of the prey.

Definition 7. An important feature of marine predators is that they have a specific memory of their successful foraging position. In MPA, this feature is simulated by saving the optimum solution of the previous iteration, and the performance of each solution of the current iteration is compared with the previous one. If it is better, the previous solution will be replaced, which is helpful for fast optimization.

Proposed MPA-ELM Forecasting Module

The MPA has a good ability of optimization. The input weight matrix and bias **b** of ELM are optimized. The forecasting module is constructed. The loss of training set is considered as the fitness function of MPA, and the calculation is shown in Equation (28),

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2} \tag{28}$$

where Y is the actual value and $\hat{Y}$ is the predicted value.

The pseudo code of forecasting module is described as follows:

The flowchart of forecasting module is shown in **Figure 4**, the steps can be summarized as the following:

Step 1: Initialize positions for the prey and the predator, construct the matrices of prey and Elite according to Equations (17) and (18), and accomplish a memory saving. The position coordinates of each predator are composed of parameters **a** and **b**.

Step 2: Select the best predator. On the basis of the location information of the prey and the previous memory of predator to capture its food successfully, the best predator matrix is assigned.

Step 3: Determine the update criteria according to the range of *Iter*. If $Iter < \frac{1}{3}MaxIter$, the process is in the stage 1, execute Step 4, otherwise, execute Step 5.

Step 4: Update the solutions based on Equations (19) and (20), and then skip to Step 8.

Step 5: If $\frac{1}{3}MaxIter < Iter < \frac{2}{3}MaxIter$, the process is in the stage 2, execute Step 6, otherwise, the process is in the stage 3,execute Step 7.

Step 6: Update the positions of the prey and predator based on Equations (21–24). The first half part performs Equations (21) and (22), and the other half follows Equations (23) and (24), and then skip to Step 8.

Algorithm 1: *MPA-ELM*

Fitness function:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \hat{Y}_i\right)^2}$$

output:

$\hat{Y}$—the forecasting direct economic losses from ELM

Parameters:

Max_Iter—the maximum number of iterations
N—the number of search agents
F_i—the fitness function of Elite i
X_i—the position of *i_th* search agent
lb/ub—the lower/upper bound of variables
Iter—the current iteration number
r—the random number from 0 to 1
d—the number of dimension

/∗Set the parameters of MPA.∗/
/∗Initialize search agents of X_i ($i = 1, 2,...,$ N) randomly.**∗/**
FOR EACH *i*: $1 \le i \le N$ DO
  Evaluate the corresponding fitness function F_i
END FOR
/∗Determine the current Elite matrix.∗/
WHILE (*Iter*< *Max_Iter*) DO
  FOR EACH *i*=1: *N* DO
    /∗Calculate the objective value of all agents.∗/
    /∗Update the obtained solution.∗/
    IF (*Iter*<*Max_Iter*/3) THEN
      /∗Update all population positions in the exploration field.∗/
      $S_i = R_B \otimes (Elite_i - R_B \otimes U_i), i = 1, 2, \ldots, n$
      $U_i = U_i + P.R \otimes S_i$
    ELSE
      IF (Max_Iter/3<Iter<2∗Max_Iter/3) THEN
        FOR EACH i=1: N /2 DO
          /∗Update the position of the first half population.∗/
          $S_i = R_L \otimes (Elite_i - R_L \otimes U_i),$
          $i = 1, 2, \ldots, {}^{n}/_{2}$
          $U_i = U_i + P.R \otimes S_i$
          /∗Update the position of the second half population.∗/
          $S_i = R_B \otimes (R_B \otimes Elite_i - U_i),$
          $i = 1, 2, \ldots, {}^{n}/_{2}$
          $U_i = Elite_i + P.CF \otimes S_i,$
          $CF = (1 - \frac{Iter}{Max_Iter})^{2\frac{Iter}{Max_Iter}}$
        END FOR

        ELSE If (2∗ Max_Iter/3 <Iter< Max_Iter) THEN
          /∗Update all population positions in the final stage.∗/
          $S_i = R_L \otimes (R_L \otimes Elite_i - U_i),$
          $i = 1, 2, \ldots, n$
          $U_i = Elite_i + P.CF \otimes S_i,$
          $CF = (1 - \frac{Iter}{MaxIter})^{(2\frac{Iter}{MaxIter})}$
        END IF
      END IF
  END FOR
  /∗ Using FADs effect and update current agent based on Equation. (27).**∗/**
  ***Iter=Iter*+1**
END WHILE
RETURN X_i^*
Set the weight and bias of the ELM according to X_i^*
Input the testing data into ELM to obtain the forecasting value $\hat{Y}$

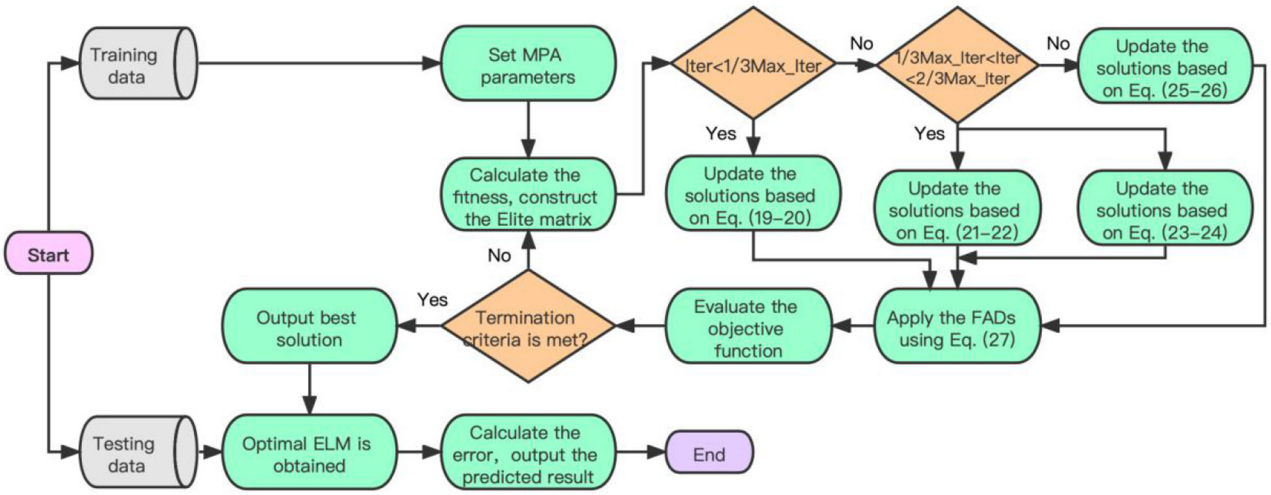


FIGURE 4 | The flowchart of marine predators algorithm (MPA)-ELM forecasting module.

Step 7: Update the positions of the prey and predator based on Equations (25) and (26), and then skip to Step 8.
Step 8: Apply the model of the FADs effects using Equation (27), and then skip to Step 9.
Step 9: Evaluate the objective function, and then skip to Step 10.
Step 10: Determine whether the termination condition is satisfied. If it is met, the program ends, and the best position parameter is the output. Otherwise, skip to step 2.
Step 11: Obtain the optimal ELM.
Step 12: Calculate the error, output the predicted result.

Module 3: Evaluation Module

In order to evaluate the effectiveness of the proposed system, an evaluation module is provided. There is no unified standard to confirm the validity of various models. Therefore, by consulting the relevant literature in the field of prediction (Wang et al., 2018; Gu et al., 2021), a variety of error metrics are adopted in this paper, including mean square error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), PMSE, PMAE, and PMAPE, as shown in **Table 2.** Where m represents the number of testing data, A and F represent the collected and predicted economic losses data, respectively. The $\overline{A}$ is the average value of the actual data, and $\overline{F}$ is the average value of forecasting results. Specifically, MSE, MAE, and MAPE can be considered to evaluate the forecasting accuracy, and the smaller value of these indicators shows better forecasting performance. Performance improvement percentage indicators namely PMSE, PMAE, and PMAPE further evaluate the improvement between different models to quantitatively describe the degree of performance improvement. The evaluation module evaluates the forecasting accuracy, the degree of performance improvement, and the forecasting ability of the system. In summary, six indicators are selected, which can focus on a comprehensive scientific evaluation.

TABLE 2 | Performance metric rules.

Metric	Definition	Equation
MSE	The average of m error squares	$MSE = \frac{1}{m}\sum_{i=1}^{m}(A_i - F_i)^2$
MAE	The mean absolute error of forecasting results	$MAE = \frac{1}{m}\sum_{i=1}^{m}\lvert A_i - F_i\rvert$
MAPE	The average of m absolute percentage errors	$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left\lvert\frac{(A_i-F_i)}{A_i}\right\rvert \times 100\%$
PMSE	The improvement rate of MSE between model 1 and model 2	$P_{MSE} = \left(\frac{MSE1-MSE2}{MSE1}\right) \times 100\%$
PMAE	The improvement rate of MAE between model 1 and model 2	$P_{MAE} = \left(\frac{MAE1-MSE2}{MAE1}\right) \times 100\%$
PMAPE	The improvement rate of MAPE between model 1 and model 2	$P_{MAPE} = \left(\frac{MAPE1-MAPE2}{MAPE1}\right) \times 100\%$

THE DIRECT ECONOMIC LOSS FORECASTING SYSTEM

A new hybrid forecasting system RS-SOM&MPA-ELM is proposed for the prediction of the direct economic losses caused by storm surge disasters. **Figure 5** illustrates the hybrid forecasting framework of our proposed new hybrid approach. The proposed hybrid approach is generalized as given below:

Step 1: Data collection and processing. Before being input to the model, the data is processed into a pattern that meets its requirements, including data normalization, discretization, and processing of the training format. At the same time, the data set is divided into training data set and testing data set.
Step 2: Attribution and sample reduction. Rough sets are proposed to reduce initial attributes, and the key features affecting the direct economic losses are selected. The decision

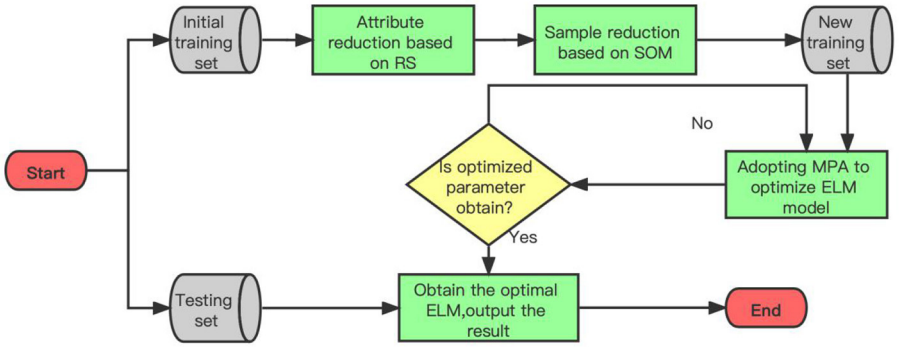


FIGURE 5 | The forecasting framework of forecasting system.

table, composed of normalized and discretized data sets of direct economic loss factors of storm surge disasters, is reduced by positive region reduction rules and by dependency reduction rules. Then, according to the features extracted from RS, disaster-related parameters of several storm surge are input into the SOM as the initial training set. Subsequently, sample sets that contribute to improving prediction performance are selected as the new training set.

Step 3: The MPA optimization. First, the parameters of ELM are initialized. Second, ELM is embedded into the MPA for calculation, including the input weight and the bias update. Finally, after the embedded model training is completed, the algorithm judgment conditions are checked to determine whether the maximum iteration times has been satisfied. If the requirement is met, the optimized parameters are output, otherwise, repeat the above process.

Step 4: Final system performance test. After the completion of the third step, the optimized model is obtained, the test data set is input into the forecasting module. Then, the result is output, and the forecasting performance is tested.

EXPERIMENTS

To verify the superior performance of the system and to ensure the diversity of data, experiments were carried out on the data of three provinces. The experimental environment is macOS 10.14.6, on a system with an Intel(R) Core CPU (Core-i5 2.6 GHz), and 8 G RAM. Different tools were used to implement the methods of this paper. Rough sets were implemented in Python 3.8.3, and all the other models were implemented in the MATLAB R2016b software package.

Experimental Data

This study involves 19 variables, including 18 characteristic variables of attribute sets and direct economic losses of storm surge disasters. This paper selected 60 relatively complete records of storm surge disasters in Guangdong, Zhejiang and Fujian provinces from 1989 to 2019 as the research objects (Because there are no complete storm surge samples collected in 2020 and 2021, the selected samples extend to 2019). To ensure the validity and dependability of the data, the relevant data obtained are from the public sources in **Table 3**.

TABLE 3 | Data sources.

Variables	Source
Maximum storm surge (cm)	Fujian Marine Disaster Bulletin, 2012–2021; Zhejiang Marine Disaster Bulletin, 2012–2021; Guangdong Marine Disaster Bulletin, 2014–2021; Nanhai Marine Disaster Bulletin, 2016–2021; China Marine Disaster Bulletin, 2020–2021, Collection of Storm Surge Disasters Historical Data in China 1949–2009, (Yu et al., 2015)
Maximum wind speed at landing (m/s)	
Central air pressure at landing (hPa)	
Damaged area of crops (1,000 HA)	
Affected area of aquaculture (1,000 HA)	
The length of marine engineering damage (km)	
The number of damaged vessels	
The disaster-affected population (10,000)	
GDP per capita (CNY)	Guangdong Statistical Yearbook, 1990–2020; Zhejiang Statistical Yearbook, 1990–2020
Population density (people/km^2)	
Proportion of primary industry in GDP (%)	
Per capita disposable income of urban households (CNY)	
Number of doctors per 10,000 people	Fujian Statistical Yearbook, 1990–2020; China social statistical yearbook, 2006–2020
Number of beds per 10,000 people	
Number of medical institutions	
Local fiscal revenue (CNY100M)	
Per capita disposable income of rural households (CNY)	
Mariculture area (1,000 HA)	China Agricultural Statistical Report, 1990–2017; China Marine Statistical Yearbook, 1993–2017
Sown area of crops (10,000 mu)	

Normalization and Discretization

Each sample in the original data table has 19 different attributes, and each attribute represents different meanings about the economic losses caused by storm surges disasters, including data

about the natural attributes of storm surges and information about the affected areas. Each group of data has different dimensions. To eliminate the dimensional influence between indicators and to speed up the training speed, the general data processing method should be the normalization before entering data into the model. The normalization calculation method is shown in Equation (29), where x_j is the input data of the j-th node in the input layer of the model, $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the input sequence, respectively, and $\bar{x}_j$ is the normalized data with the range of [0, 1].

$$\bar{x}_j = \frac{x_j - x_{\min}}{x_{\max} - x_{\min}} \tag{29}$$

To reduce the attributes based on rough set theory, it is necessary to initially discretize the data. The sample variables are divided into several intervals, and each interval is regarded as a category. This process of categorizing data variables is often called discretization. All values within each category are mapped to the same value, converting the actual value to a numeric attribute of the symbol. The equal distance partition algorithm is proposed for discretization. After discretization, the original decision table is replaced by a new decision table of numerical attributes.

Experimental Design

To verify the effectiveness and robustness of the developed prediction system, two experiments will be done, denoted as ***Experiment I*:** compared with the traditional single models, ***Experiment II*:** compared with the improved hybrid models.

To verify the superiority of the proposed RS-SOM&MPA-ELM forecasting system in predicting the economic losses caused by storm surge disasters, some forecasting models are selected as comparison models. Therefore, the single neural network algorithm models, different hybrid models, and different optimization algorithm-based models are considered. To sum up, eight comparative models are established to evaluate the developed hybrid forecasting system. These comparison models are listed in **Table 4**.

To ensure the fairness of the comparison between the models, the basic parameters of each algorithm are set according to the default values. To ensure fairness, some parameter settings in the original literature are maintained for competing models, while some parameters are shared. For example, the number of iterations or the frequency of training of all models is set to 50, and each model was performed 250 times to output statistically stable results. To judge the number of SOM clustering categories, two, three, and four values are selected to do the experiment. The results show that 60 samples are classified according to the occurrence time. To reflect the classification features to the greatest extent and to ensure the sufficient number of samples, the optimum clustering numbers of SOM is determined as 2. The BPNN and ELM adopt Sigmoid function as the activation function. The Radial Basis Function is adopted in SOM and SVM as kernel function. The regularization parameter C and kernel parameter in all SVM-based comparison models are searched in grid ranges of $[2^{-8}, 2^{8}]$ and $[2^{-5}, 2^{5}]$ with step 0.5. The specific experimental parameters are shown in **Table 5**.

Moreover, due to the uncertainty of the occurrence of storm surge disasters and the incompleteness of statistical data, the economic loss forecasting of storm surge disasters exhibited the characteristics of small sample size. In the proposed system and

TABLE 4 | Comparison models.

Experimental	Comparison models
Experimental I	BPNN
	SVM
	ELM
Experimental II	RS+SOM+BPNN
	RS+SOM+SVM
	RS+SOM+ELM
	MPA+ELM
	RS+SOM & PSO+ELM

TABLE 5 | Experimental parameter values.

Models	Experimental parameter	Default value
BPNN	Neuron number in the input layer	5
	Neuron number in the hidden layer	7
	Neuron number in the output layer	1
	Number of the hidden layer	5
	Learning velocity	0.1
	Training requirements precision	0.00004
	Maximum number of trainings	50
SVM	Penalty parameter	50
	Value of gamma in kernel function	0.2
	Setting type	3
	Value of epsilon in loss function	0.05
	Maximum number of errors	0.1
ELM	Neuron number in the input layer	19
	Neuron number in the output layer	1
	Number of the hidden layer	1
SOM	Neuron number in the input layer	5
	Neuron number in the output layer	2
	Learning rate	0.5
	Maximum number of trainings	50
MPA	Search agents number	30
	FADs	0.2
	Constant number P	0.5
	Random index of agents	0 to 1
	Maximum number of iteration times	50
PSO	Population scale	20
	Acceleration constant C1	1.49445
	Acceleration constant C2	1.49445
	Speed limit	0.1 to 0.2
	Variable value	−10 to 10
	Maximum number of iteration times	50

TABLE 6 | Forecasting performance of comparison models.

Models	MSE	MAE	MAPE (%)
Model 1	17.5536	4.1327	36.9559
Model 2	6.7561	2.4689	21.9069
Model 3	3.5866	1.8464	16.5293

Model 1: BPNN, Model 2: SVM, Model 3: ELM.

comparison models, two groups of data are proposed as the testing set, which are the economic loss samples of regional storm surge disasters in 2019 and 2018. At the same time, two storm surges are coded sequentially: the first sample and the second sample. The remaining samples constitute the initial training data set. The SOM is proposed to, respectively, select the corresponding new training set for the samples. The results are as follows. The new training set of the first sample is from sample 35 to sample 59, whereas that of the second sample is from sample 34 to sample 58. Subsequently, for the five attributes selected by RS, the test and error test should be carried out by minimizing the MAE and the MAPE of the training set to verify whether all of them are input. The results show that in the training set of test samples, regardless of which of the five factors selected by RS is removed, good performance is not obtained. Hence, five attributes are determined as the input of forecasting module.

Experiment I: Comparison With the Traditional Single Models

In *Experiment I, three* single models are adopted for forecasting through the comparison of performance forecasting standards, where BPNN and SVM are the most popular machine learning models. The ELM is the method related to the forecasting model that we use. Researchers have shown that these models have good performance in forecasting. All the performance metrics of each single model are shown in **Table 6**.

For the direct economic losses forecasting, in terms of the comparison among the results of single models, ELM has the best forecasting performance, followed by SVM and BPNN. The difference of calculation results of the three models shows that the ELM for this kind of structure is more suitable for the considered problem. Therefore, it is selected as the basic forecasting module of the proposed forecasting system. Although ELM results are better than the other two models, the forecasting results of single model are not ideal.

Experiment II: Compared With the Improved Hybrid Models

In *Experiment II,* a series of hybrid methods are selected to apply to multi factor forecasting. Taking proposed forecasting method ELM as an example, the first type of hybrid approaches only employed the forecasting module (MPA-ELM), the second type of hybrid approaches only conducted a reduction operation (RS+SOM+ELM), whereas another type of hybrid approaches only changed the optimization algorithm for the proposed model (RS+SOM&PSO+ELM).

After attribute reduction on the basis of RS, five attributes are obtained: the length of marine engineering damage C_6, the disaster-affected population C_8, proportion of primary industry in GDP C_{11}, mariculture area C_{14}, and the number of beds per 10,000 people C_{17}. Through SOM neural network training, according to the occurrence time of storm surge, 60 storm surge samples are divided into two groups: one group comprised of 32 storm surges from 1989 to 2008, and the other group comprised of 28 storm surges from 2008 to 2019. The recent storm surges were selected as the testing samples, and the storm surges samples from 2008 to 2017 were selected as the training set.

After training, the parameters of the proposed model and other models for forecasting are obtained. The MSE, MAE, and MAPE [Equation (24)] are used as indices to evaluate the forecasting performance. To ensure the reliability and stability of forecasting results, considering the inherent randomness of MPA and ELM, these models run 250 times, and the average forecasting value is taken as the final result.

The comparison between the results obtained by the above hybrid models and by the actual values is shown in **Table 7**. The MSE, MAE, and MAPE for the predicted values of each model are listed in **Table 8.** The IR between different approaches is presented in **Table 9.** The results of the tables also demonstrate the following:

(1) The experimental results show that the forecasting performance of all hybrid models is better than that of the single models.
(2) The proposed system significantly outperformed all models compared in terms of level of accuracy for the forecasting of direct economic losses caused by storm surge disasters. This verifies that the proposed forecasting system is an effective tool for forecasting direct economic losses caused by storm surge disasters, with the lowest MSE, MAE, and MAPE of 0.0133, 0.1154, and 1.0313%, respectively.
(3) **Table 9** displays the contribution of reduction module operation in attribute reduction and sample selection, in which Models 1, 2, and 3 represent BPNN, SVM, and ELM, respectively, and Models 5, 6, and 7 refer to approaches that, respectively, combined RS-SOM with the above models. For each corresponding group, Models 5–7 outperformed Models 1–3, respectively, according to the positive IR values. Through the positive PMAPE and PMAE criteria, the results clearly show that the models with the reduction module obtained higher levels of forecasting accuracy, which indicates that reduction module is very helpful to improve forecasting performance. Thus, it is particularly significant to master the features and preprocess the data.
(4) The performance of the models with the optimization algorithm is better than other models, which proves that the optimization algorithm can significantly improve the forecasting ability and stability of the model. Among them, the model with MPA algorithm is better in forecasting performance, and the optimization time is relatively short.
(5) The forecasting performance of model 9 is improved compared with models 5 and 6. The performance of the models with forecasting module is better than that with other

TABLE 7 | Predicted results of testing set.

	Comparison models						
Sample number	**MPA+ELM**	**RS+SOM+BPNN**	**RS+SOM+SVM**	**RS+SOM+ELM**	**RS+SOM+PSO+ELM**	**RS+SOM+ MPA+ELM**	**Actual value**
1	12.7953	11.8747	11.2483	11.2016	11.1858	11.1700	11.0400
2	13.2193	12.5694	11.2292	11.2173	11.2163	11.4908	11.3900

TABLE 8 | Forecasting performance of hybrid models.

Hybrid models	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**
MSE	3.2123	1.0141	0.0341	0.0279	0.0255	0.0133
MAE	1.7923	1.0071	0.1846	0.1672	0.1598	0.1154
MAPE (%)	15.9800	8.9577	1.6493	1.4900	1.4228	1.0313

Model 4: MPA+ELM, Model 5: RS+SOM+BPNN, Model 6: RS+SOM+SVM, Model 7: RS+SOM+ELM, Model 8: RS+SOM+PSO+ELM, Model 9: RS+SOM+MPA+ELM.

TABLE 9 | IR between different approaches.

Models	**PMSE (%)**	**PMAE (%)**	**PMAPE (%)**
Model5→ Model1	94.2228	75.6309	75.7611
Model6→ Model2	99.4953	92.5230	92.4713
Model7→ Model3	99.2221	90.9445	90.9857
Model9→ Model4	99.5860	93.5613	93.5463
Model9→ Model5	98.6885	88.5414	88.4870
Model9→ Model6	60.9971	37.4865	37.4704
Model9→ Model7	52.3297	30.9809	30.7852
Model9→ Model8	47.8431	27.7847	27.5162

Model 1: BPNN, Model 2: SVM, Model 3: ELM, Model 4: MPA+ELM, Model 5: RS+SOM+BPNN, Model 6: RS+SOM+SVM, Model 7: RS+SOM+ELM, Model 8: RS+SOM+PSO+ELM, Model 9:RS+SOM+MPA+ELM.

forecasting models. In general, BPNN and SVM showed strong performance in the forecasting field in the past, but poor performance in this research framework. This is primarily due to the characteristics of the disaster loss data, small sample problems, and the instability of forecasting results. The stability and forecasting performance of the system RS-SOM &MPA-ELM proposed in this paper is superior to that of the existing methods.

SUMMARY

According to the performance comparisons of both single models and hybrid models above, it is obvious that the forecasting results of RS-SOM&MPA-ELM are better than other models. Moreover, the forecasting performance of hybrid models is better than that of single models. In addition, some interesting phenomena are found during the study, as noted briefly below:

(1) The results of hybrid models are better than those of single models, and these single models cannot directly obtain satisfactory results. The main factors that affected the forecasting results are the redundancy and the non-linearity of the original influence factors. It is necessary to preprocess the direct economic loss factor table of storm surge disasters, so as to further improve the forecasting accuracy.

(2) In the contrast experiment, with or without RS-SOM, the forecasting results of the models with RS-SOM are better than those without RS-SOM. The RS can remove the redundant factors and screen out the key factors to improve the accuracy of forecasting. When the samples are clustered by SOM, there is a significant correlation between the sample classification and the occurrence time of storm surges. Other samples within 10 years (including 10 years) from the occurrence time of forecasting samples are more relevant to their data. This paper attempts to explain the reasons for this result: if the time interval between the two storm surges is longer, the gap between the industrial structure, economic level, forecasting technology level, and social management level is larger. The items causing economic losses are also different. On the contrary, the closer the occurrence time is, the higher the similarity of samples will be.

(3) In this work, risk of disaster-caused factors are not selected as critical factors, and different factors are chosen in other papers. Different methods and different samples lead to different choices, but the forecasting performance of the proposed approach is good, which verifies the rationality of the factor selection. However, it can be considered that different feature selection methods can be proposed for cross-validation in future work.

(4) Compared with the comparison models, the new forecasting system can obtain better accuracy in the forecasting of direct economic losses caused by storm surge disasters. The data presents the characteristics of small samples. Therefore, the forecasting system can be applied to other areas for small sample forecasting, such as economic loss forecasting for ice disasters, red tides, tsunamis, and other disasters, short-term time series forecasting, and so on.

(5) Although the proposed hybrid model has been verified to have a good forecasting ability in the small sample forecasting of economic losses caused by storm surge disasters, it still has some limitations and needs to be improved. First of all, in terms of data, the collection of economic loss evaluation index of storm surge disasters is limited by the practical difficulty, and the initial 18 index selection is subjective. Secondly, although the proposed optimization algorithm MPA improves the forecasting accuracy, it increases the forecasting time and the model complexity.

CONCLUSIONS AND FUTURE WORK

Storm surge disaster is the most serious source of marine disaster losses, which causes massive losses to coastal areas every year. Reasonable disaster loss assessment and forecasting help to carry out disaster management and reduce losses effectively. Therefore, direct economic loss forecasting of storm surge disasters has become an important topic. The proposed RS-SOM&MPA-ELM system is composed of three modules: one of which is the reduction module of RS-SOM, the second is the forecasting module of MPA-ELM, and the last is the evaluation module. The reduction module reduces the attributes and samples of the initial data simultaneously, to obtain the key input set of the forecasting module. In terms of model performance, the training set processed by reduction module performs better on single ELM model, the PMAPE can reach 90.9857%. The random allocation parameters of ELM are selected and optimized by MPA. Experimental results show that the performance of the system optimized by MPA is better than that of RS+SOM+ELM model. The improvement rate of MAPE between the models is 30.7852 %. Based on the data sets of storm surge disasters in Fujian, Zhejiang, and Guangdong, the proposed forecasting system is effective. The module is friendly to small sample forecasting, and the performance of the proposed system is better than other comparison models.

In the article, all the data are numerical data, while multi-source data can be added to the later system. At present, we focus on the samples under normal circumstances. In the future research, we would focus on the special sample data and pay more attention to huge disasters with extreme risks. It would be expected to establish a forecasting system with good performance for sparse samples.

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/s.

AUTHOR CONTRIBUTIONS

HG: conceptualization, methodology, software, data curation, formal analysis, investigation, and writing–original draft. CH: conceptualization, visualization, supervision, writing–reviewing, and editing. KY: supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Major Program of National Social Science Foundation of China (Grant No. 14ZDB151).

ACKNOWLEDGMENTS

The authors appreciate the valuable suggestions and objective comments made by the reviewers.

REFERENCES

Arns, A., Wahl, T., Dangendorf, S., and Jensen, J. (2015). The impact of sea level rise on storm surge water levels in the northern part of the German Bight. *Coast. Eng.* 96, 118–131. doi: 10.1016/j.coastaleng.2014.12.002

Bravo, J. M., and Ayuso, M. (2021). Forecasting the retirement age: a bayesian model ensemble approach. *Adv. Intell. Syst. Comput.* 136, 123–135. doi: 10.1007/978-3-030-72657-7_12

Chen, S., Tang, D., Liu, X., and Chunhua, H. (2018). Assessment of tropical cyclone disaster loss in Guangdong Province based on combined model. *Geom. Nat. Hazards Risk* 9, 431–441. doi: 10.1080/19475705.2018.1447024

China Agricultural Statistical Report. (1990–2017). Available online at: https://data.cnki.net/trade/Yearbook/Single/N2019120059?z=Z009 (accessed August 2, 2021).

China Marine Disaster Bulletin. (2020–2021). Available online at: https://www.mnr.gov.cn/sj/sjfw/hy/gbgg/zghyzhgb/ (accessed June 2, 2021).

China Marine Statistical Yearbook. (1993–2017). Available online at: https://data.cnki.net/area/Yearbook/Single/N2019060088? z=D20 (accessed August 2, 2021).

China social statistical yearbook. (2006–2020). Available online at: https://data.cnki.net/yearbook/Single/N2021030171 (accessed August 2, 2021).

Ding, Y., Ding, T., Rusdin, A., Zhang, Y., and Jia, Y. (2020). Simulation and prediction of storm surges and waves using a fully integrated process model and a parametric cyclonic wind model. *J. Geophys. Res. Oceans* 125, e2019JC015793. doi: 10.1029/2019JC015793

Fang, J., Liu, W., Yang, S., Brown, S., Nicholls, R. J., Hinkel, J., et al. (2017). Spatial-temporal changes of coastal and marine disasters risks and impacts in Mainland China. *Ocean Coast. Manag.* 139, 125–140. doi: 10.1016/j.ocecoaman.2017.02.003

Faramarzi, A., Heidarinejad, M., Mirjalili, S., and Gandomi, A. H. (2020). Marine predators algorithm: a nature-inspired metaheuristic. *Expert Syst. Appl.* 152, 113377. doi: 10.1016/j.eswa.2020.113377

Feng, Q., and Liu, Q. (2017). Pre-assessment for the loss caused by storm surge based on the SVM-BP neural network. *Mar. Environ. Res.* 36:7. doi: 10.13634/j.cnki.mes.2017.04.021

Fujian Marine Disaster Bulletin. (2012–2021). Available online at: http://hyyyj.fujian.gov.cn/search/?key=%E6%B5%B7%E6%B4%8B%E7%81%BE%E5%AE%B3%E5%85%AC%E6%8A%A5. (accessed August 2, 2021).

Fujian Statistical Yearbook. (1990–2020). Available online at: http://tjj.fujian.gov.cn/xxgk/ndsj/ (accessed August 2, 2021).

Giffard-Roisin, S., Yang, M., Charpiat, G., Bonfanti, C. K., and Monteleoni, C. (2020). Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data. *Front. Big Data* 3, 1. doi: 10.3389/fdata.2020.00001

Gu, B., Zhang, T., Meng, H., and Zhang, J. (2021). Short-term forecasting and uncertainty analysis of wind power based on long short-term memory, cloud model and non-parametric kernel density estimation. *Renew. Energy* 164, 687–708. doi: 10.1016/j.renene.2020.09.087

Guangdong Marine Disaster Bulletin. (2014–2021). Available online at: http://nr.gd.gov.cn/gkmlpt/search?keywords=%E5%B9%BF%E4%B8%9C%E7%9C%81%E6%B5%B7%E6%B4%8B%E7%81%BE%E5%AE%B3%E5%85%AC%E6%8A%A5 (accessed August 2, 2021).

Guangdong Statistical Yearbook. (1990–2020). Available online at: https://data.cnki.net/Yearbook/Single/N2020110011 (accessed August 2, 2021).

Guo, T., and Li, G. (2020). Study on methods to identify the impact factors of economic losses due to typhoon storm surge based on confirmatory factor analysis. *Nat. Hazard.* 100, 515–534. doi: 10.1007/s11069-019-03823-w

Hao, Y., Niu, X., Wang, J. I. (2021). Impacts of haze pollution on China's tourism industry: a system of economic loss analysis. *J. Environ. Manage.* 295, 113051. doi: 10.1016/j.jenvman.2021.113051

Hao, Y., and Tian, C. (2019). A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting. *Appl. Energy* 238, 368–383. doi: 10.1016/j.apenergy.2019.01.063

Huang, G-B., Zhu, K.-Y., and Siew, C.-K.. (2004). "Extreme learning machine: a new learning scheme of feedforward neural networks," in *IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)* (Budapest). doi: 10.1109/IJCNN.2004.1380068

Jena, R., Pradhan, B., Beydoun, G., Alamri, A. M., Ardiansyah, Nizamuddin, et al. (2020). Earthquake hazard and risk assessment using machine learning approaches at Palu, Indonesia. *Sci. Total Environ.* 749, 141582. doi: 10.1016/j.scitotenv.2020.141582

Kim, T., Song, J., and Kwon, O. S. (2020). Pre- and post-earthquake regional loss assessment using deep learning. *Earthquake Eng. Struct. Dyn.* 49, 657–678. doi: 10.1002/eqe.3258

Kohonen, T. (1990). The self-organizing map. *IEEE. Proc Icnn*. 1, 1–6. doi: 10.16/S0925-2312(98)00030-7

Lakshmi, D. D., Murty, P. L. N., Bhaskaran, P. K., Sahoo, B., Kumar, T. S., Shenoi, S. S. C., et al. (2017). Performance of WRF-ARW winds on computed storm surge using hydodynamic model for Phailin and Hudhud cyclones. *Ocean Eng.* 131, 135–148. doi: 10.1016/j.oceaneng.2017.01.005

Li, B., Zhang, Y., Zhang, S., and Li, W. (2021). Prediction of grain yield in Henan province based on grey bp neural network model. *Discrete Dyn. Nat. Soc.* 2021, 1–13. doi: 10.1155/2021/9919332

Liu, B., Lai, M., Wu, J. L., Fu, C., and Binaykia, A. (2020). Patent analysis and classification prediction of biomedicine industry: som-kpca-svm model. *Multimedia Tool. Appl.* 79, 10177–10197. doi: 10.1007/s11042-019-7422-x

Liu, W., and Xu, Y. (2020). Randomised learning-based hybrid ensemble model for probabilistic forecasting of PV power generation. *IET Gener. Transm. Distrib.* 14, 5909–5917. doi: 10.1049/iet-gtd.2020.0625

Nahornyi, V., Panda, A., Straser, V., and Pandova, I. (2021). Forecasting an vibration by monitoring the dynamics of changes its precursors of various physical nature. *MM Sci.* 2, 4396–4399. doi: 10.17973/MMSJ.2021_6_2021019

Nanhai Marine Disaster Bulletin (2016–2021). Available online at: http://scs.mnr.gov.cn/scsb/gbytj/list.shtml (accessed 2 August, 2021).

Neumann, J. E., Emanuel, K., Ravela, S., Ludwig, L., Kirshen, P., Bosma, K., et al. (2015). Joint effects of storm surge and sea-level rise on US Coasts: new economic estimates of impacts, adaptation, and benefits of mitigation policy. *Clim. Change* 129, 337–349. doi: 10.1007/s10584-014-1304-z

Niu, X., and Wang, J. (2019). A combined model based on data preprocessing strategy and multi-objective optimization algorithm for short-term wind speed forecasting. *Appl. Energy* 241, 519–539. doi: 10.1016/j.apenergy.2019.03.097

Ohz, A., Klein, A. H. F., and Franco, D. (2020). "A multiple linear regression-based approach for storm surge prediction along South Brazil," in *Climate Change, Hazards and Adaptation Options*, eds W. Leal Filho, G. Nagy, M. Borga, P. Chávez Muñoz, and A. Magnuszewski (Cham: Springer), 27–50. doi: 10.1007/978-3-030-37425-9_2

Pawlak, Z. (1982). Rough sets. *Int. J. Comput. Inf. Sci.* 11, 341–356. doi: 10.1007/BF01001956

Pulinets, S., Krankowski, A., Hernandez-Pajares, M., Marra, S., Cherniak, I., Zakharenkova, I., et al. (2021). Ionosphere sounding for pre-seismic anomalies identification (inspire): results of the project and perspectives for the short-term earthquake forecast. *Front. Earth Sci.* 9, 610193. doi: 10.3389/feart.2021.610193

Qi, P., and Du, M. (2018). Multi-factor evaluation indicator method for the risk assessment of atmospheric and oceanic hazard group due to the attack of tropical cyclones. *Int. J. Appl. Earth Obs. Geoinf.* 68, 1–7. doi: 10.1016/j.jag.2018.01.015

Sahin, S. (2019). Forecasting of Turkey's greenhouse gas emissions using linear and nonlinear rolling metabolic grey model based on optimization. *J. Cleaner Prod.* 239, 118079. doi: 10.1016/j.jclepro.2019.118079

Sahoo, B., and Bhaskaran, P. K. (2018). Multi-hazard risk assessment of coastal vulnerability from tropical cyclones—a GIS based approach for the Odisha coast. *J. Environ. Manag.* 206, 1166–1178. doi: 10.1016/j.jenvman.2017.10.075

Sawant, S., Mohite, J., Sakkan, M., and Pappula, S. (2019). "Near real time crop loss estimation using remote sensing observations," in *8th International Conference on Agro-Geoinformatics, Agro-Geoinformatics* (Istanbul). doi: 10.1109/Agro-Geoinformatics.2019.8820217

Soltani, K., Ebtehaj, I., Amiri, A., Azari, A., Gharabaghi, B., and Bonakdari, H. (2021). Mapping the spatial and temporal variability of flood susceptibility using remotely sensed normalized difference vegetation index and the forecasted changes in the future. *Sci. Total Environ.* 770, 145288. doi: 10.1016/j.scitotenv.2021.145288

Wang, D., He, W., and Shi, R. (2019a). How to achieve the dual-control targets of china's CO_2 emission reduction in 2030? future trends and prospective decomposition. *J. Cleaner Prod.* 213, 1251–1263. 1 doi: 10.1016/j.jclepro.2018.12.178

Wang, J., Yang, W., Du, P., and Li, Y. (2018). Research and application of a hybrid forecasting framework based on multi-objective optimization for electrical power system. *Energy* 148, 59–78. doi: 10.1016/j.energy.2018.01.112

Wang, S., Mu, L., Qi, M., Yu, Z., Yao, Z., and Zhao, E. (2021). Quantitative risk assessment of storm surge using GIS techniques and open data: a case study of Daya Bay Zone, China. *J. Environ. Manage.* 289, 112514. doi: 10.1016/j.jenvman.2021.112514

Wang, S., Zhao, Y., and Wiedmann, T. (2019b). Carbon emissions embodied in china–australia trade: a scenario analysis based on input–output analysis and panel regression models. *J. Cleaner Prod.* 220, 721–731. doi: 10.1016/j.jclepro.2019.02.071

Xiong, T., Li, C., and Bao, Y. (2018). Seasonal forecasting of agricultural commodity price using a hybrid stl and elm method: evidence from the vegetable market in china. *Neurocomputing* 275, 2831–2844. doi: 10.1016/j.neucom.2017.11.053

Yan, Z., Liang, B., Wu, G., Wang, S., and Li, P. (2020). Ultra-long return level estimation of extreme wind speed based on the deductive method. *Ocean Eng.* 197, 106900. doi: 10.1016/j.oceaneng.2019.106900

Yang, W., Sun, S., Hao, Y., and Wang, S. (2022). A novel machine learning-based electricity price forecasting model based on optimal model selection strategy. *Energy* . 238, 121989. doi: 10.1016/j.energy.2021.121989

Yang, W., Wang, J., Lu, H., Niu, T., and Du, P. (2019a). Hybrid wind energy forecasting and analysis system based on divide and conquer scheme: a case study in China. *J. Clean. Prod.* 222, 942–959. doi: 10.1016/j.jclepro.2019.03.036

Yang, W., Wang, J., Niu, T., and Du, P. (2019b). A hybrid forecasting system based on a dual decomposition strategy and multi-objective optimization for electricity price forecasting. *Appl. Energy* 235, 1205–1225. doi: 10.1016/j.apenergy.2018.11.034

Yi, A., Yang, M., and Li, Y. (2021). Macroeconomic uncertainty and crude oil futures volatility–evidence from china crude oil futures market. *Front. Environ. Sci.* 9, 636903. doi: 10.3389/fenvs.2021.636903

Yin, K., Zhang, Y., and Li, X. (2017). Research on storm-tide disaster losses in China using a new grey relational analysis model with the dispersion of panel data. *Int. J. Environ. Res. Public Health* 14, 1330. doi: 10.3390/ijerph14111330

Yin, K. D., and Sun, W. J. (2011). Study on economic loss evaluation index system of storm surge disaster, Fish. *Econ. China*. 29:4. doi: 10.3969/j.issn1009-590X.2011.03.013

Yin, K. D., Wang, B., and Ma, J. (2011). Empirical Study on economic early warning index system of storm surge disaster, China Science and Technology *Forum Fam. Plan. West. Hemisph.* 8. doi: 10.3969/j.issn.1002-6711.2011.07.017

Young, C. C., Liu, W. C., and Wu, M. C. (2017). A physically based and machine learning hybrid approach for accurate rainfall-runoff modeling during extreme typhoon events. *Appl. Soft Comput.* 53, 205–216. doi: 10.1016/j.asoc.2016.12.052

Yu, F., Dong, J., and Ye, L. (2015). *Collection of storm surge disasters historical data in China 1949–2009, 1st Edn.* Beijing: China Ocean Press.

Zeng, D., Zhang, H., Li, Q., and Ellingwood, B. R. (2021). Tropical cyclone damage assessment of distributed infrastructure systems under spatially correlated wind speeds. *Struct. Saf.* 91, 102080. doi: 10.1016/j.strusafe.2021.102080

Zhao, X., Li, H., Ding, L., and Liu, M. (2019). Research and application of a hybrid system based on interpolation for forecasting direct economic losses of marine disasters. *Int. J. Disaster Risk Reduct.* doi: 10.1016/j.ijdrr.2019.101121

Zhao, X., Li, H., Ding, L., Wang, W., and Xue, Y. (2020). Forecasting direct economic losses of marine disasters in China based on a novel combined model. *Int. J. Disaster Risk Reduct.* 51, 101921. doi: 10.1016/j.ijdrr.2020.101921

Zhejiang Marine Disaster Bulletin. (2012–2021). Available online at: http://search.zj.gov.cn/jrobotfront/search.do?websiteid=330000000000014andpg=andp=1andtpl=2330andcategory=andq=%E6%B5%99%E6%B1%9F%E7%9C%81%E6%B5%B7%E6%B4%8B%E7%81%BE%E5%AE%B3%E5%85%AC%E6%8A%A5andx=0andy=0 (accessed August 2, 2021).

Zhejiang Statistical Yearbook. (1990–2020). Available online at: https://data.cnki.net/Yearbook/Single/N2020090355. (accessed August 2, 2021).

Zhi, G., Liao, Z., Tian, W., and Wu, J. (2020). Urban flood risk assessment and analysis with a 3D visualization method coupling the PP-PSO algorithm and building data. *J. Environ. Manage.* 268:110521. doi: 10.1016/j.jenvman.2020.110521

Zhou, L., Hu, W., Jia, Z., Li, X., Li, Y., Su, T., et al. (2020). Integrated visualization approach for real-time and dynamic assessment of storm surge disasters for China's seas. *ISPRS Int. J. Geo-Inf.* 9, 51. doi: 10.3390/ijgi9010051

Žust, L., Fettich, A., Kristan, M., and Ličer, M. (2020). HIDRA 1.0: deep-learning-based ensemble sea level forecasting in the Northern Adriatic. *Geosci. Model Dev. Discuss.* 14, 2057–2074. doi: 10.5194/gmd-2020-355

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.

Copyright © 2022 Guo, Yin and Huang. 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 in Environmental Science

ORIGINAL RESEARCH
published: 25 February 2022
doi: 10.3389/fenvs.2022.833374

Design of a Combined System Based on Multi-Objective Optimization for Fine Particulate Matter ($PM_{2.5}$) Prediction

Lu Bai[1], *Hongmin Li*[2]*, *Bo Zeng*[3] *and Xiaojia Huang*[4]

[1]Department of Mathematics, University of Macau, Macau, China, [2]College of Economics and Management, Northeast Forestry University, Harbin, China, [3]Collaborative Innovation Center for Chongqing's Modern Trade Logistics and Supply Chain, Chongqing Technology and Business University, Chongqing, China, [4]School of Information Science and Engineering, Lanzhou University, Lanzhou, China

OPEN ACCESS

Edited by:
Yiliao Song,
University of Technology Sydney, Australia

Reviewed by:
Tong Niu,
Zhengzhou University, China
Danxiang Wei,
Macau University of Science and Technology, China

***Correspondence:**
Hongmin Li
lhm@nefu.edu.cn

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Environmental Science

Received: 11 December 2021
Accepted: 17 January 2022
Published: 25 February 2022

Citation:
Bai L, Li H, Zeng B and Huang X (2022) Design of a Combined System Based on Multi-Objective Optimization for Fine Particulate Matter ($PM_{2.5}$) Prediction. Front. Environ. Sci. 10:833374. doi: 10.3389/fenvs.2022.833374

Air pollution forecasting plays a pivotal role in environmental governance, so a large number of scholars have devoted themselves to the study of air pollution forecasting models. Although numerous studies have focused on this field, they failed to consider fully the linear feature, non-linear feature, and fuzzy features contained in the original series. To fill this gap, a new combined system is built to consider features in the original series and accurately forecast $PM_{2.5}$ concentration, which incorporates an efficient data decomposition strategy to extract the primary features of the $PM_{2.5}$ concentration series and remove the noise component, and five forecasting models selected from three types of models to obtain the preliminary forecasting results, and a multi-objective optimization algorithm to combine the prediction results to produce the final prediction values. Empirical studies results indicated that in terms of RMSE the developed combined system achieves 0.652 6%, 0.810 1%, and 0.775 0% in three study cities, respectively. Compared to other prediction models, the RMSE improved by 60% on average in the study cities.

Keywords: combined forecasting model, air pollution forecasting, improved extreme learning machine, data decomposition, multi-objective optimization approach, fuzzy computation and forecasting

1 INTRODUCTION

Atmosphere pollutants can cause a variety of diseases (Organization, 2014, March 25; Glencross et al., 2020), and cause other environmental problems (Grennfelt et al., 2020; Manisalidis et al., 2020), endangering human survival. To alleviate the impacts of atmosphere pollution, support environmental management, more scholars are focusing on air pollution forecasting.

Air pollution forecasting is a complex task since there are multiple influences on pollutant concentrations, such as weather, wind speed and direction, geographic location, pollution emission and absorption, and policies, etc. Therefore, the concentration series are chaotic and usually contain both linear and non-linear features (Niska et al., 2004). In the past decades, the forecasting of air

pollution has attracted wide academic interest, and much effort has been made to forecast concentration using various approaches. Generally speaking, these approaches can be divided into four categories: individual models, hybrid models, combined models, and meteorological models. The meteorological models are based on the physical and chemical processes of pollutants in the atmosphere. This type of model is the subject of atmospheric research. For individual methods, the concentration series are modeled and forecast by one type of model, such as the traditional statistical model, Auto-Regressive Integrated Moving-Average (ARima), neuron network model, Back-propagation Neural Network (BPnn), etc. Research on this type of model was mainly concentrated before 2010. Such as Niska et al. (2004) used a parallel genetic algorithm to select the inputs for the multi-layer perceptron model to forecast hourly concentrations of nitrogen dioxide. Goyal et al. (2006) compared the performance of three statistical models for forecasting the concentration of respirable suspended particulate matter. These three models are multiple linear regression, ARima, and the combination of ARima and Multiple linear regression. The prediction results show that the combination of ARima and Multiple linear regression performs better. Kurt et al. (2008) built an online forecasting system by utilizing BPnn to predict the concentrations of SO_2, PM_{10} and CO.

With the development of forecasting methods, a new type of forecasting method, the hybrid model, has been proposed and widely used. The hybrid models can advance forecasting by combining different forecasting techniques, such as combining statistical models and machine learning methods. This combination can compensate for the limitations of individual methods by taking advantage of different methods. Zhu et al. (2017) decomposed the original data into several intrinsic mode functions (IMFs, containing the important information) and noise series. Then, they built two hybrid forecasting models to forecast the daily air quality index, including least square support vector regression, Holt-Winters additive model, Grey model, and seasonal ARima. By combining the Hampel identifier, empirical wavelet transform, Elman neural network, and Outlier-robust extreme learning machine, a novel hybrid algorithm was proposed in (Liu et al., 2019), which improved the forecasting accuracy of fine particle concentrations. Similarly, using a data preprocessing module and an optimal forecasting module, Wang et al. (2020a) proposed a new well-performing hybrid model to forecast daily air quality, which combines Hampel identifier, Variational mode decomposition, Sine cosine algorithm, and Extreme learning machine to forecast daily air quality.

With the development of different forecasting techniques, combined forecasting has gradually become the research focus of scholars. The main idea of the combined models is to combine the forecasting results of several individual models. Yang et al. (2020) proposed a combined forecasting system combining Complementary ensemble empirical mode decomposition (CEEMD), BPnn, Extreme learning machine, and Double Exponential Smoothing, then used fuzzy theory and Cuckoo search algorithm to determine aggregation weights to obtain final results. Based on the wavelet transform and neural networks, Liu et al. (2021) constructed a new combined model. In their study, discrete Wavelet transform was used to decompose the NO_2 concentration series. Next to the Long short-term memory neural network (LSTM), Gated recurrent units and Bi-directional LSTM were utilized to forecast NO_2 concentration. Finally, they applied two numerical weighting methods combining the three single forecasting results.

However, these forecasting models have various problems. Because of their simple structure and convenient calculation, statistical models have been widely used, but the linear mapping and poor extrapolation limit the forecasting performance of such models (Wang et al., 2020c). Artificial intelligence methods are widely used own to their strong learning ability and ability to handle nonlinear features, but such methods tend to fall into local optima and overfitting. Moreover, their performance is dependent on artificially set hyperparameters (Niu and Wang, 2019). To avoid the defects of the individual models, several hybrid models have been developed. However, hybrid models still do not always perform best using only one single predictor, since the single model cannot capture various features contained in the series (Yang et al., 2020). Therefore, the combination models gradually developed. However, previous combined models usually combine a certain type of model. This combination can only continuously extract one type of feature in the series, and still cannot analyze the multiple features contained in the series. This paper summarizes the above-mentioned types of models in **Table 1**. To fill this gap, a novel combined model containing a data decomposition module, a forecasting module consisting of different types of forecasting models, and a combination module weighted by multi-objective optimization algorithms is proposed in this paper. More specifically, the complete ensemble empirical mode decomposition with adaptive noise (cEEMDan) strategy is used for data decomposition to reduce the influence of the noise in the original series. Whereafter, five predictors from three types of models are introduced to construct the forecasting module. These five predictors are one statistical model, three neuron networks, and a hesitant fuzzy forecasting model. The multi-optimization algorithm is utilized to aggregate the forecasting results of five individual models to obtain the final forecasting results.

Based on the above content, the main contributions and innovations of this research are summarized as follows:

1) A novel combined forecasting system is proposed by combining with data decomposition strategy, forecasting models, and multi-objective optimization algorithm. To obtain better forecasting performance, the strategy of "decomposition and ensemble" is introduced to capture different features and remove the noise of the original data, five individual models are used to forecast the decomposed data, and a multi-objective optimization algorithm is utilized to obtain the optimal weights of individual models and integrate them. The empirical experiments demonstrated that the proposed combined forecasting system can provide accurate prediction results for $PM_{2.5}$ concentration forecasting, and can provide data support for decision-making.

TABLE 1 | Summary of the different types of models.

Models	References	Variables	Results	Advantages	Dis-advantages
Meteorological Models					
ADMS-Urban	Dėdelė and Miškinytė (2019)	PM_{10}	According to the analysis of PM_{10} in the study cities, the ADMS-Urban model takes into account the different characteristics of the sites and can be applied to the exposure estimates in the cohort studies		
AERMOD	Mousavi et al. (2021)	CO, CO_2, SO_2	According to the experimental results, the CO concentration of 8 h and the SO_2 concentration of 1 h in the cold season may aggravate the impact on the breathing air of residents around the studied refinery	No historical weather data is required, the accuracy is high, and the causal relationship between the input and output in the model is clear, which makes the model more readable	The models are very computationally intensive and time-consuming, and the quality of the input data has a significant impact on the prediction results, as even small data deviations can lead to large differences in the results
CRTM and WRF-Chem	Cheng et al. (2019)	PM_{10}	The results show that assimilation of Lidar data can effectively improve the prediction effect. The predicted $PM_{2.5}$ concentration of the constructed model is closer to the observed value, and the low deviation of the model is significantly reduced		
Statistical models					
ARima	Zhang et al. (2018)	$PM_{2.5}$	The trend of fluctuations in PM2.5 concentrations in the forecast period is similar to the trend in the first two of the forecast period, which is a seasonal fluctuation	The structure of the statistical model is simple, so it is easy to implement and easy to calculate	This kind of models need a large amount of historical data. The statistical models cannot analyze non-linear series, and have poor extrapolation
MLR and ARima and MLR-ARima	Goyal et al. (2006)	PM	According to the experimental results, the prediction performance of the combination of ARima and multiple regression is better		
Neuron networks					
MLP and GA	Niska et al. (2004)	NO_2	The results show that the GA is able to reduce computation by eliminating irrelevant inputs and search for feasible high-level architectures	The neuron networks have strong learning ability and can handle non-linear features in the data	This kind of models need a large amount of historical data. And may fall into the local optima and overfitting. Moreover, their performance is dependent on artificially set hyper-parameters
BPnn	Kurt et al. (2008)	SO_2, PM_{10}, CO	Experiments show that quite accurate predictions of air pollutant indicator levels are possible with proposed online air pollution forecasting system		
Hybrid models					
EMD-SVR-SARima and EMD-HW/GM-SARima	Zhu et al. (2017)	AQI	The proposed hybrid model can be used as an effective and simple tool for air pollution early warning and management, and can be applied to predict other pollution indices		
HI-IEWT-Enn-ORelm	Liu et al. (2019)	$PM_{2.5}$	The performance of the proposed model is improved in multi-step forecasting, while the reconstruction method solves the overfitting problem and improves the stability of the hybrid model	Hybrid models can integrate the advantages of individual models, so that the forecasting more accuracy	This kind of models not always perform best using only one type of models, since they cannot capture various features contained in the series
HI-VMD-SCA-ELM	Wang et al. (2020a)	AQI	The proposed hybrid model gives a new feasible method for air pollution forecasting, which is beneficial to air quality management		

2) Three types of forecasting models are introduced to establish the robust forecasting module. In order to fully analyze the various features contained in the series, three types of forecasting models are combined. Since there are multiple influences on air pollution, the pollutant concentration series are chaotic and usually contain linear and non-linear features. The utilized three different types of models can analyze different features in the series, the statistical model can deal with linear features, neuron networks can cope with non-linear features, and the hesitant fuzzy forecasting model is used to analyze the fuzzy features. This ensures the diversity of the system and avoids that the combined model focuses on a certain type of specific model while ignoring other features in the series.
3) A multi-objective optimization algorithm is used to weight the individual forecasting models. In this study, the final forecasting results are equal to the weighted sum of individual model forecasting results, so the weight of each model is a key to ensuring forecasting accuracy. Most previous studies used numerical weighting methods, so this paper compares several numerical weighting methods with optimization algorithms. In addition, the idea of some feature selection methods can also be regarded as a kind of weighting, so this paper also chooses two feature selection methods, Max-Relevance and Min-Redundancy (MRMR) and ReliefF, as weighting methods to participate in the comparison of weighting methods. However, after the comparison in this study, the multi-objective optimization algorithm is proven to be the best weighting method, outperforming not only numerical methods but also feature selection methods.

For the convenience of the readers, all abbreviation words are listed in **Table 2**. The remainder of this paper is organized as follows: the basic methodology of utilized methods and the system design is introduced in **Section 2**. The experimental design, the experiment results, and the analysis of the results are presented in **Section 3**. The significance test and stability test are discussed in **Section 4**. Finally, **Section 5** provides the conclusion of this study.

2 FRAMEWORK OF THE DEVELOPED COMBINED FORECASTING SYSTEM

In this section, the utilized methodologies of the combined system are introduced. These methodologies include the cEEMDan, ARima, BPnn, $\ell_{2,1}$-norm and Random Fourier Mapping-Based Extreme Learning Machine ($\ell_{2,1}$RFelm), Echo state network (ESn), Fuzzy time series forecasting based on hesitant fuzzy sets (HFs) and Multi-objective salp swarm algorithm (mSSa).

2.1 Data Decomposition

Due to various factors, the monitoring data, especially the air pollutant concentration data, will have fluctuations and noise, which will affect the further analysis of the data. Therefore, to extract the characteristics of the series, cEEMDan is used to decompose the original series.

cEEMDan is an improved method based on the Empirical Mode Decomposition (EMD) method, which adds adaptive noise series at each stage of the EMD decomposition to make the decomposition more perfect while avoiding mode mixing problem (Torres et al., 2011). EMD-series methods can decompose any complicated series into a finite of intrinsic mode functions (IMFs), and each IMF represents the implicit characteristics of the original series.

The decomposition results of EMD are some IMFs and residuals, and the decomposition process is the process of finding the IMFs. Assume the original $PM_{2.5}$ concentration series $X(t)$, $t = 1, \ldots, n$ is decomposed into k IMFs, the EMD process can be summarized as follows:

Let $\boldsymbol{a_0(t)} = X(t)$ be the signal being analyzed, find all the local maximum and minimum of a_0, and interpolate to form upper and lower envelopes, denoted as $\boldsymbol{a_0^{max}}$ and $\boldsymbol{a_0^{min}}$, respectively. Calculate the mean of upper and lower envelopes as $m_{11}(t) = [\boldsymbol{a_0^{max}}(t) + \boldsymbol{a_0^{min}}(t)]/2$. Next, extract the first detailed component as $\bar{\bar{\boldsymbol{\chi}}}_{11}(t) = \boldsymbol{a_0}(t) - m_{11}(t)$. If $\bar{\bar{\boldsymbol{\chi}}}_{11}(t)$ satisfies the two conditions of IMFs, $\bar{\bar{\boldsymbol{\chi}}}_{11}(t)$ is the first IMF, denoted as IMF_1; else, $\bar{\bar{\boldsymbol{\chi}}}_{11}(t)$ is considered as the signal, and repeat the ***Step 1*-*Step 3*** until the decomposition result $\bar{\bar{\boldsymbol{\chi}}}_{1j}$ satisfies the conditions at j-th decomposition, $\mathbf{IMF_1}(\mathbf{t}) = \mathbf{a_0}(\mathbf{t}) - \sum_{i=1}^{r} m_{1j}(t)$. And the first residue is $\hat{\boldsymbol{\gamma}}_1(t) = \sum_{i=1}^{r} m_{1j}(t)$. Set $\hat{\boldsymbol{\gamma}}_1$ as the signal to be decomposed, and keep repeating the ***Step 1*-*Step 4*** until the final residual $\hat{\boldsymbol{\gamma}}_k$ becomes a monotonic function. At the end of this decomposition, the original series can be represented as $X(t) = \sum_{i=1}^{k} \mathbf{IMF_k}(\mathbf{t}) + \hat{\boldsymbol{\gamma}}_k(t)$.

Since the EMD method is subject to mode mixing, the ensemble EMD (EEMD) method is proposed to alleviate this problem by adding white noise to the original signal. However, EEMD with high computational cost and the number of decomposed IMFs varies with the added noise. To overcome the aforementioned problem, an improved EEMD method is proposed (Wang et al., 2020b). Let $\boldsymbol{w^i}$, $i = 1, \ldots, I$ be white noise with standard deviation $\boldsymbol{\varepsilon_j}$. Based on the EMD, the process of cEEMDan can be described as following. Add white noise into the original signal, then the signals being analyzed are $\tilde{\boldsymbol{a}}_{\mathbf{0}}^{\boldsymbol{i}}(\boldsymbol{t}) = X(t) + \boldsymbol{\varepsilon_0 w^i}(\boldsymbol{t})$, $i = 1, \ldots, I$. Using EMD decompose $\tilde{\boldsymbol{a}}_{\mathbf{0}}^{\boldsymbol{i}}$ to obtain its first IMF, denoted as $\widetilde{\mathbf{IMF}}_{\mathbf{1}}^{\boldsymbol{i}}$. Then, the first IMF after cEEMDan of $X(t)$ is $\overline{\mathbf{IMF}}_{\mathbf{1}}(\boldsymbol{t}) = \frac{1}{I}\sum_{i=1}^{I} \widetilde{\mathbf{IMF}}_{\mathbf{1}}^{\boldsymbol{i}}(\boldsymbol{t})$. And the residual after first decomposition is $\tilde{r}_1(t) = X(t) - \overline{\mathbf{IMF}}_{\mathbf{1}}(\boldsymbol{t})$. Let $\tilde{r}_1$ as the signal need further decomposition, construct the signal by the $\tilde{\boldsymbol{a}}_{\mathbf{1}}^{\boldsymbol{i}}(\boldsymbol{t}) = \tilde{r}_1(t) + \varepsilon_1 E_1[\boldsymbol{w^i}(t)]$, $i = 1, \ldots, I$, where $E_1(\cdot)$ represents the first IMF obtained by EMD method. The second IMF can be calculated as $\overline{\mathbf{IMF}}_{\mathbf{2}}(\boldsymbol{t}) = \sum_{i=1}^{I} E_1[\tilde{\boldsymbol{a}}_{\mathbf{1}}^{\boldsymbol{i}}(\boldsymbol{t})]/I$. For $k = 2, \ldots, K$, calculate the k-th residue by $\tilde{r}_k(t) = \tilde{r}_{k-1}(t) - \overline{\mathbf{IMF}}_{\boldsymbol{k}}(\boldsymbol{t})$, and decompose $\tilde{\boldsymbol{a}}_{\boldsymbol{k}}^{\boldsymbol{i}}(\boldsymbol{t}) = \tilde{r}_k(t) + \varepsilon_k E_k[\boldsymbol{w^i}(t)]$, then $(k+1)$-th IMF can be computed as $\overline{\mathbf{IMF}}_{\boldsymbol{k+1}}(\boldsymbol{t}) = \frac{1}{I}\sum_{i=1}^{I} E_1[\tilde{\boldsymbol{a}}_{\boldsymbol{k}}^{\boldsymbol{i}}(\boldsymbol{t})]$, where $E_k(\cdot)$ is the k-th IMF obtained by EMD. Repeat the decomposition processes until the residue cannot be further decomposed. After decomposition, the given signal $X(t)$ can be expressed as $X(t) = \sum_{k=1}^{K} \overline{\mathbf{IMF}}_{\boldsymbol{k}}(\boldsymbol{t}) + \tilde{r}_K(t)$, where $\tilde{r}_K(t)$ is the final residue that is no longer feasible to be decomposed. Compared with

TABLE 2 | List of nomenclature.

ADMS	Atmospheric dispersion modelling system	LSTM	Long short-term memory neural network
AERMOD	American meteorological society environmental policy agency regulatory model	$\ell_{2,1}$RFelm	$\ell_{2,1}$-norm and Random fourier mapping-based extreme learning machine
AIC	Akaike information criterion	MA	Moving average model
AR	Auto-regressive model	MAE	The mean absolute error
ARima	Auto-regressive integrated moving average	MAPE	The mean absolute percentage error
BPnn	Back-propagation neural network	MLP	Multi-layer perceptron model
CEEMD	Complementary ensemble empirical mode decomposition	MLR	Multiple linear regression
cEEMDan	Complete ensemble empirical mode decomposition with adaptive noise	MRMR	Max-relevance and min-redundancy
CRTM	Community radiative transfer model	mSSa	Multi-objective salp swarm algorithm
DM	Diebold-mariano test	ORelm	Outlier-robustness extreme learning machine
EEMD	Ensemble empirical Mode decomposition	PM	Particulate matters
EMD	Empirical mode decomposition	PRD	The pearl river delta in China
Enn	Elman neural network	QD	Quartile deviation
ESn	Echo state network	RMSE	The root mean squared error
GA	Genetic algorithm	SARima	Seasonal ARima
GM	Grey model	SCA	Sine cosine algorithm
GZ	Guangzhou	SD	Standard deviation
HFs	Fuzzy time series forecasting based on hesitant fuzzy sets	SVR	Support vector regression
HI	Hample Identifier	SZ	Shenzhen
HW	Holt-winters	VMD	Variational mode decomposition
IEWT	Inverse empirical wavelets transform	VR	Variance ratio
IMFs	Intrinsic mode functions	WRF-Chem	Weather research and forecasting model coupled to chemistry
LA	Lichtenberg algorithm	ZH	Zhuhai

cEEMD, cEEMDan has reduced the computational cost (Torres et al., 2011), and Wang et al. (2014) has proved that the computational complexity of EEMD is equivalent to $K\mathcal{O}(TlogT)$, where T is the number of the sample. Therefore, the computational complexity of cEEMDan is less than $K\mathcal{O}(TlogT)$.

2.2 Individual Forecasting Methods

In this study, three different types of methods are utilized to predict the concentration of $PM_{2.5}$. More details are introduced in the following subsections.

2.2.1 Conventional Statistical Method

This kind of method is based on statistics, with the advantages of low complexity and fast computational speed, and has a strong model interpretation. One of the most popular and important models is the Auto-regressive Integrated Moving Average (ARima), which has been widely used in time series forecasting (Pai and Lin, 2005; Ariyo et al., 2014; Benvenuto et al., 2020).

For the ARima model, future values are considered as a linear combination of past values and errors, and the mathematical form of the model for predicting is expressed as follows (Pai and Lin, 2005):

$$X(t) = \bar{\bar{\Psi}}^0 + \bar{\bar{\Psi}}^1 X(t-1) + \bar{\bar{\Psi}}^2 X(t-2) + \cdots + \bar{\bar{\Psi}}^p X(t-p) + \bar{\bar{\Psi}}^t - \bar{\bar{\theta}}^1 \Upsilon_{t-1} - \bar{\bar{\theta}}^2 \Upsilon_{t-2} - \cdots - \bar{\bar{\theta}}^q \Upsilon_{t-q}, \tag{1}$$

cwhere $X(t)$, …, X $(t–p)$ are actual values, Υ_t, … ,Υ_{t-q} are random errors, $\bar{\bar{\Psi}}^0$ is the trend component, p and q are the order of the auto-regressive model (AR) and moving average model (MA), respectively. For ARima, the complexity is depended on the order of AR p and the order of MA q. When the number of sample is T, the computational complexity of ARima is $\mathcal{O}((T-p)p^2 + (T-q)q^2)$ (Gavirangaswamy et al., 2013).

2.2.2 Fuzzy Computation and Forecasting

The fuzzy time series forecasting method was first proposed by Song et al. (Song and Chissom, 1993) based on the fuzzy set theory (Zadeh, 1996). It has been continuously developed in recent decades and has been widely applied for forecasting in many fields (Singh, 2007; Cheng et al., 2016; Wang et al., 2021a). As an extension of the fuzzy sets, Torra et al. introduced the concept of hesitant fuzzy sets in 2009 (Torra and Narukawa, 2009). The specific operation steps of HFs are described as follows (Bisht and Kumar, 2016; Cheng et al., 2016; Wang et al., 2021a).

Define the universe of discourse as $\boldsymbol{U} = (X_{\min}–\sigma, X_{\max} + \sigma)$. Here $X_{\min}$ and $X_{\max}$ are the minimum and maximum of the training set, σ is the standard deviation of X. Next, using equal and unequal intervals, and triangular membership function to fuzzify the universe of discourse. The length of equal intervals is determined by the distance between the maximum and minimum values in the time series, and the length of unequal intervals is determined by using the cumulative probability distribution approach (Lu et al., 2015; Bisht and Kumar, 2016). Suppose it is divided into J intervals, each interval defined by three parameters, $\bar{\bar{\boldsymbol{x}}}_{lj}$ and $\bar{\bar{\boldsymbol{x}}}_{rj}$ for feet of intervals, and $\bar{\bar{\boldsymbol{x}}}_{mj}$ for the tip of intervals. Two expresses mathematical formula of the triangular membership function (Wang et al., 2021a):

$$f(x) = \begin{cases} 0 & x < \underline{\underline{\boldsymbol{x}}}_{lj}, \\ \dfrac{x - \underline{\underline{\boldsymbol{x}}}_{lj}}{\bar{\bar{\boldsymbol{x}}}_{mj} - \underline{\underline{\boldsymbol{x}}}_{lj}} & \underline{\underline{\boldsymbol{x}}}_{lj} \le x \le \bar{\bar{\boldsymbol{x}}}_{mj}, \\ \dfrac{\underline{\underline{\boldsymbol{x}}}_{rj} - x}{\underline{\underline{\boldsymbol{x}}}_{rj} - \bar{\bar{\boldsymbol{x}}}_{mj}} & \bar{\bar{\boldsymbol{x}}}_{mj} < x \le \underline{\underline{\boldsymbol{x}}}_{rj}, \\ 0 & x > \underline{\underline{\boldsymbol{x}}}_{rj}. \end{cases} \tag{2}$$

After this step, the membership degrees of x_i to equal intervals and the unequal intervals can be obtained, denoted as $\boldsymbol{md_e}$ and $\boldsymbol{md_u}$, respectively. Then, compute the weights of equal intervals and unequal intervals using the following formula (Bisht and Kumar, 2016),

$$\begin{cases} w_e^j = \dfrac{\grave{\boldsymbol{d}}_{ej}}{\grave{\boldsymbol{d}}_{ej} + \grave{\boldsymbol{d}}_{uj}}, \\ w_u^j = 1 - w_{ej}, \end{cases} \tag{3}$$

where d_{ej} and d_{uj} are the lengths of j-th equal and unequal intervals, w_e^j and w_u^j are the weights of j-th equal and unequal intervals, respectively. Determine the membership of every element by using aggregate hesitant fuzzy elements, and build a fuzzy set using a novel aggregation operator, which is defined as follows (Wang et al., 2021a):

$$\boldsymbol{O}(x_1, x_2, \ldots, x_n) = 1 - (1 - \boldsymbol{md_e^{ij}})^{w_e^j} (1 - \boldsymbol{md_u^{ij}})^{w_u^j}, i = 1, \ldots, n;\ j = 1, \ldots, J, \tag{4}$$

where $\boldsymbol{md_e^{ij}}$ is the membership degree of x_i to j-th equal interval, so as $\boldsymbol{md_u^{ij}}$ is the membership degree of x_i to j-th unequal interval, w_e^j represents the weight of j-th equal interval and w_u^j represents the weight of j-th unequal interval. Specifically, $w_e^j \in [0, 1]$, $\sum_{j=1}^{J} w_e^j = 1$.

Following example introduces the specific aggregation process for aggregation:

Let $X = \{x_1, x_2, x_3\}$ be a reference set. $\boldsymbol{\mathcal{H}} = \{< x_1, \{0.2, 0.4\} >, < x_2, \{0.5, 0.25\} >, < x_3, \{0.3, 0.4\} >\}$ is a hesitant fuzzy set on X, and taking w = (0.4, 0.6). Applying the aggregation method motioned above, the fuzzy elements can be obtained as follows:

$$\boldsymbol{h}(\boldsymbol{x_1}) = 1 - (1 - 0.2)^{\frac{2}{5}} (1 - 0.4)^{\frac{3}{5}} \approx 0.33$$
$$\boldsymbol{h}(\boldsymbol{x_2}) = 1 - (1 - 0.5)^{\frac{2}{5}} (1 - 0.25)^{\frac{3}{5}} \approx 0.41$$
$$\boldsymbol{h}(\boldsymbol{x_3}) = 1 - (1 - 0.3)^{\frac{2}{5}} (1 - 0.4)^{\frac{3}{5}} \approx 0.64$$

Therefore, the fuzzy set $\boldsymbol{\mathcal{A}}$ is established as $\boldsymbol{\mathcal{A}} = \{< x_1, 0.33 >, < x_2, 0.41 >, < x_3, 0.64 >\}$.

After determining the membership of every element, establish fuzzy logical relationships and fuzzy logical relationship groups. The fuzzy logical relationships are established by the rule: If $\boldsymbol{A_i}$ and $\boldsymbol{A_{i+1}}$ are the fuzzy values at time t and $t + 1$ respectively, the fuzzy logical relation is denoted as $\boldsymbol{A_i} \rightarrow \boldsymbol{A_{i+1}}$. Here, $\boldsymbol{A_i}$ is called the current state and $\boldsymbol{A_{i+1}}$ is the next state. Then, the same left-hand side of the fuzzy logical relationships is classified to form several fuzzy logical relationship groups. The main idea of forecasting is to infer the next state based on the current state. Based on the fuzzy logical relationship groups, a matrix $P_{m\times m}$ can be generated, each element in P represents the frequency of $\boldsymbol{A_i} \rightarrow \boldsymbol{A_{i+1}}$ that with the same fuzzy logical relationship. According to the max-min composition operations on fuzzy logical relationship, the fuzzy output can be obtained and defuzzify by $\hat{\boldsymbol{\Gamma}}_{\mathbf{i}} = P_i M$, here M is the combined midpoint of the triangular membership functions for equal and unequal intervals respectively, calculated as follows (Bisht and Kumar, 2016):

$$M = \frac{M_e w_e + M_u w_u}{w_e + w_u}, \tag{5}$$

where M_e and M_u is the mid points of the equal and unequal intervals. As the introduction above, the computational complexity of HFs is $\mathcal{O}(Jn)$, J is the number of the interval and n represents the number of sample.

Summarizing all this activity, **Table 3** is given to show the implementation of the HFs.

TABLE 3 | The process of the HFs.

1: Give the data set $(\mathcal{X}, \mathcal{Y})$ and J
2: Define universe of discourse $\boldsymbol{U} = [\mathcal{X}_{min} - Std(\mathcal{X}), \mathcal{X}_{max} + Std(\mathcal{X})]$
3: Divided $\boldsymbol{U}$ into J equal intervals
4: Divided $\boldsymbol{U}$ into J unequal intervals
5: **for** $i = 1, \ldots, T$ **do**
6: **for** $j = 1, \ldots, J$ **do**
7: Compute membership degree of each sample by (2)
8: **end for**
9: **end for**
10: Compute $w_e^j = \frac{\grave{d}_{ej}}{\grave{d}_{ej} + \grave{d}_{uj}}$ and $w_u^j = \frac{\grave{d}_{uj}}{\grave{d}_{ej} + \grave{d}_{uj}}$
11: Build a fuzzy set by (4)
12: Determine the fuzzy logic relationship group
13: Defuzzify and compute the forecasting outputs by (5) and $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{i}} = P_i M$

2.2.3 Machine Learning Technique

The methods based on machine learning have strong learning ability and can handle the non-linear components in the time series, so they have been widely used in some fields (Gündüz et al., 2019; Henrique et al., 2019; Volk et al., 2020; Wang et al., 2021b). In this study, three different networks were selected to analyze the series, since the features of the series are uncertain.

($\mathcal{A}$) *Back-Propagation Neural Network*

Back-propagation neural network (BPnn) is a three-layer feed-forward network with an input layer, a hidden layer, and an output layer. Each layer takes inputs only from the previous layer and sends the outputs only to the next layer. Define the input vector as $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_N}\}$, and the output vector as $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_N}\}$. Assume the input layer has I neurons, the hidden layer has H neurons, and the output layer has one neuron, the network can be constructed as **Figure 1B**, and the training processes are described as follows.

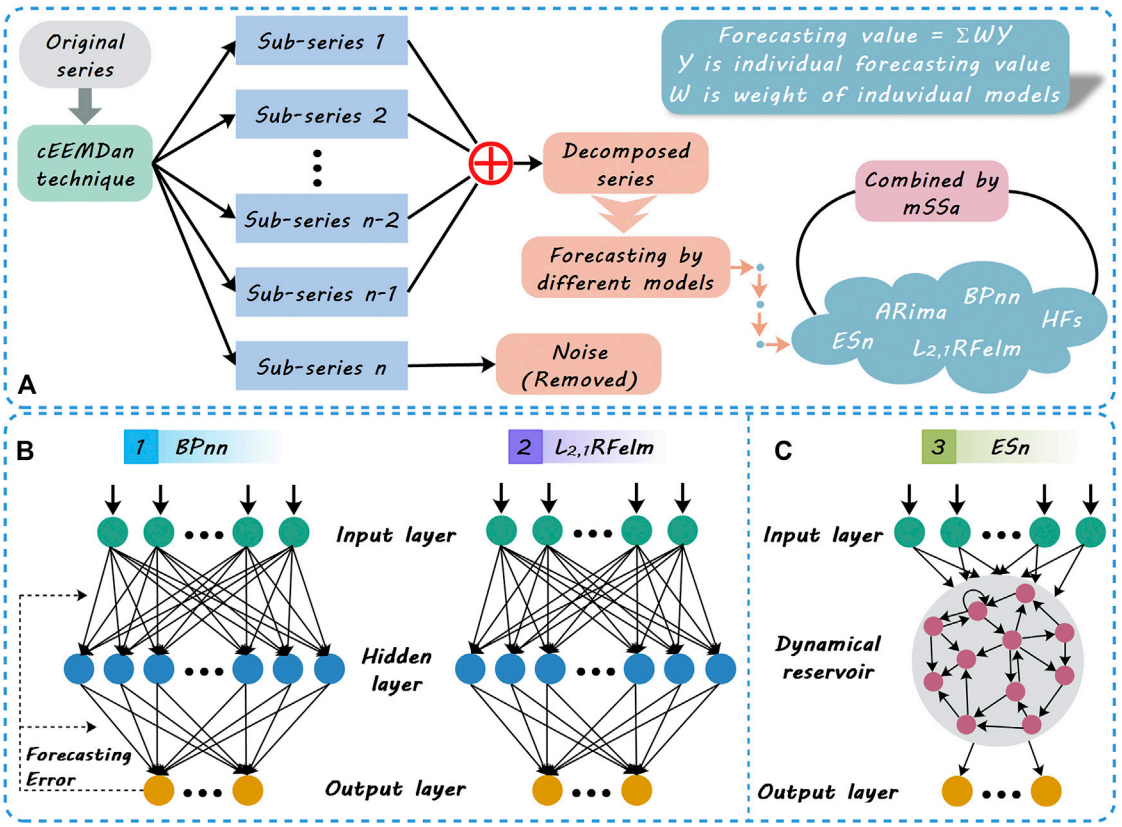


FIGURE 1 | (A) is the flowchart of the proposed combined system; **(B)** and **(C)** are the structures of the neural networks used in this paper, where BPnn and $\ell_{2,1}$RFelm have the same structure and different solution processes.

Calculate outputs of all neurons in hidden layer (Hecht-Nielsen, 1992; Wang Y. et al., 2021):

$$\begin{cases} \boldsymbol{h}_h^i = \sum_{i=1}^{I} w_{ih} x_n + \bar{\boldsymbol{\upsilon}}_h, h = 1, 2, \ldots, H, \\ \boldsymbol{h}_h^o = f(\boldsymbol{h}_h^i), h = 1, 2, \ldots, H, \end{cases} \tag{6}$$

Where, $\boldsymbol{h}_h^i$ is the activation value of the h-th node of hidden layer, $\boldsymbol{h}_h^o$ represents the output value of h-th hidden neuron, w_{ih} denotes the connection weight between i-th input neuron and h-th hidden neuron, $\bar{\boldsymbol{\upsilon}}_h$ represents the bias of h-th hidden neuron, and $f(\cdot)$ is the activation function. Then, determine the output of the network as $\boldsymbol{O}^o = g(\sum_{h=1}^{H} w_h \boldsymbol{h}_h^o + \bar{\bar{\boldsymbol{\upsilon}}})$, where $\boldsymbol{O}^o$ is the output value of output neuron, w_h is the weight between h-th hidden neuron and output neuron, $\bar{\bar{\boldsymbol{\upsilon}}}$ represents the bias of output neuron, and g $(\cdot)$ is the activation function. Obtain the minimum global error by the "error feedback" training mechanism. The global error is $\mathcal{E} = \sum_{n=1}^{N} (\boldsymbol{O}_n^o - \boldsymbol{Y}_n)^2/2$, $\boldsymbol{O}_n^o$ represents n-th output of network. For more details, please refer to (Hecht-Nielsen, 1992). For each iteration, the computational complexity of both the forward propagation process and the backward propagation process is $\mathcal{O}(N(H+1)(F+1) + N(H+1)(O+1))$, where F is the dimension of the input set and O represents the dimension of the output set. In this study, $F = 4$, $O = 1$, so the computational complexity of the algorithm is $T_{bpnn}\mathcal{O}(NH)$, here T_{bpnn} is the number of iterations.

($\mathcal{B}$) $\ell_{2,1}$-Norm and Random Fourier Mapping-Based Extreme Learning Machine

$\ell_{2,1}$RFelm is an improved feed-forward neural network with a single hidden layer, which was proposed by Zhou et al. (2016). In this method, Random Fourier Mapping is used to improve the extendibility of the network by approximating the activation function in ELM. And $\ell_{2,1}$-norm is used to make the hidden layer more compact and discriminative by cutting irrelevant neurons.

To predict the $PM_{2.5}$ concentration of the day, the $PM_{2.5}$ concentrations of the past 4 days are used. So, the original concentration series $X = \{x_1, x_2, \ldots, x_T\}$ is reconstructed as follows:

$$\begin{aligned} \boldsymbol{\mathcal{X}} &= [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N] = \begin{bmatrix} x_1 & x_2 & \cdots & x_{T-4} \\ x_2 & x_3 & \cdots & x_{T-3} \\ x_3 & x_4 & \cdots & x_{T-2} \\ x_4 & x_5 & \cdots & x_{T-1} \end{bmatrix}, \\ \boldsymbol{\mathcal{Y}} &= [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N] = [x_5, x_6, \ldots, x_{T-1}, x_T]. \end{aligned} \tag{7}$$

Then, the main processes of this method can be introduced as follows:

Randomly initialize the connection weights $\boldsymbol{\mathcal{W}}$ between the input layer and hidden layer and the bias $\boldsymbol{\mathcal{B}}$ of the hidden layer,

assume the hidden layer has H neurons, these two matrix are represented as follows:

$$\boldsymbol{\mathcal{W}} = [W_1, \ldots, W_h, \ldots, W_H]^{\mathrm{T}} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ \vdots & \vdots & \vdots & \vdots \\ w_{H1} & w_{H2} & w_{H3} & w_{H4} \end{bmatrix},$$
$$\boldsymbol{\mathcal{B}} = [b_1, \ldots, b_h, \ldots, b_H], \quad h = 1, 2, \ldots, H, \tag{8}$$

here w_{H1} represents the weight between the first input neuron and H-th hidden neuron, b_h represents the bias of h-th hidden neuron. Then the output matrix of the hidden layer is

$$\mathcal{H} = \begin{bmatrix} g(W_1\boldsymbol{X_1} + b_1) & g(W_1\boldsymbol{X_2} + b_1) & \cdots & g(W_1\boldsymbol{X_N} + b_1) \\ g(W_2\boldsymbol{X_1} + b_2) & g(W_2\boldsymbol{X_2} + b_2) & \cdots & g(W_2\boldsymbol{X_N} + b_2) \\ \vdots & \vdots & \ddots & \vdots \\ g(W_H\boldsymbol{X_1} + b_H) & g(W_H\boldsymbol{X_2} + b_H) & \cdots & g(W_H\boldsymbol{X_N} + b_H) \end{bmatrix}. \tag{9}$$

In this method, the Random Fourier Mapping $g(\cdot)$ is used to approximate the kernel function, so $\boldsymbol{\mathcal{W}}\boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{B}}$ can be mapped into a Random Fourier feature space. The specific mapping is defined as below (Rahimi and Recht, 2007):

$$g(x) = \frac{1}{\sqrt{N}}[\cos(w_1^{\mathrm{T}}x), \ldots, \cos(w_N^{\mathrm{T}}x), \sin(w_1^{\mathrm{T}}x), \ldots, \sin(w_N^{\mathrm{T}}x)]^{\mathrm{T}}. \tag{10}$$

Then, calculate the output of the network and solve parameter. Let the connection weight between the hidden layer and the output layer is $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_H]^{\mathrm{T}}$, then the output function of this network is $\sum_{h=1}^{H}\beta_h g(W_h\boldsymbol{X_n} + \bar{\boldsymbol{\delta}}_h) = O_i,\ n = 1, 2, \ldots, N$. In $\ell_{2,1}$RFelm, the only parameter need to solve is $\boldsymbol{\beta}$. Based on the given data and the initial parameters, the objective function of this network is

$$\begin{aligned} &\underset{\boldsymbol{\beta},\,\varepsilon}{\mathrm{Min}}\ \frac{1}{2}\|\boldsymbol{\beta}\|_{2,1} + \frac{1}{2}\tilde{\mathrm{C}}\sum_{i=1}^{N}\|\Upsilon_i\|^2, \\ &\mathbf{s.t.} g(\boldsymbol{X_i})\boldsymbol{\beta} = y_i - \Upsilon_i, i = 1, 2, \ldots, \mathrm{N}, \end{aligned} \tag{11}$$

chere ε represents the training error, $\tilde{\mathrm{C}}$ is the penalty coefficient, and $\|\boldsymbol{\beta}^{\mathrm{T}}\|_{2,1}$ is $\ell_{2,1}$-norm of $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_{2,1} = \sqrt{\sum_{h=1}^{H}\beta_h^2}$. Finally, β can be obtained $\hat{\boldsymbol{\beta}} = (\mathbf{D}/\tilde{\mathrm{C}} + \mathcal{H}^{\mathrm{T}}\mathcal{H})^{-1}\mathcal{H}^{\mathrm{T}}\boldsymbol{\mathcal{Y}}^{\mathrm{T}}$, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{hh} = 1/(2\|\boldsymbol{\beta}\|_2)$, and at the beginning of the iterative, $\mathbf{D}$ is an identity matrix. For more details of solve process, please refer to (Zhou et al., 2016). In this study, the computational complexity of $\ell_{2,1}$RFelm is mainly contributed by the process of computing $(\mathbf{D}/\tilde{\mathrm{C}} + \mathcal{H}^{\mathrm{T}}\mathcal{H})^{-1}$. Thus the computational complexity of $\ell_{2,1}$RFelm is $T_{rfelm}\mathcal{O}(H^3)$, T_{rfelm} is the number of the iterations.

(*C*) *Echo State Network*

Echo state network (ESn) is an improved recurrent neural network and was proposed in 2004 (Jaeger and Haas, 2004). Without output feedback connections, an ESn consists of an input layer with I neurons, L internal neurons possessing internal states, and one output neuron. The structure of ESn is shown in **Figure 1C**. Given a training set $[\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}]$ the main steps of ESn are as follows (Qiao et al., 2016; Wang et al., 2019). Randomly generate a reservoir weight matrix $\boldsymbol{W}$ with the predefined sparsity and size. In order for the reservoir to have echo-state property, the singular values of reservoir weight matrix of the reservoir must be scaled to within 1, so scaled $\boldsymbol{W}$ as $\tilde{\mathbf{W}} = (\boldsymbol{\alpha}/\boldsymbol{\Psi})\mathbf{W}$, here $0 < \boldsymbol{\alpha} < 1$ and $\boldsymbol{\Psi}$ is the spectral radius of $\boldsymbol{W}$. Next, randomly generate the weight matrix between input layer and reservoir, denoted as $\boldsymbol{W^{in}}$. And initialize the reservoir states $\mathbf{\pounds\ (0)}$. Calculate the state of reservoir by using dynamic equation, $\mathbf{\pounds\ (n+1)} = \boldsymbol{\mathcal{F}}(\tilde{\boldsymbol{W}}\ \mathbf{\pounds\ (n)} + \boldsymbol{W^{in}X_{n+1}})$, here $\mathbf{\pounds\ (n)}$ and $\mathbf{\pounds\ (n+1)}$ are reservoir states, $\boldsymbol{\mathcal{F}}(\cdot)$ is activation function, $\boldsymbol{X_{n+1}}$ represent $(n + 1)$-th sample input. Finally, calculate the network output $\hat{y}_{n+1} = \boldsymbol{\mathcal{G}}(\boldsymbol{W^{out}}\mathbf{\pounds\ (n+1)})$, where $\boldsymbol{W^{out}}$ represents weight matrix between reservoir and output layer, $\boldsymbol{\mathcal{G}}(\cdot)$ is activation function. The only trainable part of the ESn is the output weight matrix W^{out}, and can be commonly obtained as $\boldsymbol{W^{out}} = (\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\mathcal{X}})^{-1}\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\mathcal{Y}}$. As shown above, the computational complexity of ESn is largely proportional to the state updating process, the complexity of this process is equal to $\mathcal{O}(LT)$, where T is the number of sample.

2.3 Optimization of Combination Weights

Mirjalili et al. proposed a novel swarm intelligence optimization algorithm in 2017, which was inspired by the behavior of salps looking for food (Mirjalili et al., 2017a). Their study has shown that this method can approximate the Pareto optimal solution with high convergence and coverage. It has merits among the current optimization algorithms and is worth applying to different problems (Mirjalili et al., 2017a). Therefore, this method (mSSa) is used to find the optimal combined weight of different forecasting models in this study. More details are introduced as follows.

2.3.1 Multi-Objective Optimization

Multi-objective optimization is concerned with mathematical optimization problems involving more than one objective function to be optimized simultaneously (Haimes et al., 2011). The multi-objective optimization problem can represent as follows:

$$\begin{cases} \text{Minimize} & [\mathbf{Obf}_1(x), \mathbf{Obf}_2(x), \ldots, \mathbf{Obf}_o(x)], \\ \text{subject to} & x \in \mathcal{S}, \end{cases} \tag{12}$$

where $\mathcal{S}$ is the feasible search space, o is the number of objective function, and $\mathbf{Obf}_i$ is i-th objective function.

The purpose of multi-objective optimization is to find the set of acceptable solutions (Ngatchou et al., 2005). Hence, the definitions related to the Pareto-optimal solutions are introduced.

Definition 1. Pareto domination Given two vectors $\vec{\mathbf{X}} = (x_1, x_2, \ldots, x_n)$ and $\vec{\mathbf{Y}} = (y_1, y_2, \ldots, y_n)$, vector $\vec{\mathbf{Y}}$ dominates $\vec{\mathbf{X}}$ or called vector $\vec{\mathbf{X}}$ is dominated by vector $\vec{\mathbf{Y}}$ denoted as $\vec{\mathbf{Y}} \prec \vec{\mathbf{X}}$ if and only if $\forall\, i \in [1, o], [\mathbf{Obf}_i(\vec{\mathbf{Y}}) \leq \mathbf{Obf}_i(\vec{\mathbf{X}})] \wedge \exists\, i \in [1, o], [\mathbf{Obf}_i(\vec{\mathbf{Y}}) < \mathbf{Obf}_i(\vec{\mathbf{X}})]$, where $\mathbf{Obf}_{\mathrm{i}}(\cdot)$ represents i-th objective function.

Definition 2. Pareto optimal set A set including all the non-dominated solutions is called Pareto optimal set. The mathematical description is $\mathbf{P}_s := \left\{x, z \in \vec{\mathbf{X}} \mid \nexists\ z \prec x\right\}$.

TABLE 4 | Descriptive statistics of data sets.

Study areas	Data sets (number of obs)	Central tendency			Variability			Distribution	
		Mean	Median	Mode	SD	Range	QD	Kurt	Skew
GZ	All (3,600)	16.032 7	12.741 9	2.174 6	11.266 6	60.566 3	11.538 2	5.341 5	1.570 0
	Training (2,522)	14.150 6	10.196 0	2.174 6	11.561 9	60.566 3	9.107 9	7.125 8	2.077 9
	Test (1,078)	20.435 8	17.870 1	5.978 0	9.142 0	42.083 8	13.389 5	2.750 5	0.758 1
SZ	All (3,600)	12.562 1	8.560 5	1.951 5	11.187 9	70.718 5	10.637 4	7.197 1	1.983 4
	Training (2,522)	9.942 4	6.761 4	1.951 5	10.228 2	70.718 5	6.068 8	12.981 9	3.001 5
	Test (1,078)	18.690 8	16.489 5	2.886 7	10.942 2	53.130 8	13.574 4	3.671 8	0.991 9
ZH	All (3,600)	13.887 2	10.585 2	1.857 4	72.293 2	70.435 9	10.193 8	6.782 1	1.832 5
	Training (2,522)	11.636 6	8.895 4	1.857 4	72.293 2	70.435 9	6.397 2	10.645 8	2.588 2
	Test (1,078)	19.152 4	17.081 3	4.335 2	62.078 1	57.742 9	11.922 8	4.704 9	1.155 1

Note SD: standard deviation; QD: quartile deviation; Kurt. kurtosis; Skew. skewness.

TABLE 5 | Experimental parameter settings of different individual models.

Method	Meaning	Value
cEEMDan	Noise standard deviation	0.5
	Number of realizations	200
	Maximum number of sifting iterations allowed	10
ARima	The lag order	10 (GZ), 8 (SZ), 8 (ZH)
	The degree of differencing	1 (GZ), 1 (SZ), 1 (ZH)
	The order of the moving average	7 (GZ), 10 (SZ), 10 (ZH)
HTS	Number of interval	23 (GZ), 24 (SZ), 23 (ZH)
BPnn	Maximum number of iteration times	100
	Learning rate	0.1
	Training accuracy goal	0.000 01
	Neuron number of input layer	4
	Neuron number of hidden layer	9
	Neuron number of output layer	1
$\ell_{2,1}$RFelm	Penalty coefficient	5
	Maximum iterations	50
	Number of neurons in hidden layer	15
ESn	Reservoir dimension	20
	Spectral radius	0.2
	Leaking rate	0.5
	Connectivity	0.2
	Readout regularization	0.05
mSSa	Size of archive	100
	Size of population	30
	Maximum iterations	50
	Individual value range	[−5, 5]

2.3.2 Process of Multi-Objective Salp Swarm Algorithm

The individuals in a salp chain are divided into two groups: the front of the chain is the leader, the others are followers. Assume $\boldsymbol{O}$ indicates the dimension of search space, $\boldsymbol{N}$ denotes the number of salp chains, then the location of all the salps can be defined as a matrix:

$$\mathcal{P}_t = \begin{bmatrix} p_1^1(t) & p_2^1(t) & \dots & p_O^1(t) \\ p_1^2(t) & p_2^2(t) & \dots & p_O^2(t) \\ \vdots & \vdots & \dots & \vdots \\ p_1^N(t) & p_2^N(t) & \dots & p_O^N(t) \end{bmatrix}. \tag{13}$$

here t represents t-th iteration. The position of each salp is a candidate solution. Next, calculate fitness of each salp chain $\mathbf{Fit}\,[\vec{p}^{\,j}(t)] = \left\{obf_1[\vec{p}^{\,j}(t)], \quad obf_2[\vec{p}^{\,j}(t)], \dots, obf_O[\vec{p}^{\,j}(t)]\right\}$, $j = 1, 2, \dots, N$, where $\mathbf{Fit}[\vec{p}^{\,j}(t)]$ represents the fitness of j-th salp chain at t-th iteration, $obf_o[\vec{p}^{\,j}(t)]$ is the value of o-th objective function of j-th salp chain at t-th iteration. Then, determine the non-dominated salp chains according to Definition 1, and update the archive (Pareto optimal set, Definition 2). Select a salp chain as a food source from the archive, denoted as $\mathbf{F}$. After that, leaders p_1 guides the salp swarm toward the food source in an $\boldsymbol{O}$-dimensional search space. The positions of the leaders are updated as follows (Mirjalili et al., 2017a):

$$p_1^i(t+1) = \begin{cases} \mathbf{F}_i(t) + \boldsymbol{\tau}_1\left[\left(\overline{\overline{p_1}} - \underline{\underline{p}}_1\right)\boldsymbol{\tau}_2 + \underline{\underline{p}}_1\right], & \boldsymbol{\tau}_3 \geq 0, \\ \mathbf{F}_i(t) - \boldsymbol{\tau}_1\left[\left(\overline{\overline{p_1}} - \underline{\underline{p}}_1\right)\boldsymbol{\tau}_2 + \underline{\underline{p}}_1\right], & \boldsymbol{\tau}_3 < 0, \end{cases} \tag{14}$$

TABLE 6 | Forecasting results of individual models and combined systems based on the original data and decomposed data.

	GZ			SZ			ZH		
	MAE	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**
(a1) Individual models without cEEMDan									
ARima	1.3628	1.8714	7.4792	1.8384	2.6862	12.0589	1.6974	2.3692	10.2120
BPnn	2.1184	2.9313	11.3664	3.0927	4.4142	19.3875	2.6765	3.8144	15.7448
$\ell_{2,1}$RFelm	2.1446	2.9753	11.4354	2.9969	4.2871	18.9209	2.6160	3.7088	15.4394
ESn	2.1333	2.9944	11.4066	2.9570	4.1990	18.7598	2.6277	3.7268	15.4739
HFs	1.7617	2.2905	9.6849	1.6525	2.0928	10.7961	2.1028	3.0032	11.4203
(a2) Individual models combined with cEEMDan									
C-ARima	0.4834	**0.6495**	2.6661	0.6196	0.9590	4.0504	0.5907	0.8305	3.5916
C-BPnn	1.0783	1.4473	5.8968	1.5525	2.1486	9.9900	1.4011	1.8842	8.3322
C-$\ell_{2,1}$RFelm	0.9548	1.3033	5.1095	1.3786	1.8823	8.8777	1.2266	1.6663	7.1065
C-ESn	1.0301	1.4782	5.6234	1.3923	1.9313	9.2527	1.2675	1.8130	7.4409
(b1) Combined system without ceemdan									
FIX	1.4641	2.0242	6.8942	1.8370	2.3831	8.2947	1.8483	2.5445	8.3929
MAX	14.7246	15.9534	63.3391	16.5771	17.7847	71.4600	18.1277	19.4904	80.8421
MIN	1.5737	2.2184	7.3353	1.8109	2.3946	8.2507	1.7003	2.3171	7.9920
MIX	10.6795	11.6622	45.6769	2.7319	3.5157	11.5748	6.3011	7.1793	27.4367
MRMR	1.4599	2.0244	6.8277	1.7541	2.2987	7.9835	1.8007	2.4644	8.3023
ReliefF	1.2977	1.7443	6.4876	1.4309	1.8380	6.3749	1.6705	2.3664	7.3116
LA	1.3942	2.3467	6.8619	1.6659	2.0776	7.9034	1.4789	2.2477	6.3263
mSSa	1.1232	1.5157	5.8992	1.2227	1.5390	6.4305	1.1772	1.6712	5.2612
(b2) Combined system includes cEEMDan									
C-FIX	0.7764	1.0859	3.7301	0.9307	1.1954	4.2834	0.9690	1.3614	4.4141
C-MAX	3.3711	3.7705	14.2808	9.8564	10.5824	42.5217	11.9701	12.8744	53.4489
C-MIN	0.7383	0.9943	3.4696	0.7709	1.0495	3.6179	1.2171	1.6056	5.6501
C-MIX	16.0585	17.3482	69.3189	11.8266	12.6870	51.0301	13.5027	14.5101	60.3564
C-MRMR	0.6598	0.9300	3.1344	0.7951	1.0481	3.6887	0.8203	1.1252	3.7768
C-ReliefF	0.7723	1.0531	3.8831	0.9448	1.1974	4.1829	0.8833	1.2485	3.9945
C-LA	0.5027	0.6874	2.4168	0.5837	0.8361	2.8366	0.5760	0.7900	2.7692
C-mSSa	**0.4776**	0.6526	**2.3576**	**0.5670**	**0.8101**	**2.7879**	**0.5642**	**0.7750**	**2.7404**

"C-" represents the forecasting models combined with cEEMDan.
The bold numbers indicate the optimal value of the indicators.

where $p_1^i(t+1)$ is the position of leader in the i-th chain at $(t+1)$-th iteration, $\mathbf{F}_i(t)$ represents the food source position in the i-th dimension at t-th iteration. $\underline{p}_1$ and $\overline{p_1}$ are the lower bound and the upper bound of p_1.

In **Eq. 14**, $\boldsymbol{\tau}_1$ is a parameter that controls the balance of exploration and exploitation, $\boldsymbol{\tau}_2$ is a random number in (0, 1) that determines the distance to move, and $\boldsymbol{\tau}_3$ is also a random number in (0, 1) that determines the direction of movement. The coefficient $\boldsymbol{\tau}_1$ is defined as $\boldsymbol{\tau}_1 = 2e^{-(4t/\boldsymbol{T}_{mSSa})^2}$, where t is the number of the current iteration and $\boldsymbol{T}_{mSSa}$ represents the number of maximum iteration. Whereafter, the positions of the followers are mathematically updated as $p_j^i(t+1) = \frac{1}{2}[p_j^i(t) + p_j^{i-1}(t)]$, $\forall 2 \le j,\ i = 1, 2, \ldots, N$. Finally, repeat the processes of calculating fitness, updating the archive, selecting food source and updating the salps location until satisfied with the end condition.

If the archive is not full, the non-dominated solutions are saved to the archive after comparison according to Definition 1, otherwise, before storage deletes some solutions (Mirjalili et al., 2017a). According to the principle of improving the distributivity of solutions in the archive, use the Roulette Wheel mechanism to remove the densest solutions. The probability of the solution being removed can be calculated as $\mathbf{P}_r = N_l/c$, where N_l is the number of l-th solution in the archive, and c is a constant greater than 1 (Mirjalili et al., 2017b).

According to the introduction of mSSa, the computational complexity of this method is $\mathcal{O}(O \times N + cof \times N + 2N^2)$ at one iteration, where cof is the computational complexity of the objective function. In this study, the Mean Square Error and the Standard deviation of the error are set as objective function. The complexity of the first objective functions is $\mathcal{O}(T^2)$ and the second objective function is $\mathcal{O}(T)$. Therefore, the complexity of one iteration of mSSa is $\mathcal{O}(O \times N + (T^2 + T) \times N + 2N^2)$, here T is the number of samples.

2.4 The Proposed Combined Forecasting System

Using the aforementioned methods and strategy, a novel combined pollutant concentration forecasting system based on the data decomposition strategy, several individual forecasting models, and a multi-objective optimization algorithm is designed.

Assume there are M models to predict the pollutant concentration, the forecasting results are denoted as



FIGURE 2 | The forecasting results of the different models, where "C-" represents the forecasting models combined with cEEMDan.

$\hat{Y}_m$, $(m = 1, 2, \ldots, M)$, and the weight coefficients of each forecasting result are ω_1, ω_2, $\ldots$, ω_M, then the combined system can be expressed in mathematical form as

$$\begin{cases} \mathbf{F} = \sum_{m=1}^{M} \omega_m \hat{Y}_m, \\ \sum_{m=1}^{M} \Omega_m = 1, \end{cases} \tag{15}$$

here $\mathbf{F}$ is the final forecasting result.

The main steps of this proposed system are listed as follows, and the flowchart of this study is described in **Figure 1**.

Pre-processing of original data. Since the original series are fluctuating, it is difficult to analyze its features. Therefore, the strategy of "decomposition and ensemble" is utilized to distinguish different characteristics and noise in the original series. And then, the noise is filtered out to reconstruct a more stable series. The parameters of this method are shown in **Table 5**.

Forecasting by individual models. Since the features hidden in the series are not certain, three types of methods were used to analyze the series and implement forecasting. These methods contain a traditional statistical model (ARima), a hesitant fuzzy time series forecasting model, and machine learning models (BPnn, $\ell_{2,1}$RFelm, ESn). In the three machine learning models, BPnn and $\ell_{2,1}$RFelm have the same network structure but different solving strategies, BPnn and ESn have different network structures but the same solving strategy, and $\ell_{2,1}$RFelm and ESn have different network structures and solving strategies.

Construction of the combined system. In order to obtain more accurate forecasting results, use mSSa to conduce the optimal combined weights of the individual models. More specifically, take the predicted values obtained by each individual model as input and the true concentration values as output to form a training set. Then, the optimization algorithm is trained based on this set and finally obtains the optimal weight vector. Afterward, the forecasting results of such individual models are combined together by using optimal weight to obtain the final forecasting value.

3 EMPIRICAL ANALYSIS

In this study, the concentration of $PM_{2.5}$ is forecast by the proposed combined system. This section mainly introduces the experimental process and analyzes the forecasting results.

3.1 Data Description

Three $PM_{2.5}$ concentration data sets collected from the Pearl River Delta (PRD) region in China are selected as illustrative examples to verify the effectiveness of the proposed combined prediction system, including Guangzhou (GZ), Shenzhen (SZ),

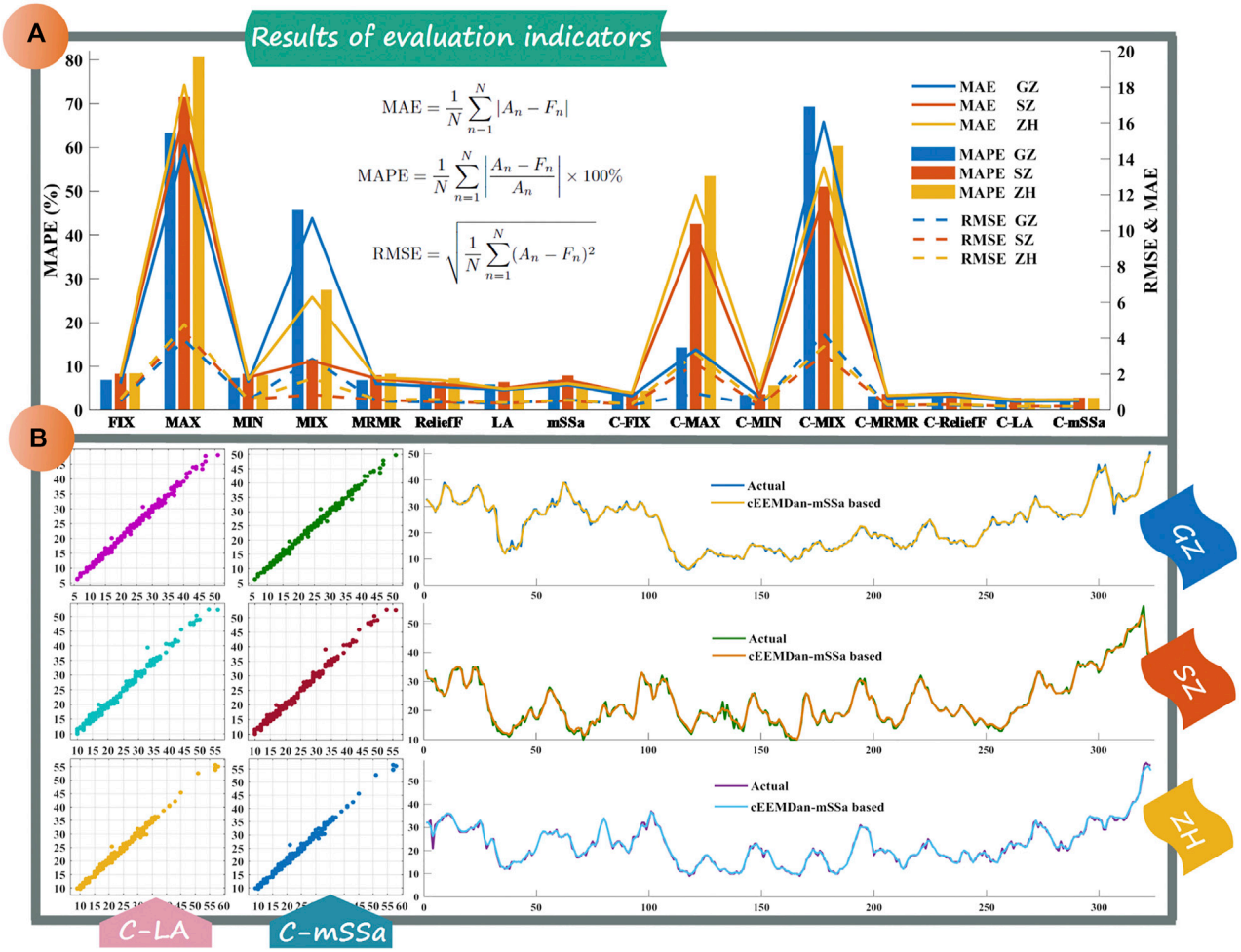


FIGURE 3 | The forecasting results of the different combined methods. **(A)** is the results of evaluation indicators of three study cities. **(B)** is the forecasting results of C-LA and C-mSSa, where "C-" represents the cEEMDan.

and Zhuhai (ZH). There are few missing data in these series, and the moving median method with a window length of 10 is used to fill in the missing data. Some statistical indicators for these three data sets are presented in **Table 4**. Considering the availability of data, the hourly concentrations were collected from 2020.04.29 to 2020.09.25, and these data were divided twice. In the forecasting module, the original data sets were divided into training sets and test sets, and the train to test ratio of each study city is Tr_1:Te_1 = 7 : 3. And in the combination module, Te_1 was divided into training set Tr_2 and test set Te_2, the division ratio is 7:3.

3.2 Evaluation Metrics

In previous studies, numerous metrics have been utilized to evaluate model performance. To scientifically assess the proposed system, three metrics are selected as evaluation criteria, including two scale-dependent indicators and a percentage indicator. Details are as follows.

3.2.1 Scale-dependent Indicators

The unit of this type of indicator is the same as the unit of original data, so it can not be used to compare two series with different units. Two commonly used scale-dependent measures are Mean absolute error and Root mean squared error, they are based on absolute errors and squared errors, respectively (Hyndman and Athanasopoulos, 2018).

A. Mean Absolute Error

The mean absolute error (MAE) is a commonly used indicator to evaluate the deviation between forecast values and true values (Khair et al., 2017):

$$\mathbf{MAE} = \frac{1}{N}\sum_{n-1}^{N}|A_n - F_n|, \tag{16}$$

where N is the sample size, A_n represents the actual value of n-th sample, and F_n indicates the n-th forecast value. This metric can avoid the cancellation of the positive and negative

predicted errors. The lower the value of MAE, the better the model is. MAE = 0 indicates that there is no error in the forecasting.

B. Root Mean Squared Error

The root mean squared error (RMSE) is a commonly used measure of the forecasting results of machine learning models. Its equation is shown in (**Eq. 17**) (Wang Y. et al., 2021)

$$\mathbf{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(A_n - F_n)^2}. \tag{17}$$

Same to the MAE, the lower the value of RMSE, the better the prediction. But RMSE is more sensitive to extreme values. Therefore, if the difference between RMSE and MAE is large, the greater the possibility of large errors existing in forecasting.

3.2.2 Percentage Indicator

The frequently used percentage indicator is the mean absolute percentage error (MAPE). It is often used in practice since it is a very intuitive explanation in terms of relative error and is unit-free. Its equation is shown as follows (Khair et al., 2017):

$$\mathbf{MAPE} = \frac{1}{N}\sum_{n=1}^{N}\frac{A_n - F_n}{A_n} \times 100\%. \tag{18}$$

Compared to MAE, this indicator is normalized by actual value, and useful when the size or size of a prediction variable is significant in evaluating the accuracy of forecasting (Khair et al., 2017). However, when there is 0 in the actual value, this indicator can not be used. MAPE = 0% indicates a perfect model, while MAPE = 100% indicates a poor model.

3.3 Parameter Settings

Different parameters of the model will lead to different results, so the analysis of the predicted results should be based on the parameters used. The model parameters used in this paper are shown in **Table 5**. For ARima, the optimal lag order, the optimal degree of difference, and the optimal order of the moving average are determined based on the Akaike Information Criterion (AIC). And all the empirical experiments are implemented on MATLAB R2020a, run on the Windows 10 professional operating system.

3.4 Experiments and Results Analysis

In this study, three comparisons are implemented based on the data from GZ, SZ, and ZH in China. The first comparison is implemented to verify the effectiveness of the data decomposition strategy, the second comparison compares the different combination methods, and the last comparison compares the individual forecasting methods with the combined forecasting system. The forecasting performance lists in **Table 6** and the specific results are analyzed as follows.

3.4.1 Comparison I

This comparison is set to compare the forecasting accuracy between the models combining the cEEMDan and models without combining cEEMDan. The comparisons are divided into two categories, one for individual models and one for the combined system. The first category contains comparisons of ARiam vs. C-ARima, BPnn vs. C-BPnn, $\ell_{2,1}$RFelm vs. C-$\ell_{2,1}$RFelm, and ESn vs. C-ESn. Here, the hesitant fuzzy time series forecasting method has fuzzed the original series and constructed a transition matrix based on the fuzzy logic relationship group to forecast pollution concentration. These operations have compressed and filtered the information of the original series, so the hesitant fuzzy time series forecasting experiment based on the composed data is no longer carried out. The second category contains comparisons of FIX vs. C-FIX, MAX vs. C-MAX, MIN vs. C-MIN, MRMR vs. C-MRMR, ReliefF vs. C-ReliefF, LA vs. C-LA, and mSSa vs. C-mSSa.

1) From the results in **Table 6** (a1) and (a2), it can be found that the forecasts based on the decomposed data are more accurate than based on the original data. Take the results from Guangzhou as an example. The maximum MAPE of the forecasts based on decomposed data ($\mathrm{MAPE}_{GZ}^{C-BPnn}$ = 5.896 8%) is lower than the minimum MAPE of the forecasts based on the original data ($\mathrm{MAPE}_{GZ}^{ARima}$ = 7.479 2%). And the MAE values of the forecasts based on the original data are all greater than 1.1 ($\mathbf{MAE_{GZ} > 1.1}$), but the MAE values of the forecasts based on the decomposed data are all less than 1.1 ($\mathrm{MAE}_{GZ}^{C} < \mathbf{1.1}$), especially the ($\mathrm{MAE}_{GZ}^{C-ARima} < \mathbf{0.5}$), which is the best performance among all the forecasting models. The value of RMSE also shows the same result. The values of RMSE for the forecasts based on the original data are all greater than the values of RMSE for the forecasts based on the decomposed data ($\mathrm{RMSE}_{GZ} > \mathrm{RMSE}_{GZ}^{C}$), which indicates that the forecasting values based on the decomposed data are closer to the true values. The sub-figures in **Figure 2** show the predicted results of these models.
2) The strategy of "decomposition and ensemble" to remove noise contributes to improving the forecasting accuracy. The figures in **Table 6** (b1) and (b2) show the forecasting results of combined systems. Take ZH as an example, the values of the indicators of the mSSa combination method are $(1.177\,2,\ 1.671\,2,\ 5.261\,2\%)_{\mathrm{MAE,\ RMSE,\ MAPE}}$. But, the results obtained by the proposed cEEMDan-mSSa based method are $(0.564\,2,\ 0.775\,0,\ 2.740\,4\%)_{\mathbf{MAE,\ RMSE,\ MAPE}}$, these three values are lower compared to the index results of mSSa based combined method. The same relationship can be found in the indicator results for GZ and SZ.

Then, by comparing the remaining figures, it can be found that the values of indicators for systems without combining data decomposition strategy are smaller than the values of combining data decomposition strategy except for the MIX combined method. Take GZ as an example, all the values of MAE are greater than 1 of the method without combining cEEMDan ($\mathbf{MAE}_{GZ} > 1$), but the values of these indicators are less than 1 for the method combining cEEMDan except for MAX and MIX combined methods ($\mathbf{MAE}_{GZ}^{C} < 1$). So as the values of **RMSE** and **MAPE**, the figures for the methods without combining cEEMDan are greater than the figures for methods combining cEEMDan. Therefore, it can be considered that no matter which combination method, the forecasting based on the decomposed data is more accurate.

Remark: Through the comparisons between the models combining the cEEMDan and models without combining cEEMDan, what can be found is that the data decomposition strategy can effectively improve the prediction ability of the model.

3.4.2 Comparison II

This comparison is set to compare the combination methods. These methods contain four numerical methods (FIX[1], MAX[2], MIN[3], MIX[4]), two feature selection methods (MRMR, ReliefF), and two optimization algorithms, the Lichtenberg algorithm (LA) and mSSa. The results in **Table 6** (b1) and (b2), and **Figure 3** demonstrate that after data decomposition, the forecasting accuracy is improved. Moreover, the proposed combined model performance is best. The detailed analyses are as follows.

1) The multi-objective optimization method is the best weighting method. For the results in **Table 6** (b1), it can be seen that the indicators' values of mSSa are the smallest. Take GZ as an example, the indicators' values of mSSa are $(1.123\,2, 1.515\,7, 5.899\,2\%)_{\mathbf{MAE,\ RMSE,\ MAPE}}$, the minimum indicators' values of the numerical methods are $\mathrm{MAE}_{GZ}^{FIX} = 1.464\,1$, $\mathrm{RMSE}_{GZ}^{FIX} = 2.024\,2$, $\mathrm{MAPE}_{GZ}^{FIX} = 6.894\,2$, and the minimum indicators' values of the feature selected methods are $\mathrm{MAE}_{GZ}^{ReliefF} = 1.297\,7$, $\mathrm{RMSE}_{GZ}^{ReliefF} = 1.744\,3$, $\mathrm{MAPE}_{GZ}^{ReliefF} = 6.487\,6$. Based on these indicators' values, it can be seen that the mSSa method has the best forecasting results. So as the results in SZ and ZH, the indicators' values obtained by mSSa method are smaller than the value of other methods.
2) Check the results in **Table 6** (b2), take SZ as an example, the MAE of numerical methods are $(0.9307, 9.8564, 0.7709, 11.8266)_{\mathbf{FIX,\ MAX,\ MIN,\ MIX}}$, MAE of feature selected methods are $(0.7951, 0.9448)_{\mathbf{MRMR,\ ReliefF}}$, and for the optimization methods are $(0.5837, 0.5670)_{\mathbf{LA,\ mSSa}}$. And $\min(\mathrm{MAE}) = \mathbf{MAE}_{SZ}^{C-mSSa} = 0.5670$. The same result can be obtained in GZ and ZH. Based on the results shown in the tables, it can be considered that the mSSa optimization algorithm is optimal as a weighting method.
3) The forecasts of the proposed combined system are more accurate than the mSSa based system. As the forecasting results shown in **Table 6** (b1) and (b2), the MAE values of the proposed combined system in the three study cities are $\mathbf{MAE}^{C-mSSa} = (0.477\,6, 0.567\,0, 0.564\,2)_{\mathbf{GZ,SZ,ZH}}$, these values are less than 0.6, but the MAE values of the system based on the original data are greater than 1.1 for three study cites ($\mathbf{MAE}^{mSSa} = (1.123\,2,\ 1.222\,7,\ 1.177\,2)_{\mathbf{GZ,SZ,ZH}}$). Moreover, the MAPE values of the proposed system are $\mathbf{MAPE}^{C-mSSa} = (2.357\,6\%,\ 2.787\,9\%,\ 2.740\,4\%)_{\mathbf{GZ,SZ,ZH}}$, compared to the mSSa-based system they are improved by $(60.04, 56.65, 47.91\%)_{\mathbf{GZ,SZ,ZH}}$[5]. Since the smaller the values of the three metrics, the better the forecasting. Therefore, the results of these metrics indicate that the proposed combined system is performing better than the other system. The same conclusion can be drawn from the values of RMSE.

Remark: The optimization algorithm combination methods are performing better than the other combination methods, especially better than the numerical combination methods. The weights determined by the numerical methods only consider part of the samples, so when the data fluctuates greatly, this type of method cannot get good forecasting results. And the weights determined by the feature selection methods and the optimization algorithms consider all the samples, including samples with large fluctuations, so the impact of large fluctuations can be reduced during the forecasting process.

3.4.3 Comparison III

This experiment compares the forecasting performance of the individual forecasting models and the combined forecasting system. The proposed forecasting system performs better than the individual forecasting models. Almost all the indicators' values in the **Table 6** (b1) and (b2) are smaller than those in the **Table 6** (a1) and (a2), except for the MAX combination method and MIX combination method. Based on the data of SZ, the $\mathbf{min(MAPE_{SZ})} = \mathbf{MAPE}_{SZ}^{C-ARima} = 4.0504$, but this value is still greater than the $\mathrm{MAPE}_{SZ}^{C-mSSa} = 2.7879\%$. The results of the other two metrics of SZ also show the same relationship. The $\min(\mathrm{MAE}_{SZ})$ and $\min(\mathrm{RMSE}_{SZ})$ are (0.5670, 0.8101), and all are obtained by the proposed forecasting system. These results indicate that the proposed combined forecasting system outperforms the individual forecasting models. The metric results of ZH can also draw the same conclusion as SZ. The results in GZ are a little different. The $\min(\mathrm{RMSE}_{GZ}) = \mathrm{RMSE}_{GZ}^{C-ARima} = 0.649\,5$, and $\mathrm{RMSE}_{GZ}^{C-mSSa} = 0.652\,6$, which is only 0.0031 different from the result of C-ARima. Therefore, the performance of the combined forecasting system can be regarded as better than the performance of the individual models.

In summary, the following conclusions can be drawn. The data decomposition strategy can significantly improve forecasting accuracy. These experimental results show that the forecasting results of all methods combined with cEEMDan, except MIX, are more accurate than the methods not combined with cEEMDan. In addition, the mSSa method has the best forecasting results among these combined methods, thus proving the forecasting performance of the proposed system is best.

Remark: For forecasting, data preprocessing is important. In this study, a powerful data decomposition strategy was used to decompose the original data series, and then discarded the noise component of the series. This processing improves the accuracy of the

[1]**FIX** represents a weighting method with fixed weights, and the weight of each forecasting model is 0.2.

[2]**MAX** represents the method of using the maximum forecasting error to assign weights, and the weight of each forecasting model is the reciprocal of the maximum forecasting error obtained by each model in the training set.

[3]**MIN** is opposite to MAX, using the minimum value of the forecasting error is used as the basis for weighting, the weight of wach methode is caluculated as $w_i = e_i / \sum_{i=1}^{5} e_i, e_i = 1/me_i, i = 1, \ldots, 5$, where me_i represents minimum error of i-th model.

[4]For **MIX** weighting method, the weight of each model is obtained by following equation: $w_i = mean(|e_{in}|/A_n)$, $i = 1, \ldots, 5$; $n = 1.\cdots, N$, here the e_{in} is forecasting errors of i-th model, A_n is the actual value of $PM_{2.5}$ concentration.

[5]The improved percentage is calculated as follows: $\mathbf{P}_{metric} = (V_m^{Model1} - V_m^{Model2})/V_m^{Model1}$. Such as the improved percentage of GZ's MAPE is $((5.899\,2 - 2.357\,6)/5.899\,2) \times 100\%$.

TABLE 7 | DM test results of different models.

	Based on the original data			—	Based on the decomposed data		
	GZ	SZ	ZH		GZ	SZ	ZH
(a) Forecasting models							
ARima	5.0977*	8.6243*	6.3830*	C-ARima	1.8092**	1.1831***	1.2560***
BPnn	6.8607*	10.4830*	8.5372*	C-BPnn	7.2128*	8.3443*	9.5224*
$\ell_{2,1}$RFelm	7.0826*	9.5058*	9.0820*	C-$\ell_{2,1}$RFelm	6.6667*	7.9443*	8.2492*
ESn	5.9861*	9.2922*	8.1848*	C-ESn	2.1994**	7.8597*	2.5456**
HFs	8.5816*	11.6504*	5.2527*	—	—	—	—
(b) Combined systems							
FIX	7.2134*	9.0465*	7.7204*	C-FIX	4.5435*	5.3059*	11.5207*
MAX	22.7947*	21.5754*	20.7482*	C-MAX	16.9403*	21.2405*	14.4119*
MIN	6.3787*	6.1052*	5.9078*	C-MIN	5.1935*	4.8392*	12.5494*
MIX	21.4592*	10.6720*	14.4818*	C-MIX	23.2960*	21.4463*	16.1001*
MRMR	6.7769*	9.0465*	8.4095*	C-MRMR	3.9308*	4.1445*	11.9718*
ReliefF	8.2291*	8.4603*	4.9277*	C-ReliefF	5.2545*	5.4203*	10.6837*
LA	2.2041**	9.7650*	5.5175*	C-LA	1.6563**	2.2618**	0.5763***
mSSa	6.7769*	8.6787*	5.8400*	C-mSSa	—	—	—

** indicates the 1% significance level $Z_{0.01/2} = 2.58$; ** indicates the 5% significance level $Z_{0.05/2} = 1.96$; *** indicates the 10% significance level $Z_{0.10/2} = 1.64$.*
"C-" represents the forecasting models combined with cEEMDan.
Indicates that the DM test has not been performed. Since HFs have compressed the original series, the forecasting based on the decomposed data has not been performed. And the C-mSSa is the system proposed in this paper, so the DM test has not been performed on itself.

TABLE 8 | Results of the model stability test.

	Based on the original data			—	Based on the decomposed data		
	GZ	SZ	ZH		GZ	SZ	ZH
(a) Forecasting models							
ARima	0.966 4	0.965 8	0.898 1	C-ARima	0.984 1	0.988 1	0.972 1
BPnn	0.990 2	0.984 1	0.991 5	C-BPnn	0.997 1	0.993 6	0.964 4
$\ell_{2,1}$RFelm	0.964 3	0.994 3	0.926 2	C-$\ell_{2,1}$RFelm	0.997 9	0.957 7	0.993 5
ESn	0.928 8	0.912 6	0.837 6	C-ESn	0.956 4	0.993 2	0.914 4
HFs	0.796 8	0.802 8	0.895 1	—	—	—	—
(b) Combined systems							
FIX	0.899 6	0.894 7	0.862 5	C-FIX	0.931 8	0.952 1	0.942 0
MAX	0.117 9	0.073 4	0.032 9	C-MAX	0.700 9	0.322 5	0.212 2
MIN	0.935 0	0.923 4	0.879 4	C-MIN	0.990 9	0.986 3	0.999 5
MIX	0.270 1	0.773 3	0.478 9	C-MIX	0.085 2	0.232 5	0.155 7
MRMR	0.925 4	0.917 9	0.875 5	C-MRMR	0.952 6	0.967 6	0.954 5
ReliefF	0.830 0	0.839 1	0.852 7	C-ReliefF	0.897 2	0.897 2	0.934 8
LA	0.837 4	0.859 9	0.939 3	C-LA	0.938 9	0.975 6	0.879 4
mSSa	0.946 0	0.992 1	0.990 6	C-mSSa	0.986 1	0.998 6	0.985 9

"C-" represents the forecasting models combined with cEEMDan, that is the forecasting models based on the decomposed data.
Indicates that the stability test has not been performed. Since HFs have compressed the original series, the forecasting based on the decomposed data has not been performed. And the C-mSSa is the system proposed in this paper, so the stability test has not been performed on itself.

forecasting, and this conclusion is reached in two experiments. For combination, the multi-objective optimization method works better, and the numerical methods are the worst, and the performance is unstable. When the results of other methods become better, the numerical method performs worse.

4 TEST OF FORECASTING SYSTEM

In order to verify the significance and stability of the proposed forecasting system, the Diebold-Mariano test (DM) (Francis and Roberto, 1995) and the variance ratio (VR) are introduced in this study. The related details and results are described in this section.

4.1 Diebold-Mariano Test

DM is a hypothesis testing method to analyze the difference in prediction accuracy. According to the constructed DM statistics, it can be judged whether the difference of the prediction method is significant. In this test, the null hypothesis (H_0) and the alternative hypothesis (H_1) are as follows:

$$\begin{aligned} \boldsymbol{H_0}&: \bar{\bar{\boldsymbol{E}}}\left[\tilde{\boldsymbol{\mathcal{L}}}\left(\breve{\boldsymbol{\delta}}_{\boldsymbol{1}}^{\boldsymbol{t}}\right)\right] = \bar{\bar{\boldsymbol{E}}}\left[\tilde{\boldsymbol{\mathcal{L}}}\left(\breve{\boldsymbol{\delta}}_{\boldsymbol{2}}^{\boldsymbol{t}}\right)\right] \\ \boldsymbol{H_1}&: \bar{\bar{\boldsymbol{E}}}\left[\tilde{\boldsymbol{\mathcal{L}}}\left(\breve{\boldsymbol{\delta}}_{\boldsymbol{1}}^{\boldsymbol{t}}\right)\right] \neq \bar{\bar{\boldsymbol{E}}}\left[\tilde{\boldsymbol{\mathcal{L}}}\left(\breve{\boldsymbol{\delta}}_{\boldsymbol{2}}^{\boldsymbol{t}}\right)\right] \end{aligned} \tag{19}$$

here $\breve{\delta}_1^t$ and $\breve{\delta}_2^t$ represent the forecasting errors of forecasting model 1 and forecasting model 2 at t-th, $\tilde{\mathcal{L}}(\cdot)$ represents the loss function. Then, the DM statistic is constructed as follows (Huang et al., 2021):

$$\mathbf{DM} = \frac{\sum_{t=1}^{n}[\tilde{\mathcal{L}}(\breve{\delta}_1^t) - \tilde{\mathcal{L}}(\breve{\delta}_2^t)]\Big/ n}{\sqrt{S^2/n}} \tag{20}$$

where S^2 denotes the variance estimation of $\breve{\delta}_1^t - \breve{\delta}_2^t$.

Given a certain significance level α, the critical value $Z_{\alpha/2}$ can get, if the absolute value of DM statistic is greater than the $Z_{\alpha/2}$, the null hypothesis H_0 is rejected, and the result that two forecasting methods have significant differences.

Table 7 gives the DM test results of different forecasting models. This study compares 24 forecasting models or systems with the proposed system. Compared with the forecasting model without cEEMDan, the proposed forecasting system is significantly better, since the values of DM statistic are greater than the critical value of 1% significance level. After combined with cEEMDan, the forecasting ability of individual forecasting models has been improved, but the DM test results show that their predictive ability is still inferior to the proposed forecasting system, since the lowest value of DM test is between the critical value of 10% significance level and the critical value of 15% significance level. The DM values of **Table 7 (b)** also show that the proposed forecasting system is significantly superior than the other combined forecasting system, especially the system without data decomposition strategy.

4.1.1 Stability Test

In order to validate the stability of models, the variance ratio **(Vr)** is introduced. **Vr** combines the variances of the forecasting value and the true value to illustrate the stability of the forecasting model. The greater the value of Vr, the higher the forecasting stability of the method (Huang et al., 2021).

$$Vr = min(Var_{forecasting}\Big/ Var_{actual}, Var_{actual}\Big/ Var_{forecasting}), \tag{21}$$

here, $Var_{forecasting}$ and Var_{actual} are the variances of the forecasting values and actual values.

The Vr results are shown in **Table 8**. The Vr values of the proposed system in the three cities are $(0.986\,1,\ 0.998\,6,\ 0.985\,9)_{GZ,SZ,ZH}$. Although the Vr values of the proposed system are not the largest among all forecasting models and systems, these three values are all greater than 0.98, while the Vr values of most other forecasting models and systems are less than 0.98, indicating that the proposed forecasting system is relatively stable. Combined with the results of the forecasting evaluation metric shown in **section 3**, it shows that the proposed forecasting system has high prediction accuracy and relatively high stability.

5 CONCLUSION

Based on the multi-objective optimization algorithm and data decomposition strategy, an effective combined forecasting system is proposed to forecast the $PM_{2.5}$ concentration from Guangzhou, Shenzhen, and Zhuhai in China. The proposed system mainly contains three modules, the data preprocessing module, the individual model forecasting module, and the combination forecasting module. In the first module, the strategy of "decomposition and ensemble" is applied to remove the noise in the original series. In the individual model forecasting module, ARima, BPnn, $\ell_{2,1}$RFelm, ESn, and HFs are applied to forecast $PM_{2.5}$ concentration respectively. These five models are from different kinds of forecasting models and are used to analyze different features in the $PM_{2.5}$ concentration series. ARima is a classical traditional statistical forecasting method; BPnn, $\ell_{2,1}$RFelm, and ESn are neural networks with different characteristics; hesitant fuzzy time series model is a fuzzy-based forecasting model. By comparing eight weighting methods from three categories, the best combination method is found as a multi-objective optimization weighting method.

The developed combined forecasting system has been successfully applied in $PM_{2.5}$ concentration forecasting. Based on the forecasting evaluation indicators, the forecasting performance of the proposed system is validated. Specifically, compared the models forecasting results based on data before and after the preprocessing of cEEMDan in Comparison I. In Comparison II, compare the system employing diverse combination methods. Compere between the individual models and the combined models in Comparison III. After these comparative experiments, it can be observed that the MAE and MAPE values of the proposed system are always lower than the values of individual models and other combination methods. For RMSE in Guangzhou, the value of the proposed system is slightly higher than the minimum RMSE value, but overall, the forecasting performance of the proposed system is still the best. Therefore, the proposed combined forecasting system, which combines different types of individual forecasting models, has high practical application potential in air pollution concentration forecasting.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.mep.gov.cn/.

AUTHOR CONTRIBUTIONS

LB: Conceptualization, Methodology, Software, Writing-Original draft preparation. HL: Conceptualization, Software, Validation. BZ: Conceptualization, Supervision, Writing-Reviewing and Editing. XH: Conceptualization, Writing—Review and Editing, Data curation.

FUNDING

This work was supported by Major Program of National Social Science Foundation of China (Grant No. 17ZDA093).

REFERENCES

Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. (2014). "Stock price Prediction Using the Arima Model," in 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation (Cambridge, UK: IEEE), 106–112. doi:10.1109/uksim.2014.67

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., and Ciccozzi, M. (2020). Application of the Arima Model on the Covid-2019 Epidemic Dataset. *Data in brief* 29, 105340. doi:10.1016/j.dib.2020.105340

Bisht, K., and Kumar, S. (2016). Fuzzy Time Series Forecasting Method Based on Hesitant Fuzzy Sets. *Expert Syst. Appl.* 64, 557–568. doi:10.1016/j.eswa.2016.07.044

Cheng, S.-H., Chen, S.-M., and Jian, W.-S. (2016). Fuzzy Time Series Forecasting Based on Fuzzy Logical Relationships and Similarity Measures. *Inf. Sci.* 327, 272–287. doi:10.1016/j.ins.2015.08.024

Cheng, X., Liu, Y., Xu, X., You, W., Zang, Z., Gao, L., et al. (2019). Lidar Data Assimilation Method Based on Crtm and Wrf-Chem Models and its Application in pm2.5 Forecasts in Beijing. *Sci. Total Environ.* 682, 541–552. doi:10.1016/j.scitotenv.2019.05.186

Dėdelė, A., and Miškinytė, A. (2019). Seasonal and Site-specific Variation in Particulate Matter Pollution in lithuania. *Atmos. Pollut. Res.* 10, 768–775. doi:10.1016/j.apr.2018.12.004

Francis, X., and Roberto, S. (1995). Comparing Predictive Accuracy. *J. Business Econ. Stat.* 13, 134. doi:10.1080/07350015.1995.10524599

Gavirangaswamy, V. B., Gupta, G., Gupta, A., and Agrawal, R. (2013). "Assessment of Arima-Based Prediction Techniques for Road-Traffic Volume," in Proceedings of the fifth international conference on management of emergent digital EcoSystems, 246–251. doi:10.1145/2536146.2536176

Glencross, D. A., Ho, T.-R., Camiña, N., Hawrylowicz, C. M., and Pfeffer, P. E. (2020). Air Pollution and its Effects on the Immune System. *Free Radic. Biol. Med.* 151, 56–68. doi:10.1016/j.freeradbiomed.2020.01.179

Goyal, P., Chan, A. T., and Jaiswal, N. (2006). Statistical Models for the Prediction of Respirable Suspended Particulate Matter in Urban Cities. *Atmos. Environ.* 40, 2068–2077. doi:10.1016/j.atmosenv.2005.11.041

Grennfelt, P., Engleryd, A., Forsius, M., Hov, Ø., Rodhe, H., and Cowling, E. (2020). Acid Rain and Air Pollution: 50 Years of Progress in Environmental Science and Policy. *Ambio* 49, 849–864. doi:10.1007/s13280-019-01244-4

Gündüz, D., de Kerret, P., Sidiropoulos, N. D., Gesbert, D., Murthy, C. R., and van der Schaar, M. (2019). Machine Learning in the Air. *IEEE J. Selected Areas Commun.* 37, 2184–2199.

Haimes, Y. Y., Hall, W. A., and Freedman, H. T. (2011). *Multiobjective Optimization in Water Resources Systems: The Surrogate worth Trade-Off Method*. Amsterdam, Netherlands: Elsevier.

Hecht-Nielsen, R. (1992). "Theory of the Backpropagation Neural Network**Based on "nonindent" by Robert Hecht-Nielsen, Which Appeared in Proceedings of the International Joint Conference on Neural Networks 1, 593-611, June 1989. 1989 IEEE," in *Neural Networks for Perception* (Amsterdam, Netherlands: Elsevier), 65–93. doi:10.1016/b978-0-12-741252-8.50010-8

Henrique, B. M., Sobreiro, V. A., and Kimura, H. (2019). Literature Review: Machine Learning Techniques Applied to Financial Market Prediction. *Expert Syst. Appl.* 124, 226–251. doi:10.1016/j.eswa.2019.01.012

Huang, X., Wang, J., and Huang, B. (2021). Two Novel Hybrid Linear and Nonlinear Models for Wind Speed Forecasting. *Energ. Convers. Manage.* 238, 114162. doi:10.1016/j.enconman.2021.114162

Hyndman, R. J., and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts.

Jaeger, H., and Haas, H. (2004). Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *science* 304, 78–80. doi:10.1126/science.1091277

Khair, U., Fahmi, H., Hakim, S. A., and Rahim, R. (2017). Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error. *J. Phys. Conf. Ser.* 930, 012002. doi:10.1088/1742-6596/930/1/012002

Kurt, A., Gulbagci, B., Karaca, F., and Alagha, O. (2008). An Online Air Pollution Forecasting System Using Neural Networks. *Environ. Int.* 34, 592–598. doi:10.1016/j.envint.2007.12.020

Liu, B., Yu, X., Chen, J., and Wang, Q. (2021). Air Pollution Concentration Forecasting Based on Wavelet Transform and Combined Weighting Forecasting Model. *Atmos. Pollut. Res.* 12, 101144. doi:10.1016/j.apr.2021.101144

Liu, H., Xu, Y., and Chen, C. (2019). Improved Pollution Forecasting Hybrid Algorithms Based on the Ensemble Method. *Appl. Math. Model.* 73, 473–486. doi:10.1016/j.apm.2019.04.032

Lu, W., Chen, X., Pedrycz, W., Liu, X., and Yang, J. (2015). Using Interval Information Granules to Improve Forecasting in Fuzzy Time Series. *Int. J. Approximate Reasoning* 57, 1–18. doi:10.1016/j.ijar.2014.11.002

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., and Bezirtzoglou, E. (2020). Environmental and Health Impacts of Air Pollution: a Review. *Front. Public Health* 8, 14. doi:10.3389/fpubh.2020.00014

Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., and Mirjalili, S. M. (2017a). Salp Swarm Algorithm: A Bio-Inspired Optimizer for Engineering Design Problems. *Adv. Eng. Softw.* 114, 163–191. doi:10.1016/j.advengsoft.2017.07.002

Mirjalili, S., Jangir, P., and Saremi, S. (2017b). Multi-objective Ant Lion Optimizer: a Multi-Objective Optimization Algorithm for Solving Engineering Problems. *Appl. Intell.* 46, 79–95. doi:10.1007/s10489-016-0825-8

Mousavi, S. S., Goudarzi, G., Sabzalipour, S., Rouzbahani, M. M., and Mobarak Hassan, E. (2021). An Evaluation of Co, Co2, and So2 Emissions during Continuous and Non-continuous Operation in a Gas Refinery Using the Aermod. *Environ. Sci. Pollut. Res.* 28, 56996–57008. doi:10.1007/s11356-021-14493-2

Ngatchou, P., Zarei, A., and El-Sharkawi, A. (2005). "Pareto Multi Objective Optimization," in Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems (Arlington, VA, USA: IEEE), 84–91.

Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., and Kolehmainen, M. (2004). Evolving the Neural Network Model for Forecasting Air Pollution Time Series. *Eng. Appl. Artif. Intelligence* 17, 159–167. doi:10.1016/j.engappai.2004.02.002

Niu, X., and Wang, J. (2019). A Combined Model Based on Data Preprocessing Strategy and Multi-Objective Optimization Algorithm for Short-Term Wind Speed Forecasting. *Appl. Energ.* 241, 519–539. doi:10.1016/j.apenergy.2019.03.097

Organization, W. H. (2014). *7 Million Premature Deaths Annually Linked to Air Pollution*. Geneve, Switzerland: WHO.

Pai, P.-F., and Lin, C.-S. (2005). A Hybrid Arima and Support Vector Machines Model in Stock price Forecasting. *Omega* 33, 497–505. doi:10.1016/j.omega.2004.07.024

Qiao, J., Li, F., Han, H., and Li, W. (2016). Growing echo-state Network with Multiple Subreservoirs. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 391–404. doi:10.1109/TNNLS.2016.2514275

Rahimi, A., and Recht, B. (2007). "Random Features for Large-Scale Kernel Machines," in Proceedings of the 20th International Conference on Neural Information Processing Systems (New York: Curran Associates Inc.), 1177–1184.

Singh, S. R. (2007). A Simple Method of Forecasting Based on Fuzzy Time Series. *Appl. Math. Comput.* 186, 330–339. doi:10.1016/j.amc.2006.07.128

Song, Q., and Chissom, B. S. (1993). Fuzzy Time Series and its Models. *Fuzzy sets Syst.* 54, 269–277. doi:10.1016/0165-0114(93)90372-o

Torra, V., and Narukawa, Y. (2009). "On Hesitant Fuzzy Sets and Decision," in 2009 IEEE International Conference on Fuzzy Systems (Jeju, South Korea: IEEE), 1378–1382. doi:10.1109/fuzzy.2009.5276884

Torres, M. E., Colominas, M. A., Schlotthauer, G., and Flandrin, P. (2011). "A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise," in 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP) (Prague, Czech Republic: IEEE), 4144–4147. doi:10.1109/icassp.2011.5947265

Volk, M. J., Lourentzou, I., Mishra, S., Vo, L. T., Zhai, C., and Zhao, H. (2020). Biosystems Design by Machine Learning. *ACS Synth. Biol.* 9, 1514–1533. doi:10.1021/acssynbio.0c00129

Wang, J., Du, P., Hao, Y., Ma, X., Niu, T., and Yang, W. (2020a). An Innovative Hybrid Model Based on Outlier Detection and Correction Algorithm and Heuristic Intelligent Optimization Algorithm for Daily Air Quality index Forecasting. *J. Environ. Manag.* 255, 109855. doi:10.1016/j.jenvman.2019.109855

Wang, J., Li, H., Wang, Y., and Lu, H. (2021a). A Hesitant Fuzzy Wind Speed Forecasting System with Novel Defuzzification Method and Multi-Objective

Optimization Algorithm. *Expert Syst. Appl.* 168, 114364. doi:10.1016/j.eswa.2020.114364

Wang, J., Li, Q., and Zeng, B. (2021b). Multi-layer Cooperative Combined Forecasting System for Short-Term Wind Speed Forecasting. *Sustainable Energ. Tech. Assessments* 43, 100946. doi:10.1016/j.seta.2020.100946

Wang, J., Niu, T., Lu, H., Yang, W., and Du, P. (2019). A Novel Framework of Reservoir Computing for Deterministic and Probabilistic Wind Power Forecasting. *IEEE Trans. Sustain. Energ.* 11, 337–349.

Wang, J., Wang, Y., Li, Z., Li, H., and Yang, H. (2020b). A Combined Framework Based on Data Preprocessing, Neural Networks and Multi-Tracker Optimizer for Wind Speed Prediction. *Sustain. Energ. Tech. Assessments* 40, 100757. doi:10.1016/j.seta.2020.100757

Wang, J., Yang, W., Du, P., and Niu, T. (2020c). Outlier-robust Hybrid Electricity price Forecasting Model for Electricity Market Management. *J. Clean. Prod.* 249, 119318. doi:10.1016/j.jclepro.2019.119318

Wang, Y.-H., Yeh, C.-H., Young, H.-W. V., Hu, K., and Lo, M.-T. (2014). On the Computational Complexity of the Empirical Mode Decomposition Algorithm. *Physica A: Stat. Mech. its Appl.* 400, 159–167. doi:10.1016/j.physa.2014.01.020

Wang, Y., Wang, J., Li, Z., Yang, H., and Li, H. (2021c). Design of a Combined System Based on Two-Stage Data Preprocessing and Multi-Objective Optimization for Wind Speed Prediction. *Energy* 231, 121125. doi:10.1016/j.energy.2021.121125

Yang, H., Zhu, Z., Li, C., and Li, R. (2020). A Novel Combined Forecasting System for Air Pollutants Concentration Based on Fuzzy Theory and Optimization of Aggregation Weight. *Appl. Soft Comput.* 87, 105972. doi:10.1016/j.asoc.2019.105972

Zadeh, L. A. (1996). "Fuzzy Sets," in *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers by Lotfi A Zadeh* (Singapore: World Scientific), 394–432. doi:10.1142/9789814261302_0021

Zhang, L., Lin, J., Qiu, R., Hu, X., Zhang, H., Chen, Q., et al. (2018). Trend Analysis and Forecast of pm2.5 in Fuzhou, china Using the Arima Model. *Ecol. Indicators* 95, 702–710. doi:10.1016/j.ecolind.2018.08.032

Zhou, S., Liu, X., Liu, Q., Wang, S., Zhu, C., and Yin, J. (2016). Random Fourier Extreme Learning Machine with $\ell_{2,1}$-Norm Regularization. *Neurocomputing* 174, 143–153. doi:10.1016/j.neucom.2015.03.113

Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., and Che, J. (2017). Daily Air Quality index Forecasting with Hybrid Models: A Case in china. *Environ. Pollut.* 231, 1232–1244. doi:10.1016/j.envpol.2017.08.069

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.

Copyright © 2022 Bai, Li, Zeng and Huang. 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 in Environmental Science

ORIGINAL RESEARCH
published: 15 March 2022
doi: 10.3389/fenvs.2022.854150

A Quantitative Assessment on Ecological Compensation Based on Water Resources Value Accounting: A Case Study of Water Source Area of the Middle Route of South-To-North Water Transfer Project in China

Junfei Chen[1,2,3]*, *Qian Wang*[1] and *Qian Li*[1]

[1]Business School, Hohai University, Nanjing, China, [2]Yangtze Institute for Conservation and Development, Hohai University, Nanjing, China, [3]Jiangsu Research Base of Yangtze Institute for Conservation and High-Quality Development, Nanjing, China

OPEN ACCESS

Edited by:
Yiliao Song,
University of Technology Sydney, Australia

Reviewed by:
Grigorios L. Kyriakopoulos,
National Technical University of Athens, Greece
Haseeb Hassan,
Shenzhen University, China

***Correspondence:**
Junfei Chen
chenjunfei@hhu.edu.cn

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: *13 January 2022*
Accepted: *21 February 2022*
Published: *15 March 2022*

Citation:
Chen J, Wang Q and Li Q (2022) A Quantitative Assessment on Ecological Compensation Based on Water Resources Value Accounting: A Case Study of Water Source Area of the Middle Route of South-To-North Water Transfer Project in China. Front. Environ. Sci. 10:854150. doi: 10.3389/fenvs.2022.854150

As an economic means to adjust the contradiction between ecology and development, ecological compensation plays an important role in promoting the good operation of interbasin water transfer projects and the sustainable development of regional economy. The accounting of ecological compensation is the key and difficult point of ecological compensation as well as the basis of ecological compensation policy and practice. Watershed ecological compensation based on water resources value accounting is an early exploration field of ecological compensation research, and water resources value calculation needs to consider both water quantity and water quality comprehensively. Taking the water source area of the Middle Route of South-to-North Water Transfer Project (SNWTP) as an example, this article tries to establish the payment standard of watershed ecological compensation from the perspective of water resources value. The results show that: 1) The water resources value of the six core regions in the water source area has shown an overall upward trend since 2000, and the northern regions have demonstrated higher value of water resources than the southern regions. 2) The LSTM neural network model is used to forecast the value of water resources in the six regions from 2020 to 2022, and it is found that the value of water resources would increase in the next few years. 3) The compensation price of the six regions in the water source area is predicted in the range of 0.5–1.5 yuan/m^3 from 2020 to 2022, and an upward trend in the ecological compensation amount is forecast. Based on the above conclusions, this article puts forward suggestions to establish an ecological compensation accounting system in line with the Middle Route water source area from the perspectives of ecological compensation legislation, allocation of ecological compensation amount, and introduction of market mechanism.

Keywords: water resources value, ecological compensation accounting, the middle route of water transfer, machine learning, China

1 INTRODUCTION

Ecological compensation is an important economic measure to ensure the sustainability of water quality and quantity in the interbasin water transfer project, which is beneficial to realize the efficiency and equity in the process of water supply and consumption in water source area and water-receiving area. One of the key points of ecological compensation lies in the accounting of ecological compensation. Different accounting methods produce different implementation effects, which have very important practical significance for the sustainability of water transfer projects and economic development of water source area. The existing studies on ecological compensation accounting methods mainly are carried out from the perspectives of ecosystem service value, beneficiary's profit, conservator's input and loss, payment, and compensation willingness (Xepapadeas, 2011; Hu, 2016). The common compensation accounting can be divided into protection compensation and cross-basin compensation (Wang and Hou, 2013). So far, compensation accounting methods in water source area have included ecosystem service value method (Zhang et al., 2002), water quality accounting method (Xu et al., 2008), water resources value method and willingness to pay method (Costanza, 2012), water ecological restoration cost method (Martinez-Paz et al., 2013), etc. And the ecosystem service value method is the most important accounting method. In the 1970s, the value assessment of ecosystem services begins to receive more attention from different countries and regions. Value quantity and physical quantity are mostly used to measure the value of ecosystem services internationally. Pimentel et al. (1995) proposed two methods (i.e., ecosystem optimal model and maintenance of biological balance model) to estimate the functional value of global biodiversity based on willingness to pay. Ouyang and Wang, 2000a; Ouyang et al., 2000b) divided the value of ecosystem services into four categories, namely direct use-value, indirect use-value, selection value, and heritage value, and calculated the value of six terrestrial ecological services. In the study of global environment and natural services, ecosystem service value is divided into use-value and non-use-value (Rawlins et al., 2018). Specifically, use-value includes direct use-value, choice use-value, and indirect use-value, while non-use-value is divided into existence value and heritage value (Gomez-Baggethun et al., 2010). By simulating the flow path and quantity of water supply services, Xu et al. (2019) provided a scientific theoretical basis for ecosystem service management and ecological compensation. Ma and Wang (1984) studied the complex ecosystem of Society-Economy-Nature, which marked ecologists' march into the field of economics. Subsequently, empirical studies based on specific regions and basins began to emerge. Zhang et al. (2001) studied the service value evaluation method and divided it into three categories, namely actual market, alternative market, and simulated market. Zi (2010) evaluated the economic value of ecosystem services of water tourism resources in Heilongjiang Province, and believed that reasonable water tourism resources planning should be made with emphasis on the protection of wetlands, rivers, lakes, and other water bodies. Based on the existing achievements in this field, Xie (2012) compiled a table of equivalent factors of ecosystem service value in China. Based on the Research Report of the United Nations Environment Programme (UNEP) and combined with the theory of ecological economics, Ouyang et al. (2013) obtained the ecosystem service value evaluation methods of alternative market and simulated market. According to the existing problems in the ecological environment of the Chishui River, Qiu and Zhai (2014) established the ecological compensation mechanism. Based on the perspective of agricultural development, Hu (2015) concluded that the factors affecting the ecological compensation mechanism mainly include the loss of development opportunities, environmental protection cost, and public awareness. Yang et al. (2020) analyzed the current situation of ecological compensation in the Yellow River basin, and described the framework of water-related ecological compensation in the Yellow River Basin in detail. Xu et al. (2021) constructed a universal framework for interregional ecological compensation on the basis of considering differences among regional development.

The value of water resources was first discussed in the 1970s, and it was expressed as the maximum payment amount willing to transfer a unit of water. In the late 1980s, with the deepening of water resources crisis, water resources price has caused an upsurge of research by scholars. Lyman (1992) found that the maximum price of water resources fluctuated twice as much as the non-maximum price. Jiang (1998a) discussed the value basis of water resources from different perspectives, and his most significant contribution was to complete the calculation of the water resources value with the method of fuzzy mathematics. Wang and Qu (2001) incorporated the water resources indexes into the national economic accounting system. Piper (2003) explored the relationship between water quality and cost, and proposed the mechanism of water price affecting water quality. Wang et al. (2003) and Shen (2006) put forward the concept of "triple water price" and tried to use the general equilibrium model to calculate the actual water price in China. Subsequently, the value accounting of water resources began to cross departments and regions, and gradually tended to be market-oriented. Zhao et al. (2007) established a theoretical model of emission rights trading market based on experimental methods, and proposed the system construction scheme of resources trading market with optimal efficiency. Wei et al. (2008) analyzed the different profit levels of water conservancy departments under different water prices, so as to explore the maximization of income under the condition of the lowest water price. Du (2015) explored the improvement scheme of the real water price setting mechanism in view of the problems existing in the current water price scheme of SNWDP in China. Jia et al. (2018) used the improved fuzzy comprehensive evaluation model to calculate the maximum affordable water price of residents. Taking Taipu River as an example, Yang et al. (2019) adopted the game theory method to study the incentive policy, ecological compensation, and water quality accounting, and explored the establishment of trans-regional cooperation mechanism of water resources.

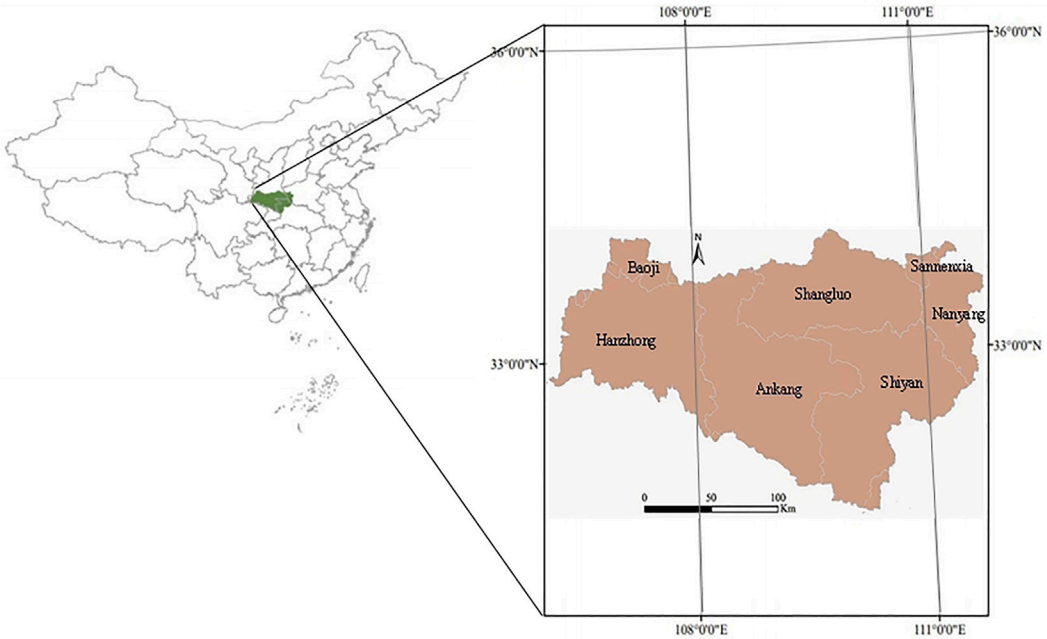


FIGURE 1 | Water source district administrative zoning map. Note: The map was generated by ArcGIS 10.5. URL link: https://www.esri.com/en-us/arcgis/products/districting-for-arcgis/overview.

To sum up, there are many studies on ecological compensation accounting and value of water resources, but few ones connect the two to discuss, resulting in the lack of effective connection between the existing ecosystem service value accounting methods and water resources social-economic pricing. Considering that it is more easy to accept the cost (i.e., value of water resources) paid for protecting water resources as the compensation standard by both the water source area and the water receiving area, the paper takes the water source area of the Middle Route of SNWDP in China as the research object, and studies the ecological compensation accounting from the perspective of water resources value accounting, which could promote the research on the ecological compensation accounting method of water source area. The novelty of this study is that the machine learning method (i.e., K-means clustering model and LSTM model) in computer science is applied to the evaluation of water resources value in the field of ecosystem. It gives full play to the advantages of machine learning method in dealing with complex nonlinear social science topics and makes the research results more consistent with the reality.

2 MATERIALS AND METHODS

2.1 Study Area

The water source area of the Middle Route of the SNWDP mainly refers to the basin area above the Danjiangkou Reservoir, which is located at the junction of Shaanxi, Henan, and Hubei provinces. The geographical coordinates of the water source area boundary are between 106°30′-112°18′E and 31°20′-34°10′N. The water source area of the middle route covers a total area of 130,906 square kilometers, with a total population of 24.26 million. The core cities in the water source area mainly include Shiyan of Hubei Province, Hanzhong, Ankang and Shangluo of Shaanxi Province, Nanyang and Sanmenxia of Henan Province (**Figure 1**). This area belongs to the north subtropical subhumid monsoon climate zone, with uneven distribution of precipitation. The average annual temperature is 16°C, and the annual average precipitation is about 800 mm. Han River is the main river, crossing the whole territory from west to east. The river water system in this area is relatively abundant, and large-scale water conservancy projects such as Xi River Reservoir and Shimen Reservoir have been built.

To protect the ecological environment of the water source area, all polluting enterprises and mines on both sides of the water source area have been closed, and garbage treatment and sewage purification plants in some cities with relatively concentrated populations have been set up by the local government. An area of 7,681 square kilometers has been brought under water and soil erosion control, more than 20,000 ha of basic farmland have been newly built and 270,000 ha of afforestation have been planted, which not only effectively controlled water and soil loss, but also promoted local economic development. In addition, the government of the water source region has actively implemented the natural forest protection project and stopped logging of natural forest commodities. Sixteen state-owned forest farms in 199 townships have been included in the project to protect natural forests, and provincial-level protected forests such as Hualongshan and Yinghu Wetland have been established. In these protected forests, the animal and plant communities are rich and complete, and the ecological environment is good as a whole. The drainage area is large and the rainfall is abundant;

TABLE 1 | Evaluation index system of water resources value.

Dimension	Indicator	Description	Attribute
Nature	X_1: Surface water resources (10×10^7 m^3)	Extract the data of surface water resources	–
	X_2: Groundwater resources (10×10^7 m^3)	Extract the data of groundwater resources	–
	X_3: Annual precipitation (10×10^7 m^3)	Collect rainfall monitoring data from national meteorological stations	–
	X_4: Water quality	Collect water quality monitoring data	+
	X_5: Total soil erosion (t)	Calculate the total amount of soil erosion per unit area per year	+
	X_6: Carbon fixation and oxygen release (t)	Calculate the carbon retention per unit area per unit time	+
	X_7: Surface runoff coefficient	Calculate the ratio of surface runoff to rainfall	+
Economy	X_8: Per capita GDP (yuan)	Calculate the ratio of GDP to population	–
	X_9: Per capita income of residents (yuan)	Extract the data of per capita income	–
	X_{10}: Per capita consumption expenditure (yuan)	Extract the per capita expenditure data	–
	X_{11}: Per capita urban maintenance and construction funds (yuan)	Extract the per capita urban construction cost	+
	X_{12}: The registered urban unemployment rate (%)	Extract the unemployment rate data	+
	X_{13}: Water consumption per 10,000 yuan GDP (m^3)	Extract the water consumption of 10,000 yuan GDP	–
	X_{14}: Utilization ratio of water resources (%)	Calculate the ratio of water supply to total water resources	–
Society	X_{15}: Population density (people/km^2)	Extract the population density data	+
	X_{16}: Urban area green coverage rate (%)	Extract the greening coverage rate data	+
	X_{17}: Ratio of students in ordinary colleges and universities (%)	Calculate the ratio of the number of students in colleges and universities to total population	–
	X_{18}: Number of beds in public health institutions (PCs./1,000 persons)	Extract the data of medical beds	–
	X_{19}: Poverty rate (%)	Calculate the ratio of population receiving minimum living security in urban and rural areas to total population	+
	X_{20}: Per capita water resources (m^3)	Extract the per capita water resources	–
	X_{21}: Sewage treatment rate (%)	Extract the sewage treatment data	+

Note: "+" represents that the indicator is positive, and "–" represents that the indicator is negative.

therefore, the water supply is sufficient, the self purification capacity is strong, and the water quality is excellent (Li et al., 2021).

The construction of the SNWDP has improved the irrigation conditions, promoted the development of fishery, forest industry, and its processing industry, and promoted the adjustment of agricultural industrial structure in the water source area. Take Nanyang, one of the core cities in the water source area, as an example. After the completion of the supporting water conservancy infrastructure related to the SNWDP, the water diversion for Nanyang can be increased by nearly 600 million cubic meters every year, with an effective irrigation area of 2,673 square kilometers, accounting for 31% of the total irrigation area of the city, which can effectively alleviate the current situation of agricultural water shortage and improve crop output in Nanyang (Wu, 2009; Yu, et al., 2021).

2.2 The Index System Construction

Referring to the existing evaluation index system of water resources value research (Fu, 2008; Hajkowicz and Higgins, 2008; Shen et al., 1998), the value of the water resources evaluation system for ecological compensation is constructed. And 21 indicators from three dimensions of nature, economy, and society are selected, as shown in **Table 1**. Specifically, the natural dimension indicators are considered from the perspectives of water quantity, water quality, and water development, which reflect the ecological environmental value of water resources. The economic dimension indicators reflect the economic value of water resources by considering the utilization efficiency of water resources, per capita income, and economic scale. Social dimension indicators are considered from the aspects of population, policy, social background, etc.

2.3 Methods

This part mainly consists of three parts: 1) introduce the methods of water resources value measurement and comprehensive evaluation, 2) introduce the method to forecast the value of water resources in the next few years, and 3) introduce the accounting method of ecological compensation.

2.3.1 Value of Water Resources Measurement Model Based on K-Means Clustering Algorithm

The K-means clustering algorithm (Zhao et al., 2021; Liu et al., 2022) is used to measure the value of water resources in this study. Select the initialized n samples as the initial clustering center and divide n data into K sets (i.e., $C = \{C_1, C_2 \cdots C_k\}$) to minimize the Within-Cluster Sum of Squares (WCSS), so as to form cluster C_i.

$$\arg\min_{s} \sum_{i=1}^{n} \sum_{x \in S_i} \|x - \mu_i^2\| \tag{1}$$

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{2}$$

$$C_i^{(t)} = \left\{x: \left\|x - m_i^{(t)}\right\|^2 \le \left\|x - m_j^{(t)}\right\|^2 \forall i, 1 \le j \le n\right\} \tag{3}$$

$$dist\left(x_i, x_j\right) = \sqrt{\sum_{d=1}^{n} \left(x_{i,d} - x_{j,d}\right)^2} \tag{4}$$

where x is the sample value, μ_i is the centroid of all points in $C^{(t)}$, $C^{(t)}$ is the new cluster formed after t times of update, $m^{(t)}$ is the sample mean vector, and $dist\ (x_i, x_j)$ represents the dissimilarity of x_i and x_j.

The clustering obtained from **Eq. 4** causes the data object x to be allocated to the nearest centroid. Taking the center of the observed value in the clustering as the new centroid, the family partition (i.e., $C = \{C_1, C_2 \cdots C_k\}$) is obtained by reciprocating division.

2.3.2 Comprehensive Evaluation Method of Water Resources Value Based on Entropy Weight Method

Entropy weight method is an objective weight assignment method to determine the index weight according to the change range of information entropy (Zou et al., 2005). The indicators are divided into positive and negative, and different algorithms are needed for standardization.

When indicators are positive

$$x_{ij}^{'} = \frac{x_{ij} - \min\left\{x_{1j}, \cdots, x_{nj}\right\}}{\max\left\{x_{1j}, \cdots, x_{nj}\right\} - \min\left\{x_{1j}, \cdots, x_{nj}\right\}} \tag{5}$$

When indicators are negative

$$x_{ij}^{'} = \frac{\max\left\{x_{1j}, \cdots, x_{nj}\right\} - x_{ij}}{\max\left\{x_{1j}, \cdots, x_{nj}\right\} - \min\left\{x_{1j}, \cdots, x_{nj}\right\}} \tag{6}$$

Then, the information entropy and weight of indicator j are calculated as follows:

$$P\left(x'_{ij}\right) = x'_{ij} \Big/ \sum_{i=1}^{n} x'_{ij} \tag{7}$$

$$e_j = -\frac{1}{\ln m} \sum_{i=1}^{n} P\left(x_{ij}^{'}\right) \ln P\left(x_{ij}^{'}\right), j = 1, \cdots, m \tag{8}$$

$$\omega_j = \frac{1 - e_j}{\sum_{j=1}^{m} 1 - e_j}, j = 1, \cdots, m \tag{9}$$

The comprehensive score of sample i is

$$S_i = \sum_{j=1}^{m} \omega_j x_{ij}^{'} i = 1, \cdots, n \tag{10}$$

Calculate the total score of each cluster of samples, and sort them according to the score, so as to distinguish the level of water resources value.

2.3.3 Value of Water Resources Prediction Model Based on LSTM Algorithm

As an improved algorithm of recurrent neural network (RNN), long short-term memory network (LSTM) model can preserve valid data information in long-term time series, which has good support for the persistence of data information. The LSTM model has both forward and back propagation functions. By changing part of the data into training set, the parameters are constantly updated in the learning process, and the processing and discarding of historical data are controlled by internal algorithm. Therefore, the LSTM model is suitable for dealing with time series problems (Omlin and Giles, 1996; Wang et al., 2018). When establishing the value of water resources prediction model based on LSTM to predict the evaluation index X_i, m, and d are taken as the current time and the sliding window size, respectively:

$$X_i(m) = [x_i(m - d + 1) \ldots x_i(m - 1), x_i(m)] \tag{11}$$

The sequence $X_i(m)$ is adopted to predict the value (i.e., $\widehat{x}_i(m + 1)$) at time m+1.

The selection of sliding window d is very critical. In this article, referring to the existing literature (Gers et al., 2000; Wang et al., 2018) and combining with the data of the water source area, the size of the sliding window is set as 5. On the basis of first-order difference, regularization, and other preprocessing of the data, the data from 2000 to 2005 are used as the training set to predict the value of water resources grade in 2006, the data from 2001 to 2006 are used as the training set to predict the value of water resources grade in 2007, and so on, the data from 2015 to 2019 are used as the training set to predict the value of water resources grade in 2020.

The main steps of water resources value prediction using LSTM are as follows:

1) To improve the stability of the time series and solve the over-fitting phenomenon that may occur in the process of prediction, the data set is stabilized (i.e., difference) and regularized;
2) The training set D is constructed according to the size of the sliding window d;
3) Establish the value of water resources prediction model based on LSTM and initialize the model parameters;
4) The prediction model is trained by the training set D. The gradient descent method is used to back propagate, and update the model parameters until the prediction accuracy requirements are met;
5) Input the first d historical observation value sequence $X_i(m)$ of X_i into the LSTM prediction model of water resources value completed by training, and obtain the predicted value $\widehat{x}_i(m + 1)$ at time m+1.

2.3.4 Calculation Method of Ecological Compensation Amount Based on Value of Water Resources

In this article, the difference between the upper limit of water price and the cost of current water price is taken as the upper limit of compensation amount, and then the ecological compensation amount is calculated according to the value of water resources. The upper limit of water price presents the water price when it reaches the maximum water price bearing index. In the calculation of ecological compensation in water source area, to convert the dimensionless evaluation level of water resources value into the scalar value of corresponding water price, it is necessary to introduce the price vector (Jiang, 1998b; Li et al., 2010; Zhu et al., 2017). The water price is calculated as follows:

$$P = V \cdot S^T \tag{12}$$

where P is the price of water resources, V is the value level of water resources, and S^T represents the value vector of water resources.

The price of water resources is calculated according to the water price bearing index method, which reflects the affordability of consumers to pay for water commodities.

$$A = SW/AE \tag{13}$$

where A is water cost bearing index, SW is water cost expenditure, and AE is income of residents. According to the international standard of water price bearing index for developing countries, 3% of residents' per capita income is generally taken as the maximum water price bearing index for households (Gakidou et al., 2017).

At present, the current water price in China mainly includes water resources fee, water supply cost, and sewage treatment fee. The upper limit of the water resources compensation price is calculated as follows:

$$PU = E \cdot \frac{A}{B} - C - D - F \tag{14}$$

where PU is the upper limit of water resources compensation price, E is per capita disposable income of residents, B is per capita water consumption, C is water supply cost, D is water resources fee, and F is sewage treatment fee.

The compensation price upper limit PU is divided according to arithmetic series, and the compensation price vector is calculated as follows:

$$S = [0, 0.25PU, 0.5PU, 0.75PU, PU] \tag{15}$$

Finally, the ecological compensation in the water source area is calculated as follows:

$$W = P \cdot QW \cdot G \tag{16}$$

where W is the amount of ecological compensation (100 million yuan/year), P is compensation price (yuan/m^3), QW is the annual water regulation (10^8 m^3/year), and G is the water quality adjustment coefficient.

2.4 Data Sources

The data of regions in the core water source area of the Middle Route of the SNWDP from 2000 to 2019 is collected and analyzed. The data are cited from Water Resources Bulletin of the study area (2000–2019), China Statistical Yearbook (2001–2020), and China Urban Construction Yearbook (2000–2019).

3 RESULTS

3.1 Calculation of Water Resources Value in Water Source Area of the Middle Route of SNWTP

3.1.1 Classification of Water Resources Value Based on K-Means Algorithm

Since value of water resources accounting belongs to high-dimensional data clustering in clustering algorithm, dimension reduction is required. Feature extraction has become a common method, and its principle is to map the original feature set from high-dimensional space to low-dimensional space using the linear mapping method (Yu and Li, 2009). In this paper, principal component analysis (Cui et al., 2020) is adopted to extract features from high-dimensional data, and the two indicators of surface water resources and groundwater resources are finally selected to measure their dissimilarity, namely x_i and x_j in **Eq. 4**. The K-means clustering results are shown in **Figure 2**.

The optimal K value is determined by the elbow method. Based on the deviation sum of squares (SSE) index, when SSE decreases gently with the increase in K, it indicates that the K value corresponding to elbow is the optimal cluster number (Hou et al., 2005). SSE is calculated as follows:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{17}$$

The errors of water resources value accounting based on the K-means clustering algorithm under the selection of different cluster numbers are shown in **Figure 3**. When the curvature is the highest, the K value corresponding to the elbow is 5. Therefore, the value of water resources in the water source region is divided into five categories. At this time, there is no grade relationship between the five categories.

After obtaining the classification results of the water resources value data in the water source area based on the K-means algorithm, the entropy weight method could be used to comprehensively score each category. The total scores are sorted in order, namely, V_1-V_5, where V_1 represents the lowest grade and V_5 represents the highest grade. The scoring results of V_1-V_5 are shown in **Table 2**.

3.1.2 Comprehensive Evaluation of Water Resources Value Based on Entropy Weight Method

The entropy weight method was used to calculate the weight of indicators in the value of water resources evaluation system constructed based on the three dimensions of nature, economy, and society. The results are shown in **Figure 4**.

By using the entropy weight method, the value of water resources of six regions in the core water source area of the Middle Route of the SNWDP from 2000 to 2019 is evaluated (**Figures 5–7**). As shown in **Figures 5–7**, the blue dotted line represents the trend line of water resources value from 2010 to 2019, the value of water resources of the six core regions in the water source area has shown an overall upward trend from 2000 to 2019. The value of water resources in northern regions (e.g., Sanmenxia, Nanyang, and Shangluo) was higher than that in the southern regions (e.g., Shiyan, Ankang, and Hanzhong) in terms of spatial distribution. The increasing rate of water resources value in different regions varied significantly in terms of time distribution. The value of water resources of Shiyan showed an obvious upward trend, while that of Hanzhong and Ankang had a gentle upward trend.

Specifically, as can be seen from **Figure 5**, the value of water resources in Shiyan has shown an obvious upward trend since 2013. The value of water resources of Shiyan was basically stable

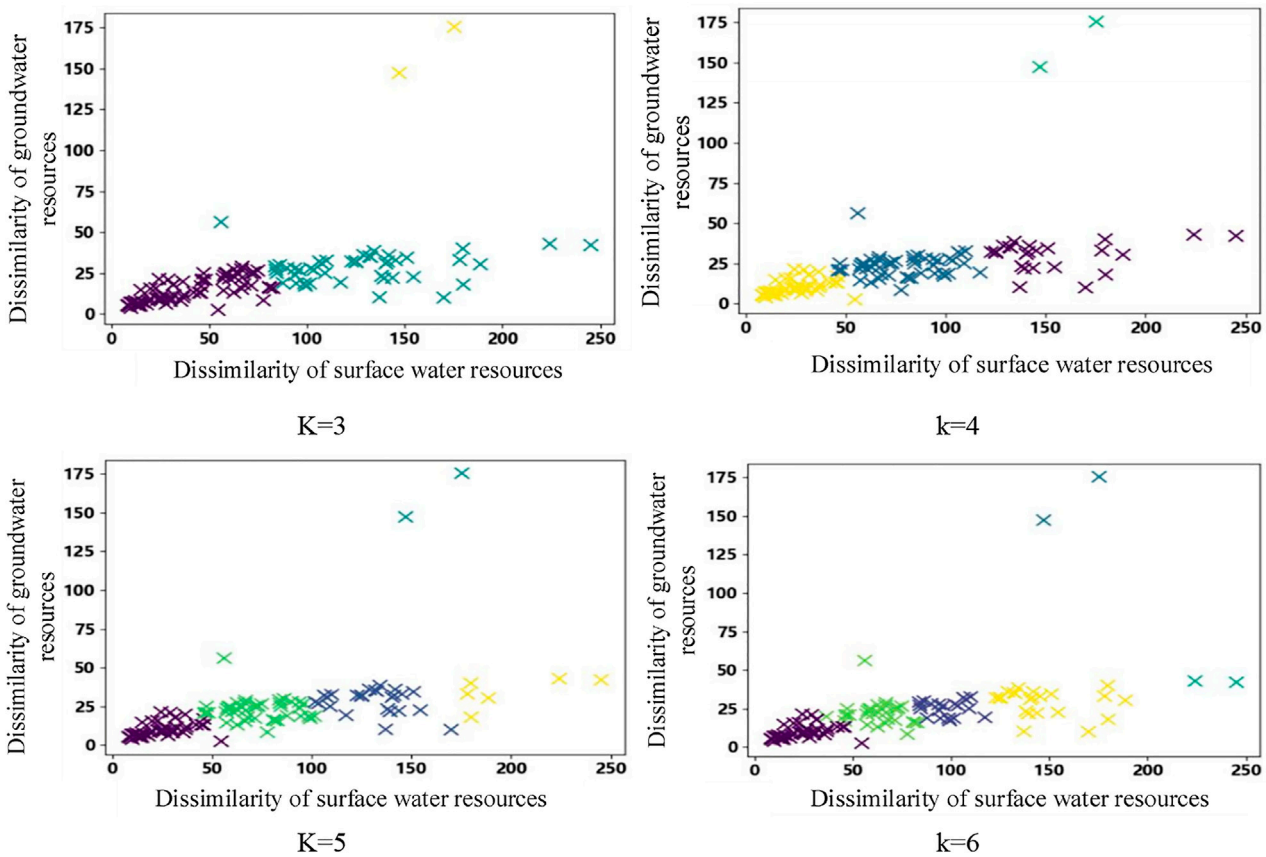


FIGURE 2 | Comparison diagram of k-means clustering results.

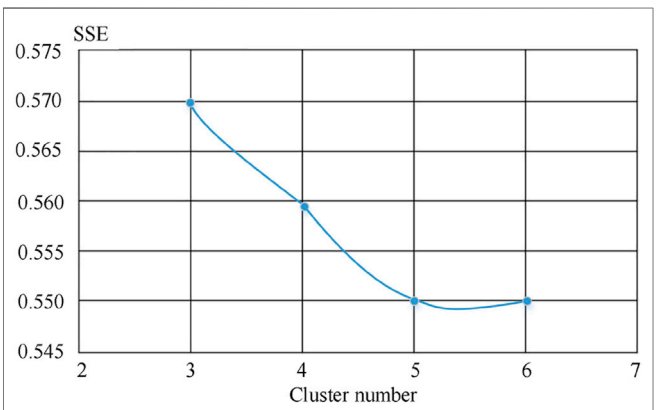


FIGURE 3 | Relation diagram of K-means clustering on the evaluation error of water resources value and cluster number.

TABLE 2 | Comprehensive evaluation results based on the entropy weight method.

Grade	V_1	V_2	V_3	V_4	V_5
Value	7.8	18.7	50.0	79.0	95

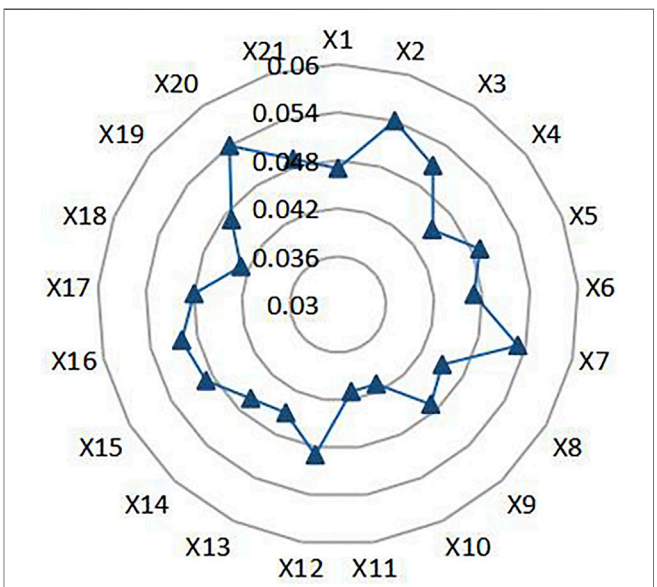


FIGURE 4 | The result of index weight evaluation based on the entropy weight method.

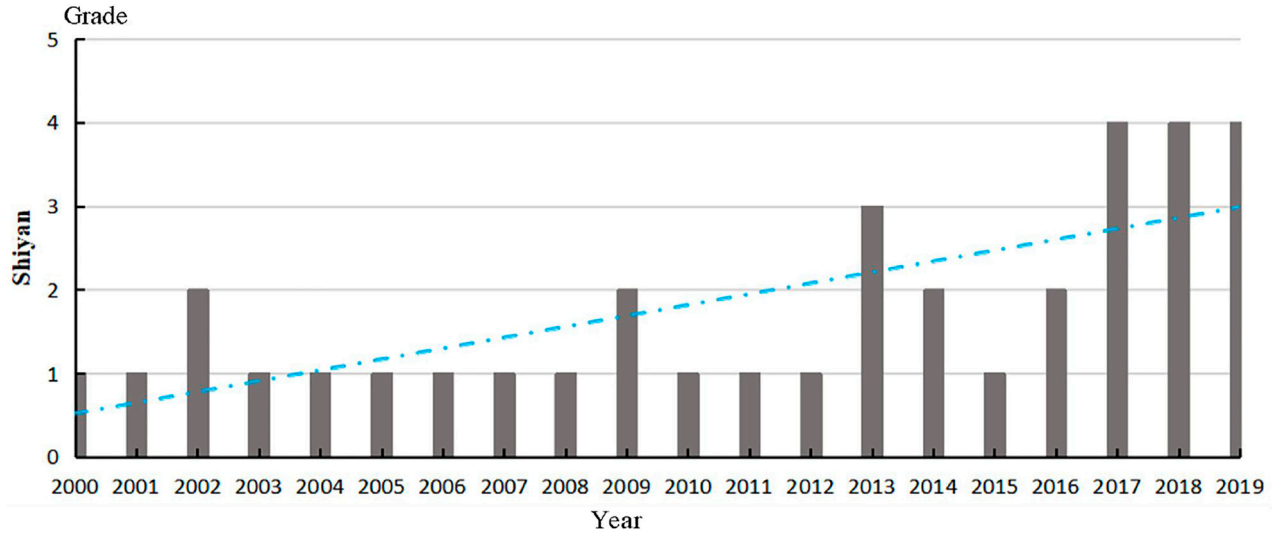


FIGURE 5 | Evaluation results of water resources value in Shiyan city, Hubei province.

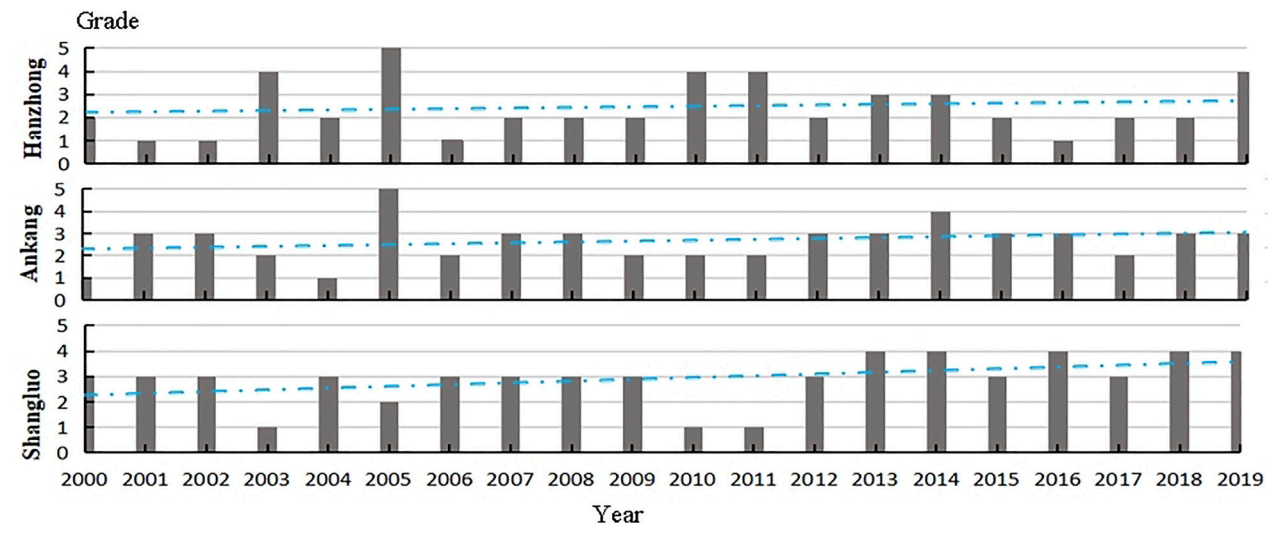


FIGURE 6 | Evaluation results of water resources value in three regions of Shaanxi province.

at V_1 level from 2000 to 2012, and only increased to V_2 level in 2002 and 2009. After 2013, it basically stayed above V_2 level, then rose and stayed at V_4 level for a long time from 2017 to 2019. There was little difference between the three regions in southern Shanxi in terms of water resources value, among which, Shangluo had the highest value of water resources, Ankang took the second place, and Shangluo had the lowest value, as shown in **Figure 6**. The value of water resources in Hanzhong was generally stable. The value of water resources in Ankang was at V_2 level before 2012. Thereafter, it increased to V_3 level. The value of water resources in Shangluo increased the most, which stabilized at V_3 level before 2013, and rose to V_4 level after 2013. The value of water resources in Sanmenxia was obviously higher than that in Nanyang (**Figure 7**). The value of water resources of Nanyang was mostly at or below V_3 level, while that of Sanmenxia reached at or above V_3 level.

3.2 Value of Water Resources Forecast in Water Source Area of the Middle Route of SNWTP

The prediction results of the water resources value in the water source area of the Middle Route of the SNWTP from 2020 to 2022 are shown in **Table 3**.

The overall level of water resources value in each region will experience an upward trend in varying degrees in the future.

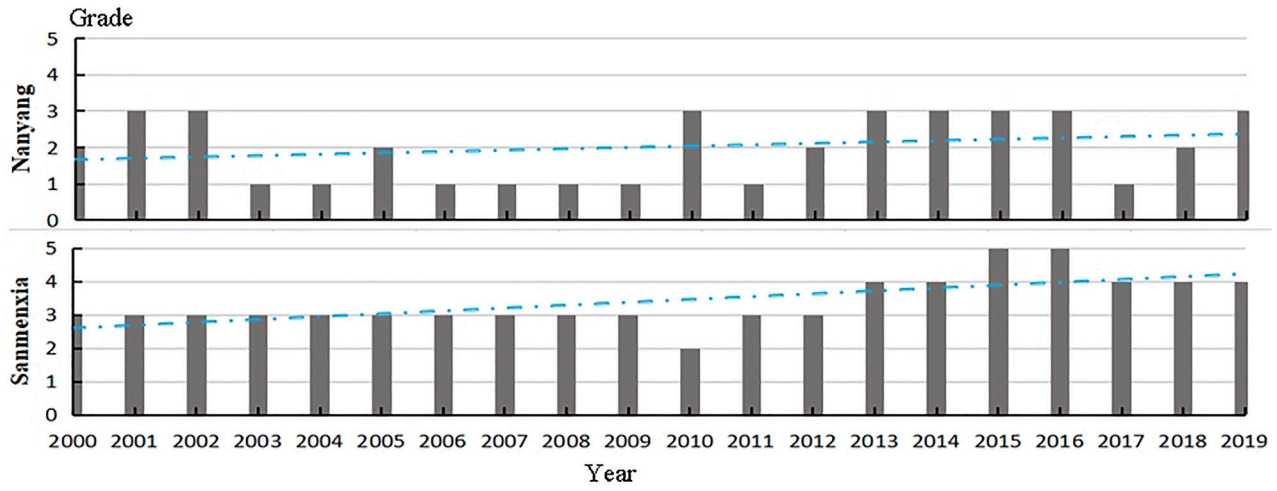


FIGURE 7 | Evaluation results of water resources value in Henan province.

TABLE 3 | Water resources value of regions in water source area in 2020–2022.

RegionYear	Shiyan	Hanzhong	Ankang	Shangluo	Nanyang	Sanmenxia
2020	V_4	V_2	V_4	V_3	V_2	V_3
2021	V_3	V_3	V_3	V_3	V_3	V_5
2022	V_5	V_4	V_3	V_3	V_3	V_5

However, the value of water resources in Hanzhong in 2022 may be lower than that in 2021, which is due to the fact that natural precipitation is the main source of water resources in Hanzhong, and its annual precipitation has an oscillation cycle of 13~14 years. After 2020, when the next rainy period is entered, the amount of water resources will increase. It should be noted that the value assessment of water resources is a high-dimensional and complex research subject, which is affected by many factors such as water quantity and quality, ecological environment, economic development, continuous investment cost of water transfer project, etc. Any change in the above factors may lead to changes in the overall valuation of water resources value.

Overall, the prediction results could basically reflect the actual situation of water resources value of core regions in the water source area from 2020 to 2022. However, with the increase in domestic and industrial water consumption, the value of water resources will be further improved in the water source area.

3.3 Accounting for Ecological Compensation in the Water Source Area of the Middle Route of SNWTP

3.3.1 Ecological Compensation Price Accounting

In terms of water supply cost and normal profit, the total cost of water supply of the Middle Route project is 3.142 billion yuan, and the average unilateral comprehensive water supply cost is 0.305 yuan/m^3 (Tan and Zhu, 1998). The Middle Route Project spans several watersheds, and the management costs vary between different provinces and regions. Generally speaking, the water resources fee of the whole project is roughly stable between 0.02 and 0.2 yuan/m^3 (Zhang et al., 2006). Combined with the experts' suggestions and the general situation of the water source area, the water resources fee is calculated at 0.2 yuan/m^3 in this paper. As for sewage treatment fee, the average treatment costs of sewage plants in Hanzhong, Ankang, and Shangluo are 0.97 yuan/m^3, 1.74 yuan/m^3, and 1.05 yuan/m^3, respectively, whereas the average collection of sewage treatment fee in these three regions is only 0.47 yuan/m^3 (Ma, 2014). Considering the situation of other regions in the water source area, 0.5 yuan/m^3 is taken as the sewage treatment fee in this paper.

The ecological compensation prices of Shiyan, Sanmenxia, Nanyang, Shangluo, Ankang, and Hanzhong are calculated based on the predicted value of water resources grade results from 2020 to 2022 (**Table 3**). Taking the data of Shiyan in 2020 as an example, the annual disposable income of residents is 21,435 yuan, and the total water consumption of the city is 9.09 yuan × 10^8 m^3, the total population is about 3.5 million. Therefore, the upper limit of compensation price in Shiyan in 2020 can be calculated as follows:

$$PU = \frac{21435 \times 3\%}{9.09 \times 10^8 / (350 \times 10^4)} - 0.305 - 0.2 - 0.5 \approx 1.468$$

The LSTM model is adopted to predict that the value of water resources of Shiyan in 2020 is at V_4 level, and the compensation price of Shiyan in 2020 is calculated as follows:

TABLE 4 | Ecological compensation prices of regions in water source areas in 2020–2022. Unit: yuan.

RegionYear	Shiyan	Hanzhong	Ankang	Shangluo	Nanyang	Sanmenxia
2020	1.101	0.227	0.573	1.440	0.445	2.018
2021	1.144	0.627	0.603	1.091	0.937	2.643
2022	1.529	0.478	0.899	1.521	1.480	2.724

TABLE 5 | The rule of water quality adjustment coefficient value.

C value	Condition
$1 < C < 2$	When the water quality is better than class I
$C = 1$	When the water quality is between class II and class I
$0 < C < 1$	When the water quality is between class III and class II
$C = -1$	When the water quality is poor V or above

$$P = [0, 0, 0, 1, 0]\,[0, 0.367, 0.734, 1.101, 1.468]^T = 1.101$$

The calculation results of ecological compensation prices of regions in the core water source area of the Middle Route in 2020–2022 are shown in **Table 4**.

As can be seen from **Table 4**, the compensation prices of regions in the water source area show an overall upward trend from 2020 to 2022. There would be obvious fluctuations in compensation prices in some regions. For instance, the compensation prices in Shangluo will drop in 2021. The regions with the highest and lowest compensation prices will be Sanmenxia and Hanzhong, respectively. From the perspective of time, the compensation prices in the water source area from 2020 to 2022 will have the largest change range in Nanyang, and the smallest change range in Shangluo. The compensation prices of six regions in the water source area in 2020–2022 will be mostly in the range of 0.5–1.5 yuan/m^3. Taking 1 yuan/m^3 and 2 yuan/m^3 as different compensation gradient boundaries, the compensation prices in Hanzhong and Ankang will be lower than 1 yuan/m^3 from 2020 to 2022. The compensation prices in Nanyang will be less than 1 yuan/m^3 in 2020 and 2021, and then show an upward trend in 2022. The compensation prices in Shiyan and Shangluo will be between 1 and 2 yuan/m^3 in the next few years. Sanmenxia will have the highest compensation prices, which are predicted basically to be above 2 yuan/m^3 from 2020 to 2022.

3.3.2 Ecological Compensation Accounting

As for the water quality adjustment coefficient *C*, the principle of "high quality, high price" shall be followed, and the adjustment coefficient is determined according to the water quality situation. **Table 5** shows the rule of water quality adjustment coefficient value.

According to the monitoring of relevant departments in China, the water quality in Danjiangkou reservoir area has been stable above the national class II standard for many years. Therefore, the *C* value is taken as 1.

Based on the compensation prices of water resources per cubic meter in the six regions in the water source area from 2020 to 2022, the ecological compensation amount is calculated in combination with the annual average water transfer and water quality of each city (**Table 6**). For instance, Shiyan is rich in water resources, and the average annual water inflow into Danjiangkou Reservoir is 3.62×10^9 m^3. The annual water diversion from Danjiangkou reservoir accounts for about 26% of the total water inflow. Therefore, the annual water diversion from Shiyan is about 9.41×10^8 m^3. Based on **Eq. 16**, it can be calculated that the ecological compensation amount of Shiyan in 2020 is 1.036 billion yuan.

As can be seen from **Table 6**, the accounting results of ecological compensation amount of regions in the water source area show an upward trend from 2020 to 2022, and the overall change trend is the same as that of ecological compensation prices of regions. As the calculation of ecological compensation amount needs to comprehensively consider the compensation price, water quality, and quantity, the compensation prices and amount in some regions do not match very well. For instance, the compensation prices of water resources in Sanmenxia are predicted high with a relatively low compensation amount. And the compensation prices of water resources in Shangluo are in the opposite situation that the compensation prices of water resources are predicted low with a relatively high compensation amount.

The annual compensation amount in the water source area from 2020 to 2022 will be the highest in Shangluo and the lowest in Hanzhong. According to the prediction results, the ecological compensation amount from 2020 to 2022 is divided into three gradients. The first gradient will include Ankang and Nanyang, with the compensation almost in the range of 0.5–1 billion yuan. The second gradient will include Shiyan, Hanzhong, and Sanmenxia, whose compensation is in the range of 1–2 billion yuan. The third gradient will include Shangluo, with the compensation in the range of 2–4 billion yuan. In 2021, the epidemic has affected the production and life of residents, resulting in reduced water demand. Therefore, the growth trend of ecological compensation amount in some regions slowed down or even fell back, which was also reflected in the prediction results.

4 DISCUSSION

4.1 Trend Analysis of Water Resources Value

By comparing the historical data and the calculation results of water resources value of core regions in the water source area, it can be seen that the change of water resources value in the water source area had experienced three stages.

TABLE 6 | Amount of ecological compensation for regions in water source areas in 2020–2022. Unit: 10^9 yuan.

Year/Region	Shiyan	Hanzhong	Ankang	Shangluo	Nanyang	Sanmenxia
2020	1.036	0.477	0.802	3.169	0.418	0.827
2021	0.691	1.381	0.905	2.509	0.975	1.348
2022	1.439	1.099	1.438	3.651	1.689	1.389

4.1.1 Initial Rising Stage of Water Resources Value

The first stage is from 2000 to 2005, during which the reservoir area had suitable climate, abundant water resources, and suitable ecological environment. The value of water resources was relatively stable and low, basically at V_1 or V_2 levels. The Middle Route project started in 2005. To make the water quality reach the high level, polluting enterprises have been rectified or even shut down. The value of water resources was coming into focus and increased obviously since that year.

4.1.2 Unstable Change Stage of Water Resources Value

The second stage is from 2006 to 2013. At the end of 2008, the Danjiangkou Reservoir area immigration pilot was officially launched. As of September 2010, there were 65,000 immigrants and the poor population increased. The Middle Route of the SNWDP was completed in 2013 and was put into operation by the end of 2014. Although experiencing the dual pressure of natural environment and social-economic development, the value of water resources in the water source area generally showed a trend of declining first and then rising with the emergence of importance of the Middle Route project to the allocation of national water resources.

4.1.3 Steady Rising Stage of Water Resources Value

The third stage is from 2014 to 2019, during which the value of water resources in the water source area was basically at V_3 or V_4 levels, and the value of water resources kept rising gently. To protect the water quality of the water source area, the State Council has successively approved a number of plans related to water pollution prevention, and soil and water conservation in Danjiangkou reservoir area, so as to ensure that the water environment of the core water source area can be at high level.

4.2 Analysis on Price Difference of Water Resources Ecological Compensation for SNWDP

The compensation prices in regions have produced obvious differences (**Table 7**). The average compensation prices of regions in the water source area could be obviously different due to the diversity of natural environment, social and economic development. Among regions with the average compensation prices less than 1 yuan/m^3, we take Hanzhong as an example for analysis. Hanzhong is located in the basin with abundant rainfall. It is the water conservation area of the water diversion project and the birthplace of the Han River. Covering an area of 27,247 square kilometers, Hanzhong is the largest city in the water source area of the Middle Route project, with a permanent population of 3.437 million in 2019. Per capita water consumption in Hanzhong reaches 450 m^3, ranking first among regions in the water source area. The industrial pillars of Hanzhong are equipment manufacturing, modern materials, and green food and medicine. Among the three regions in southern Shanxi, Hanzhong is least affected by the water transfer project. Therefore, the compensation price of Hanzhong is low. Among regions with average compensation prices between 1 and 2 yuan/m^3, Shangluo is taken as an example for analysis. As a northern city in the water source area, Shangluo covers an area of 19,851 square kilometers and has a permanent population of 2.38 million in 2019. The annual water transfer of Shangluo accounts for more than 20% of the total planned water transfer of the SNWDP, slightly lower than that of Ankang. In recent years, the ecological environment quality has been improved significantly, and the registered urban unemployment rate of Shangluo has stabilized at around 3%. The relocation of migrants and re-employment problems has been properly dealt with as well. Owing to the abundant rainfall and the increase of total water resources in Shangluo, the value of water resources will not be very high in the next few years. Among regions with average compensation prices higher than 2 yuan/m^3, we take Sanmenxia as an example for analysis. Sanmenxia belongs to the eastern extension of loess Plateau, and the problem of soil erosion is serious. With an area of 10,309 square kilometers, Sanmenxia is the smallest city in the water source area of the middle Line project, with a resident population of 2.28 million. The rainfall is at low level in the regions of the water source area, and the amount of water resources is less than other regions. The construction of water diversion project and water environment improvement have led to the closure of a large number of seriously polluting enterprises, increasing the poor population and the unemployment rate. The economic development is seriously affected by the water transfer project. Therefore, the compensation price is relatively high.

4.3 Measures to Improve the Accounting System of Ecological Compensation in Water Source Area

To establish a market ecological compensation accounting system in line with the water source area of the Middle Route, the following countermeasures and suggestions are put forward. First, establish and improve the legislation of ecological compensation in water source area, and define the subject and scope of accounting according to law. Ecological compensation in water source area is a long-term and arduous task. The object, scope, method, and standard of compensation should be

TABLE 7 | Average ecological compensation price of core cities in water source area in 2020~2022.

Ecological compensation price range	City
Less than 1 yuan/m^3	Hangzhong, Ankang, Nanyang
Between 1 and 2 yuan/m^3	Shiyan, Shangluo
Above 2 yuan/m^3	Sanmenxia

established in legal form, which can lay a legal foundation for the appropriate and sustainable development and utilization of the water resources and environment. The second is to establish the ecological compensation distribution accounting system in water source area. On the basis of establishing the standard of water price and water resources fee which is in line with the common interests of water source area and water receiving area, the differences of natural environment, productivity level, technological level, capacity, and scale of ecological protection input should be comprehensively considered to determine the allocation amount of ecological compensation. Taking the problem of soil erosion in southern Shanxi as an example, the rate is 44.6% in Hanzhong, 53% in Ankang, and 66% in Shangluo. The control expense of soil erosion in Shangluo is obviously higher than that in the other two regions. If such factors are not fully considered in the allocation of compensation, the enthusiasm of ecological protection in Shangluo would be reduced and the effect of compensation incentive cannot be achieved. The third is to introduce market mechanism and fully consider the specific situation of resource taxes and fees in water source area when calculating the compensation amount. In 2016, Beijing, Tianjin, Henan, and other regions became the pilots of water resources tax reform in China, with the water resources tax rate ranging from 0.4 yuan/m^3 to 1.8 yuan/m^3. When forecasting the amount of ecological compensation, the water-receiving area should make market compensation for the water source area according to the water resources tax rate of the water source area as reference.

5 CONCLUSION

Based on the relevant data of regions in the core water source area of the Middle Route of SNWDP from 2000 to 2019, this paper uses value of water resources accounting as the entry point of compensation accounting to predict the ecological compensation amount of the Middle Route water source area from 2020 to 2022. The main conclusions are summarized as follows:

1) Based on entropy weight method and K-means clustering, the value of water resources assessment model is constructed to calculate the value of water resources from 2000 to 2019 in the water source area. The results reveal that the value of water resources of the six core regions in the water source area has an overall upward trend, and the value growth curve of regions is slightly different. 2) Based on the value of water resources and its dynamic changes, the ecological compensation payment standard is established, and the value of water resources in the water source area from 2020 to 2022 is predicted by using the LSTM neural network model. It is found that the value of water resources in the water source area will increase in the future, and the number of the high value regions will increase. 3) The ecological compensation amount from 2020 to 2022 in the water source area is predicted. For example, in 2020, among the six regions in the water source area, the highest and lowest ecological compensation amounts are expected to reach 3.651 billion yuan in Shangluo and 0.418 billion yuan in Nanyang. To maintain the sustainability of the water transfer project, it is necessary to establish the ecological compensation accounting system in line with water source area of the Middle Route.

Water source area is a concept of geographical region. This paper takes the main regions in the water source area as the research scope, including Ankang, Hanzhong, and Shangluo in Shanxi Province, Shiyan in Hubei Province, Nanyang and Sanmenxia in Henan Province. Some regions (e.g., Baoji) are not included in the analysis. After consulting information from various sources, it is found that there are differences between administrative region and geographical region. However, the statistical data adopted in this paper are mostly divided by administrative regions, and less by geographical regions. Therefore, it is impossible to accurately calculate the value of water resources and ecological compensation amount within geographical regions. The calculation of the relevant data within the geographical regions shall be analyzed after the water source area is divided into independent administrative units.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.stats.gov.cn/tjsj/ndsj/; http://slt.henan.gov.cn/bmzl/szygl/szygb/; http://slt.hubei.gov.cn/sjfb/; http://slt.shaanxi.gov.cn/site/sxsslt/search.html?keyWord=%E6%B0%B4%E8%B5%84%E6%BA%90%E5%85%AC%E6%8A%A5.

AUTHOR CONTRIBUTIONS

JC: Conceptualization, Methodology, Formal analysis, Investigation, Validation, Writing-original draft, Writing-review and editing. QW: Conceptualization, Super-vision, Writing-original draft, Writing-review and editing. QL: Resources, Data curation, Software, Visualization.

FUNDING

This research was funded by the National Key R&D Program of China (Grant No. 2019YFC0409000), the National Natural Science Foundation of China (Grant Nos. 42071278, 42001250), the Fundamental Research Funds for the Central Universities (Grant No. B210207036).

REFERENCES

Costanza, R. (2012). Ecosystem Health and Ecological Engineering. *Ecol. Eng.* 45 (8), 24–29. doi:10.1016/j.ecoleng.2012.03.023

Cui, H. Y., Pan, M. Q., and Zhang, C. (2020). An Analysis on the Development Environment of Science and Technology Innovation Talents in Shenzhen Based on Principal Component Analysis. *Sci. Tech. Prog. Pol.* 37 (7), 35–42. doi:10.6049/kjjbydc.2019080296

Du, R. F. (2015). Preliminary Discussion on Water Pricing Mechanism of Middle Route Project of South to North Water Diversion. *Yangtze River* 46 (6), 7–9. doi:10.16232/j.cnki.1001-4179.2015.06.003

Fu, Q. F. (2008). The Application of Fussy Comprehensive Evaluation to Water Resources Value. *J. Northwest Univ. (Natural Sci. Edition)* 38 (2), 187–191.

Gakidou, E., Afshin, A., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., et al. (2017). Global, Regional, and National Comparative Risk Assessment of 84 Behavioural, Environmental and Occupational, and Metabolic Risks or Clusters of Risks, 1990-2016: A Systematic Analysis for the Global burden of Disease Study 2016. *Lancet* 390 (10100), 1345–1422. doi:10.1016/S0140-6736(17)32366-8

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Comput.* 12 (10), 2451–2471. doi:10.1162/089976600300015015

Gómez-Baggethun, E., De Groot, R., Lomas, P. L., and Montes, C. (2010). The History of Ecosystem Services in Economic Theory and Practice: From Early Notions to Markets and Payment Schemes. *Ecol. Econ.* 69 (6), 1209–1218. doi:10.1016/j.ecolecon.2009.11.007

Hajkowicz, S., and Higgins, A. (2008). A Comparison of Multiple Criteria Analysis Techniques for Water Resource Management. *Eur. J. Oper. Res.* 184 (1), 255–265. doi:10.1016/j.ecolecon.2009.11.00710.1016/j.ejor.2006.10.045

Hou, Y., Ding, Z., and He, P. L. (2005). Self-organizing Isometric Embedding. *J. Comp. Res. Dev.* 42 (2), 188–195. doi:10.1360/crad20050202

Hu, Q. (2015). Ecological Compensation Mechanism of Agricultural Water Resources: An Empirical Study in Zhejiang, China. *Ajfst* 8 (7), 525–530. doi:10.19026/ajfst.8.1559

Hu, Y. Y. (2016). *Eco-compensation Model, Accounting, Standard and Allocation Pattern: A Case Study of Hanjiang Watershed*. Beijing: People's Publishing House, 234–282.

Jia, Y. Z., Shen, J. Q., Wang, H., Jia, Y. B., and Sun, F. H. (2018). Research on Fuzzy Evaluation of Water Resources Value in Lanzhou City. *Yellow River* 40 (9), 68–73.

Jiang, W. L. (1998a). A Study on Model of Water Resources Value. *Resour. Sci.* 20 (1), 35–43.

Jiang, W. L. (1998b). *Value Theory of Water Resources*. Beijing: Science Press.

Li, H. E., Pang, M., Xiao, Y., and Shi, S. J. (2010). A Study on Ecological Compensation Quantity Based of Water Resource Value in Shaanxi Water Source Area. *J. Northwest Univ. (Natural Sci. Edition)* 40 (1), 149–154. doi:10.16152/j.cnki.xdxbzr.2010.01.030

Li, P. A., Jiang, Y. T., Qi, P. C., and Wang, L. M. (2021). *Research on the Time-Spatial Evolution Characteristics of Vegetation in the Water Source Area of the Middle Route Project of S-N Water Diversion*. Journal of Yangtze River Scientific Research Institute.

Liu, C., Nagler, O., Tremmel, F., Unterreitmeier, M., Frick, J. J., Patil, R. P., et al. (2022). Cluster-based Acoustic Emission Signal Processing and Loading Rate Effects Study of Nanoindentation on Thin Film Stack Structures. *Mech. Syst. Signal PROCESSING* 165, 108301. doi:10.1016/j.ymssp.2021.108301

Lyman, R. A. (1992). Peak and Off-Peak Residential Water Demand. *Water Resour. Res.* 28 (9), 2159–2167. doi:10.1029/92wr01082

Ma, S. J., and Wang, R. S. (1984). The Social-Economic-Natural Complex Ecosystem. *Acta Ecologica Sinica* 4 (1), 1–9.

Ma, Y. (2014). Analysis on Current Situation and Existing Problems of Sewage Treatment Market in Water Source Area of Middle Route of South-To-north Water Transfer in Shaanxi Province. *Water Conservancy Sci. Tech. Economy* 20 (9), 108–112.

Martínez-Paz, J. M., Perni, A., and Martínez-Carrasco, F. (2013). Assessment of the Programme of Measures for Coastal Lagoon Environmental Restoration Using Cost-Benefit Analysis. *Eur. Plann. Stud.* 21 (2), 131–148. doi:10.1080/09654313.2012.722923

Omlin, C. W., and Giles, C. L. (1996). Rule Revision with Recurrent Neural Networks. *IEEE Trans. Knowl. Data Eng.* 8 (1), 183–188. doi:10.1109/69.485647

Ouyang, Z. Y., Zhu, C. Q., Yang, G. B., Xu, W. H., Zheng, H., Zhang, Y., et al. (2013). Gross Ecosystem Product：concept, Accounting Framework and Case Study. *Acta Eco Sin* 33 (21), 6747–6761. doi:10.5846/stxb201310092428

Ouyang, Z. Y., and Wang, R. S. (2000a). Ecosystem Services and Their Economic Valuation. *World Sci-tech R. D* 22 (5), 45–50.

Ouyang, Z. Y., Wang, X. K., and Miao, H. (2000b). China's Eco-Environmental Sensitivity and its Spatial Heterogeneity. *Acta Ecologica Sinica* 20 (1), 9–12.

Pimentel, D., Harvey, C., Resosudarmo, P., Sinclair, K., Kurz, D., McNair, M., et al. (1995). Environmental and Economic Costs of Soil Erosion and Conservation Benefits. *Science* 267 (5201), 1117–1123. doi:10.1126/science.267.5201.1117

Piper, S. (2003). Impact of Water Quality on Municipal Water price and Residential Water Demand and Implications for Water Supply Benefits. *Water Resour. Res.* 39 (5), 135–155. doi:10.1029/2002WR001592

Qiu, L., and Zhai, H. J. (2014). An Ecological Compensation Mechanism of Chishui River Water Resources protection and Research. *Amm* 685, 463–467. doi:10.4028/www.scientific.net/AMM.685.463

Rawlins, J. M., De Lange, W. J., and Fraser, G. C. G. (2018). An Ecosystem Service Value Chain Analysis Framework: A Conceptual Paper. *Ecol. Econ.* 147, 84–95. doi:10.1016/j.ecolecon.2017.12.023

Shen, D. J. (2006). *Theory, Method and Practice of Water price Formulation*. Beijing: Water Resources and Hydropower Press.

Shen, D. J., Liang, R. J., Wang, H., and Jiang, Y. Z. (1998). The Water Resources Value. *J. Hydraulic Eng.* 5, 54–59.

Tan, C. L., and Zhu, Q. (1998). Analysis of Water Supply Cost and Water price in the Middle Route of South-To-North Water Transfer Project. *Ewrhi* 19 (9), 19–22.

Wang, H., Luan, B. Q., and Shen, D. J. (2003). *Theory and Practice of Water price for Sustainable Development*. Beijing: Science Press.

Wang, J. F., and Hou, C. B. (2013). Study on Implementation Framework and Compensation Pattern of basin Ecological Compensation Mechanism in china: From the Perspective of Compensation Funds Source. *China Popul. Resour. Environ.* 23 (2), 23–29.

Wang, S. M., and Qu, F. T. (2001). Water Resource Accounting and Adjustment of GDP: A Case Study of the Developed Areas in the Eastern Part of China. *J. Nanjing Agric. Univ.* 24 (2), 115–118.

Wang, X., Wu, J., Liu, C., Yang, H. Y., Du, Y. L., and Niu, W. S. (2018). Exploring LSTM Based Recurrent Neural Network for Failure Time Series Prediction. *J. Beijing Univ. Aeronautics Astronautics* 44 (4), 772–784. doi:10.13700/j.bh.1001-5965.2017.0285

Wei, L. L., Fu, Q., and Chen, L. Y. (2008). Elastic Analysis on the Water Demand of Urban Residential of Harbin. *J. Northeast Agric. Univ.* 39 (7), 34–37.

Wu, H. F. (2009). *Construction of Ecological Megalopolis and Coordinated Development of Regions*. Beijing: Social Science Academic Press.

Xepapadeas, A. (2011). The Economics of Ecosystems and Biodiversity: Ecological and Economic Foundations, Edited by Pushpam Kumar, 2010, London and Washington: Earthscan, ISBN 978-1-84971-212-5 (HB) Price £49.99 [Earthscan Have Offered a 20% Discount off the Book for EDE Readers. To Receive the Discount Enter 'EDETEEB' into the Voucher Code Box at the Checkout on the Earthscan Website at http://www.earthscan.co.uk/?tabid=102480]. *Envir. Dev. Econ.* 16, 239–242. doi:10.1017/S1355770X11000088

Xie, G. D. (2012). Realization Mechanism of Ecosystem Service Value. *Environ. Prot.* 17, 16–18. doi:10.14026/j.cnki.0253-9705.2012.17.010

Xu, D. W., Zheng, H. X., and Liu, M. Q. (2008). Measuring Method of River basin Ecological Compensation Based on River Water Quality and its Water Quantity about across Administration Area. *China Popul. Resour. Environ.* 18 (4), 189–194.

Xu, J., Xiao, Y., Xie, G., and Jiang, Y. (2019). Ecosystem Service Flow Insights into Horizontal Ecological Compensation Standards for Water Resource: A Case Study in Dongjiang Lake Basin, China. *Chin. Geogr. Sci.* 29 (2), 214–230. doi:10.1007/s11769-019-1025-3

Xu, J., Xiao, Y., Xie, G., Liu, J., Qin, K., Wang, Y., et al. (2021). How to Coordinate Cross-Regional Water Resource Relationship by Integrating Water Supply Services Flow and Interregional Ecological Compensation. *Ecol. Indicators* 126, 107595. doi:10.1016/j.ecolind.2021.107595

Yang, M.-j., Yang, K., Li, G., and Niu, X.-d. (2019). The Cooperation Mechanism of Water Resources protection in Trans-boundary River Based on Game Theory:

A Case Study of the Taipu River in the Taihu Lake Basin. *J. Nat. Resour.* 34 (6), 1232–1244. doi:10.31497/zrzyxb.20190609

Yang, X., Zhang, H., Yang, J., Li, H., and Zhu, X. (2020). Research on Ecological Compensation of Water Resources in the Yellow River basin. *Fresenius Environ. Bull.* 29 (6), 4426–4429.

Yu, T. S., Zheng, S. N., Zhu, J. Y., Tang, M. F., Dong, R. C., and Wang, Y. (2021). Evaluation on the Ecological Security Status in Nanyang City, the Water Source Region of the Middle Route of South-To-North Water Diversion Project in China. *Acta Ecoligica Sinica* 41 (18), 7292–7300. doi:10.5846/stxb202003260701

Yu, X., and Li, M. Q. (2009). Collaborative Filtering Recommendation Model Based on Effective Dimension Reduction and K-Means Clustering. *Appl. Res. Comput.* 26 (10), 3718–3720.

Zhang, Y. M., Xiong, Y. Q., and Zhang, N. (2006). Analysis of Water Resource Fee Collection in the Middle Route of South-To-North Water Transfer Project. *Water Resour. Dev. Res.* (6), 30–32. doi:10.13928/j.cnki.wrdr.2006.06.008

Zhang, Z. Q., Xu, Z. M., Cheng, G. D., and Su, Z. Y. (2002). Contingent Valuation of the Economic Benefits of Restoring Ecosystem Services of Zhangye Prefecture of Heihe River basin. *Acta Ecologica Sinica* 22 (6), 885–893.

Zhang, Z. Q., Xu, Z. M., and Cheng, G. D. (2001). Valuation of Ecosystem Services and Natural Capital. *Acta Ecologica Sinica* 21 (11), 1918–1926.

Zhao, W. H., Gao, Y., and Dai, T. S. (2007). An Optimization Model of Initial Emission Permits Distribution. *Syst. Eng.* 25 (6), 57–61.

Zhao, Z., Huang, Q. Q., and Tong, N. N. (2021). Evaluating SMEssupporting Policies during COVID-19 Pandemic with K-Means Clustering. *Data Anal. Knowledge Discov.* 5 (12), 148–157. doi:10.11925/infotech.2096-3467.2020.0320

Zhu, J. L., Wang, J., Tao, X. Y., and Wang, S. J. (2017). The Allocation of Ecological Compensation Funds in the Water Source Area of the Middle Route Scheme of South-north Water Diversion Project Based on Ecosystem Services Value. *Ecol. Economy* 33 (6), 127–132, 139.

Zi, T. (2010). Assessment on Ecosystem Service Value of the Water Tourism Resources in Heilongjiang Province, Northeast China. *Adv. Mater. Res.* 113-116, 132–136. doi:10.4028/www.scientific.net/AMR.113-116.132

Zou, Z. H., Sun, J. N., and Ren, G. P. (2005). Study and Application on the Entropy Method for Determination of Weight of Evaluating Indicators in Fuzzy Synthetic Evaluation for Water Quality Assessment. *Acta Scientiae Circumstantiae* 25 (4), 552–556. doi:10.13671/j.hjkxxb.2005.04.022

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.

Copyright © 2022 Chen, Wang and Li. 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 in Environmental Science

ORIGINAL RESEARCH
published: 25 March 2022
doi: 10.3389/fenvs.2022.873939

Multivariate Adaptive Step Fruit Fly Optimization Algorithm Optimized Generalized Regression Neural Network for Short-Term Power Load Forecasting

Feng Jiang, Wenya Zhang and Zijun Peng*

School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China

Short-term load forecasting plays a significant role in the management of power plants. In this paper, we propose a multivariate adaptive step fruit fly optimization algorithm (MAFOA) to optimize the smoothing parameter of the generalized regression neural network (GRNN) in the short-term power load forecasting. In addition, due to the substantial impact of some external factors including temperature, weather types, and date types on the short-term power load, we take these factors into account and propose an efficient interval partition technique to handle the unstructured data. To verify the performance of MAFOA-GRNN, the power load data are used for empirical analysis in Wuhan City, China. The empirical results demonstrate that the forecasting accuracy of the MAFOA applied to the GRNN outperforms the benchmark methods.

Keywords: power load, multivariate adaptive step, fruit fly optimization algorithm, generalized regression neural network, forecasting

OPEN ACCESS

Edited by:
Wendong Yang,
Shandong University of Finance and Economics, China

Reviewed by:
Shiping Wen,
University of Technology Sydney, Australia
Huabin Chen,
Nanchang University, China

****Correspondence:***
Feng Jiang
fjiang@zuel.edu.cn

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: *11 February 2022*
Accepted: *07 March 2022*
Published: *25 March 2022*

Citation:
Jiang F, Zhang W and Peng Z (2022) Multivariate Adaptive Step Fruit Fly Optimization Algorithm Optimized Generalized Regression Neural Network for Short-Term Power Load Forecasting. Front. Environ. Sci. 10:873939. doi: 10.3389/fenvs.2022.873939

INTRODUCTION

It is well known that the role of short-term power load forecasting is increasingly crucial in the management of power plants. Short-term power load forecasting mainly refers to electric load forecasting in the next few hours, 1 day to several days. Accurate short-term power load forecasting can reasonably arrange the operation of units, ensure the safety of operation of the power grids, and improve the economic benefits of power enterprises (Friedrich and Afshari, 2015; Dudek, 2016). On the contrary, inaccurate forecasts will produce unnecessary electricity and result in considerable electrical power system losses (Yang et al., 2017). Hobbs et al. (1999) pointed that the reduction of 1% in load forecasting error of 10,000 MW utility can save up to $1.6 million annually. So, it is of vital importance to achieve high accuracy for short-term power load forecasting nowadays.

With the development of computer technology, the theory of artificial neural networks (ANNs) has been applied in a wide range of fields such as power market, system engineering, and control system (Jiang et al., 2014; Liu et al., 2018; Du et al., 2019; Yang et al., 2022). The forecast of power load considers not only the load but also the factors that affect the load, so the use of ANNs has been highly concerned by researchers. For example, Xuan et al. (2021) combined the convolutional neural network (CNN) and bidirectional gated recurrent unit (Bi-GRU) to forecast the short-term load. In the meantime, the random forest was used to select features. The final result showed that this hybrid method had a higher accuracy. Wang et al. (2020) applied an extreme learning machine model to electricity price forecasting, as well as considering the influence of outliers. The Elman neural

network (ENN) model was also used to forecast the electrical power system (Zhen Wang et al., 2018). Abedinia and Amjady (2016) presented a new stochastic search algorithm to find the optimum number of neurons for the hidden layer, and they used the proposed method to predict the power load. They compared the obtained results with those of several other recently published methods, and it confirmed the validity of the developed approach. Lu et al. (2016) used the weighted fuzzy C-means clustering algorithm based on principal component analysis to determine the basis function centers, and they used the gradient descent algorithm to train the output layer weights. The proposed model was implemented on real smart meter data, and simulation results showed that the proposed method had good forecasting accuracy. Ding et al. (2016) applied variable selection and model selection to power load forecast to ensure an optimal generalization capacity of the neural network model, and the results showed that the neural network–based models outperform the time series models.

The generalized regression neural network (GRNN) is a type of ANNs based on mathematical statistics, proposed by Specht (1991). Instead of listing the equations in advance, the network uses a probability density function to predict the output. Therefore, the GRNN has strong non-linear mapping capability and quick learning speed, which is better than the radial basis function neural network. In addition, even if the number of input training samples is small, its output can converge to the optimal value, which is very suitable for solving the problem of non-linearity (Jiang and Chen, 2016; Zhu et al., 2018). It has been applied in a wide range of fields such as prediction of wind speed (Kumar and Malik, 2016), two-dimensional spectral images (Jianzhou Wang et al., 2018), automated emotion detection systems (Talele et al., 2016), short-term load forecasting (Hu et al., 2017), mineral resource estimation (Das Goswami et al., 2017), and the estimation of peak outflow (Sammen et al., 2017). The optimization of smoothing parameter is a crucial step in the application of GRNN. There are a few ways to estimate its value. For example, Agarkar et al. (2016) applied particle swarm optimization (PSO) to the smoothing parameter of GRNN, which reduced the time complexity and produced more accurate results than random selection of spread factor. Gao and Chen (2015) presented an improved GRNN algorithm, using phase space reconstruction to strike GRNN training samples, applying adaptive PSO algorithm to optimize the smoothing parameter. Zhao et al. (2020) applied PSO-GRNN for risk prediction of urban logistics and found that the model can handle the high-frequency influencing factors well. The result showed that PSO-GRNN can better improve the accuracy of prediction than others.

Recently, Pan (2012) proposed a fruit fly optimization algorithm (FOA) to optimize the financial distress model, which was based on the foraging behavior of fruit flies. This algorithm has been effectively applied in a few fields including the dual-resource constrained flexible job-shop scheduling problem (Zheng and Wang, 2016), monthly electricity consumption forecasting (Jiang et al., 2020), multidimensional knapsack problem (Meng and Pan, 2017), seasonal electricity consumption forecasting (Cao and Wu, 2016), joint replenishment problems (Wang et al., 2015), steelmaking casting problem (Li et al., 2018), and optimization of support vector regression (Samadianfard et al., 2019; Zhang and Hong, 2019; Sattari et al., 2021). With the extensive applications of FOA, more and more scholars studied the optimization of this algorithm. Hu et al. (2017) changed the step length of the fruit fly from a constant to a decrement sequence to improve the optimization abilities of FOA, and the empirical results showed that the performance of the proposed algorithm was improved. Pan et al. (2014) introduced a new control parameter that adaptively adjusted the range of search space around the location of the cluster, and the accuracy and convergence speed were improved.

In this paper, we propose a multivariate adaptive step fruit fly optimization algorithm (MAFOA) to optimize the smoothing parameter of GRNN for short-term load forecasting. We make three contributions as follows. Firstly, we consider factors that affect the power load as much as possible, such as temperature, weather type, and date type. Secondly, we propose an efficient interval partition technique to handle the structured and unstructured data. Finally, we improve the selection of step size, which has a multivariate adaptive step and can achieve high adaptability.

The remainder of this paper is organized as follows. FOA and its improvement are presented in *The Improvement of Fruit Fly Optimization Algorithm*. *Improvement of Generalized Regression Neural Network* shows the MAFOA-optimized GRNN for short-term load forecasting. We carry out the empirical analysis and compare the proposed model with other models in *Empirical Analysis*. Finally, the summary of this study is drawn in *Conclusion*.

THE IMPROVEMENT OF FRUIT FLY OPTIMIZATION ALGORITHM

Considering the problems of local optimum in the ordinary FOA, we propose the MAFOA to optimize the smoothing parameter of GRNN. In this section, we first briefly introduce the ordinary FOA in *Fruit Fly Optimization Algorithm*, and then we propose the MAFOA in *Multivariate Adaptive Step Fruit Fly Algorithm*.

Fruit Fly Optimization Algorithm

Fruit fly is a kind of flying insect, which is very sensitive to the external environment because of its superior olfactory and vision. Firstly, the olfactory organ is used to obtain the odor floating in the air. Then, it will distinguish the general direction of the food source and fly to the source of food. Finally, the fruit fly can discover the position of food by its keen vision, and then fly to the position. The process of searching food for the fruit flies can be simulated as follows (Mitić et al., 2015):

1) Randomly initialize the population size, maximal number of iterations, and position coordinates (x, y) of the group in a set interval.
2) Choose the search radius of the fruit fly. Then, determine the new position coordinates (x_i, y_i) of individual fruit fly by using

$$\begin{cases} x_i = x + L_0 \times rands(-1, 1), \\ y_i = y + L_0 \times rands(-1, 1), \end{cases} \tag{1}$$

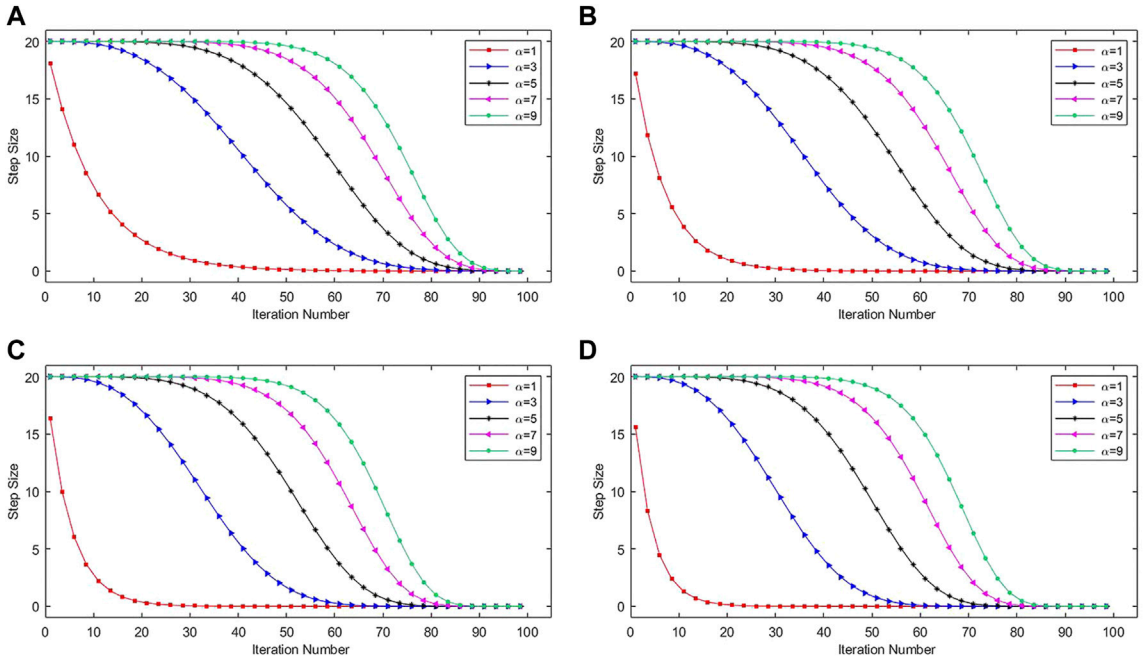


FIGURE 1 | Multivariate adaptive step size corresponding to different values of N and α: **(A)** N = 10, **(B)** N = 15, **(C)** N = 20, and **(D)** N = 25.

where L_0 is a fixed step size and $rands(-1, 1)$ is a sample of uniform distribution on $(-1, 1)$.

3) Estimate the distance (d_i) between the individual fruit fly and the coordinate origin and then calculate the judgment value (S_i) of smell concentration:

$$d_i = \sqrt{x_i^2 + y_i^2}, \tag{2}$$

$$S_i = 1/d_i. \tag{3}$$

4) Calculate the smell concentration $(smell_i)$ by substituting S_i into the fitness function (f) of the taste concentration:

$$smell_i = f(S_i). \tag{4}$$

5) Find out the best smell concentration $(smell_i)$ among the fruit fly swarm:

$$[bestsmell \quad bestindex] = best(smell_i), \tag{5}$$

where $bestsmell$ is the extreme value of $smell_i$ and $bestindex$ is the position coordinate of the individual fruit fly with best smell concentration.

6) Determine whether the smell concentration is better than the previous one. If yes, implement step 7; otherwise, repeat the process from step 2 to step 6.
7) Retain the best smell concentration value $(Smellbest)$ and the position coordinate of the individual fruit fly with the best smell concentration (x_{best}, y_{best}):

$$Smellbest = bestsmell, \tag{6}$$

$$\begin{cases} x_{best} = x(Smellbest), \\ y_{best} = y(Smellbest). \end{cases} \tag{7}$$

8) Determine whether the end condition is reached. If yes, find out the location of the best smell concentration value; otherwise, return to step 2.

Multivariate Adaptive Step Fruit Fly Algorithm

In the ordinary FOA, the individual fruit fly seeks the food source with the pre-set step size. Obviously, if the step size is too small, the search space will be limited, and it will cause the problem of local optimum. On the contrary, if the step size is too large, its local search ability will become weaker, and the convergence rate will slow down. To deal with these issues, the setting of step size should adhere to the following principles. In the initial phase of iterations, the step size should be large to ensure global optimization performance. On the contrary, in the later stage, the step size should be small to ensure local search performance.

Therefore, there are a few successful algorithms for the improvement of step size of fruit flies, such as the decreasing step fruit fly optimization algorithm (DSFOA) (Hu et al., 2017), self-adaptive step fruit fly optimization algorithm (FFOA) (Yu et al., 2016), and improved fruit fly optimization algorithm (IFFO) (Pan et al., 2014). In the DSFOA and IFFO, the step size decreased quickly in the initial phase of iteration, which cannot guarantee the global optimization performance of the algorithm. In this paper, we propose the multivariate adaptive step size, which can be demonstrated as follows:

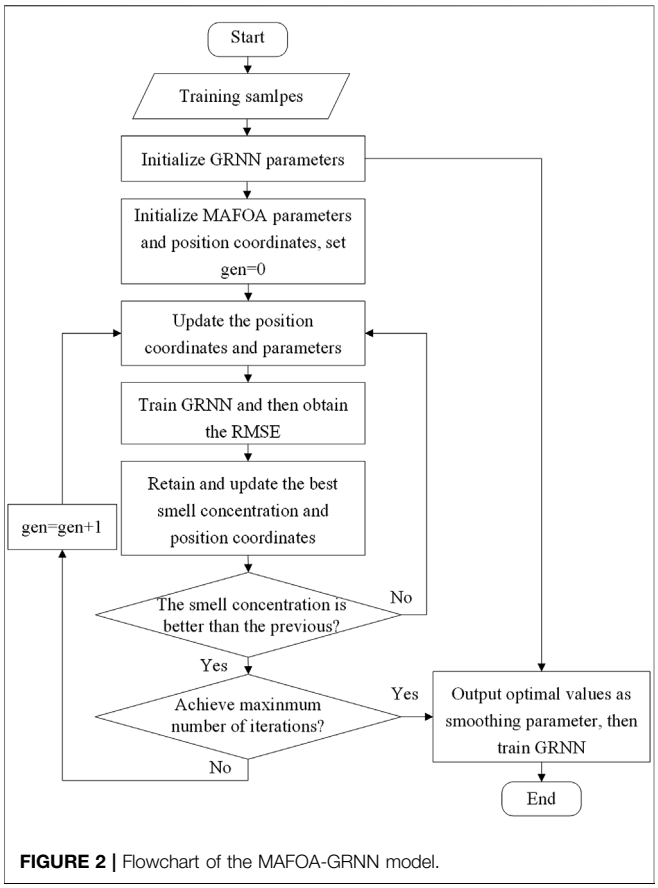


FIGURE 2 | Flowchart of the MAFOA-GRNN model.

$$L_i = L_0 \cdot \exp\left[-N\left(\frac{G_i}{G_{\max}}\right)^{\alpha}\right], \tag{8}$$

where L_0 is the initial step size, G_i is the current number of iterations, $G_{\max}$ is the maximum number of iterations, N is a positive integer, and the exponential factor α is a constant within $(0, 10)$. The positive integer N and the exponential factor α control the decreasing rate of step size and realize the better local search performance. In order to choose proper values of N and α, the convergence ability of algorithm under different parameter values is compared. The initial step size L_0 is set to 20, and the maximum number of iterations $G_{\max}$ is set to 100. **Figure 1** gives the variations of step size in **Eq. 8** corresponding to different values of α when $N = 10, N = 15, N = 20$, and $N = 25$, respectively.

As shown in **Figure 1**, the step size decreases gradually from 20 to 0 with the increasing iteration number and different values of α correspond to different step size change trends. In the initial stage of iterations, the algorithm has the largest step size, which can guarantee the global optimum. As the iteration number increases, the capability of local search is gradually enhanced to find the local optimum value, which can be seen from the rapid decline in the curves. Therefore, the dynamic step size can realize the balance of global search capability and local optimization ability.

Besides, from the subfigures in **Figure 1**, the step size changes relatively symmetrical when $\alpha = 3, \alpha = 5$, and $\alpha = 7$. When $\alpha = 1$, the curve drops sharply from the beginning, which means the step size will become small even before achieving the global optimum, and the step size cannot achieve 20 at the beginning of iteration. The moment when step size begins to decline is a bit later when $\alpha = 9$. There seems to be no difference in convergence performance when N takes different values. So in *Empirical Analysis*, we will test the performance of the proposed model with different values of N and α to search for the optimal value, and we will substitute the optimal N and α into the model for short-term load forecasting.

IMPROVEMENT OF GENERALIZED REGRESSION NEURAL NETWORK

Generalized Regression Neural Network

The GRNN is a kind of neural network using the radial basis function and has been very popular in applications in recent years. It can establish the implicit mapping relationship according to the sample data, so that the output can converge the optimal regression surface. Once the sample is determined, the only goal is the determination of smoothing parameter in the kernel function (Ozturk and Turan, 2012; Kumar and Malik, 2016).

Assuming that $f(x, y)$ is the joint probability density function of random variable X and variable Y, the observed value of X is x_0, and the regression of Y with respect to X is

$$\hat{Y}(x_0) = \frac{\int_{-\infty}^{\infty} y f(x, y) dy}{\int_{-\infty}^{\infty} f(x_0, y) dy}. \tag{9}$$

Based on the Parzen non-parametric estimation, the density function $f(x_0, y)$ can be estimated by the sampled dataset $\{x_i, y_i\}_{i=1}^{n}$:

$$f(x_0, y) = \frac{1}{n(2\pi)^{\frac{p+1}{2}}\sigma_1\sigma_2 \ldots \sigma_p\sigma_y} \sum_{i=1}^{n} e^{-d(x_0, x_i)} e^{-d(y, y_i)}, \tag{10}$$

$$d(x_0, x_i) = \sum_{j=1}^{n} \left[\left(x_{0j} - x_{ij}\right)/\sigma_j\right]^2, \tag{11}$$

$$d(y, y_i) = (y - y_i)^2, \tag{12}$$

where n is the sample size, p is the dimension of random variable X, and σ is the width coefficient of the Gaussian function, which is called the smoothing parameter.

Substituting **Eq. 10** into **Eq. 9** yields

$$\hat{Y}(x_0) = \frac{\sum_{i=1}^{n}\left(e^{-d(x_0, x_i)} \int_{-\infty}^{\infty} y e^{-d(y, y_0)} dy\right)}{\sum_{i=1}^{n}\left(e^{-d(x_0, x_i)} \int_{-\infty}^{\infty} e^{-d(y, y_0)} dy\right)}. \tag{13}$$

Note that $\int_{-\infty}^{\infty} z e^{-z^2} dz = 0$, (**Eq. 13**) can be simplified as follows:

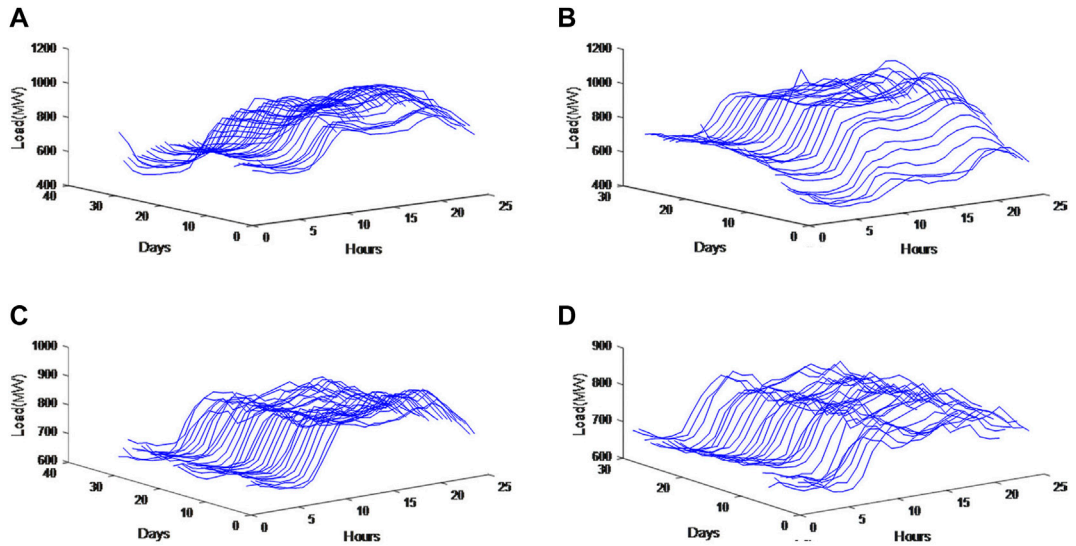


FIGURE 3 | Historical power load curve: **(A)** January 2014, **(B)** February 2014, **(C)** March 2014, and **(D)** April 2014.

$$\hat{Y}(x_0) = \frac{\sum_{i=1}^{n} y_i e^{-d(x_0,x_i)}}{\sum_{i=1}^{n} e^{-d(x_0,x_i)}}. \quad (14)$$

The predicted value in **Eq. 14** is the weighted sum of the observations of the dependent variable, and the weights are $e^{-d(x_0,x_i)}$. The GRNN is composed of input layers, pattern layers, summation layers, and output layers. Once the learning samples are determined, the structure of neural network and the connection weights between neurons are completely determined. Therefore, the GRNN does not need to adjust the connection weight values between neurons, but to adjust the transfer function of each unit by changing the smoothing parameter to obtain the best regression result, which is different from the traditional error backward propagation algorithm. Thus, a key step in the GRNN is to determine the value of the smoothing parameter.

Optimization of Generalized Regression Neural Network Based on Multivariate Adaptive Step Fruit Fly Optimization Algorithm

In this paper, the MAFOA is applied to optimize the smoothing parameters in the GRNN. The MAFOA-GRNN takes the root mean square error (RMSE) of GRNN as the fitness function of MAFOA, so as to calculate the smell concentration in each iteration. Part of the training data are used in the MAFOA to select the best parameters for the GRNN. When the algorithm reaches the maximum number of iterations, the location of the fruit fly with best smell concentration is obtained. Then, these optimal parameters will be used in the GRNN to get the optimal prediction model. The flowchart of the MAFOA-GRNN model is shown in **Figure 2**.

EMPIRICAL ANALYSIS

In this section, the power load data in Wuhan are used to test the performance of MAFOA-GRNN. The data description is introduced in *Data Description*. Then, *Data Processing* is discussed. The evaluation criteria and empirical results are further discussed in *Evaluation Criteria* and *Experimental Analysis*.

Data Description

The power load data used in this paper are hourly and obtained from a power grid in Wuhan with 2,880 observations ranging from January 1, 2014, to April 30, 2014, which are shown in **Figure 3**. In this section, we predict the power load of the last day of each month. The in-sample data are power load data of each month except the last day, and the out-of-sample data are the power load data of the last day of each month.

As shown in **Figure 3**, the short-term power load has obvious periodicity. Therefore, historical load data are an important reference for forecasting. In order to accurately predict the power load, the factors influencing the power load should be considered as much as possible. The factors related to load forecasting include date classification (weekday, weekend, holiday), daily temperature (maximum, minimum, average temperature), and weather condition.

Combining the influence factor, the improved GRNN adopts a three-layer network structure. The input variables of the GRNN are shown in **Table 1**, and the corresponding output vector is the power load value at t o'clock on day d.

Data Processing

The original load data are normalized to eliminate the impact of the dimensions between indicators. In addition, the input variables of GRNN in **Table 1** should be numerical data, so we quantify the above weather factors and date type factors.

TABLE 1 | Input variables of the GRNN.

Number	Input variables
1	Power load value at t o'clock on day $d-2$
2	Power load value at $t-1$ o'clock on day $d-2$
3	Maximum temperature on day $d-2$
4	Minimum temperature on day $d-2$
5	Weather condition on day $d-2$
6	Date type on day $d-2$
7	Power load value at t o'clock on day $d-1$
8	Power load value at $t-1$ o'clock on day $d-1$
9	Maximum temperature on day $d-1$
10	Minimum temperature on day $d-1$
11	Weather condition on day $d-1$
12	Date type on day $d-1$
13	Maximum temperature on day d
14	Minimum temperature on day d
15	Weather condition on day d
16	Date type on day d

Meanwhile, we propose an efficient interval partition technique to handle temperature and weather types:

1) Normalization of load data. All load data are normalized by using the linear transformation method, given by

$$y = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{15}$$

where $x_{\min}$ is the minimum load value in the dataset and $x_{\max}$ is the maximum load value in the dataset.

2) Quantization of temperature. In the previous studies, temperature is standardized by direct standardization (Hu et al., 2017). When the temperature changes in a suitable range, the effect of the load is small. However, when the temperature increases or decreases to a certain extent, the effect on the load will be larger gradually. Therefore, standardization may not be an appropriate choice. In this work, we propose an efficient interval partition technique. The temperature is partitioned by intervals, and different quantitative values are taken according to the situation. For example, when the temperature is 0°C, the temperature is coded as 1; when the temperature is 5°C, the temperature is coded as 0.8. The specific code value can be adjusted within a small range according to the previous prediction result. Therefore, the temperature is partitioned by intervals, as shown in **Table 2**.
3) Quantization of weather types. The weather types can be divided into six categories, as shown in **Table 2**, which can affect the power load by influencing the use of lighting equipment and other household appliances. Their corresponding quantized values are also shown in **Table 2**.
4) Quantization of date types. As a result of the social production modules, the electricity consumption generally shows the alternation of work and rest. The date types can be divided into three categories: weekday (Monday to Friday), weekend (Saturday to Sunday), and holiday (holiday or major event day). On holiday, people often go out to relax or take a rest, which has a substantial impact on the changes in power load. According to the degree of influence on power load, the date type is coded as three categories: weekday is coded as 0, weekend is coded as 0.5, and holiday is coded as 1.

Evaluation Criteria

This paper uses the normalized root mean square error (NRMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) as the evaluation criteria, given by

$$NRMSE = \frac{100}{\bar{y}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{16}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{17}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{18}$$

where $\bar{y}$ is the mean of value, $\hat{y}_i$ is the predicted value, y_i is the observation value, and N is the number of data.

Although the NRMSE, MAE, and MAPE can be used as criteria to obtain model predicted loss values, it cannot be verified whether the comparison result is statistically significant. To solve this problem, Diebold and Mariano (1994) proposed the Diebold–Mariano (DM) test to test the statistical significance of different prediction models. Assume that model B and model T do the forecasting task in period t at the

TABLE 2 | Quantitative value of meteorological factors.

Temperature (°C)	Quantitative value	Weather type	Quantitative value
–5–0	(0.7, 1.0)	Sunny	(0, 0.1)
0–5	(0.5, 0.8)	Sunny–cloudy	(0.1, 0.2)
5–10	(0.3, 0.6)	Cloudy	(0.2, 0.4)
10–15	(0.2, 0.4)	Cloudy–rainy	(0.3, 0.6)
15–20	(0.1, 0.2)	Rainy	(0.5, 0.8)
20–25	(0, 0.1)	Snowy	(0.7, 1.0)
25–30	(0.1, 0.4)		
30–35	(0.4, 0.7)		
35–	(0.1, 1.0)		

same time, and we wonder if there are significant differences in the performance between the two models. The original hypothesis is that the forecast accuracy for two models is the same, which is equivalent to the mean value of relative loss function of 0. The DM statistics is defined as follows:

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi f_d(0)}{T}}}, \tag{19}$$

where

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t \tag{20}$$

is the sample mean loss differential, in which

$$d_t = Loss_T - Loss_B \tag{21}$$

is the relative loss function, where $Loss_T$ and $Loss_B$ are the loss function of predicted errors of test model T and benchmark method B at time t, respectively.

Note that, in this paper, the mean-squared prediction error (MSPE) is used as the loss function:

$$Loss_i = \frac{1}{N}\sum_{i=1}^{N}(y_t - \hat{y}_{it})^2, \tag{22}$$

where $\hat{y}_{it}$ is the predicted value of model i at time t.

$$f_d(0) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau) \tag{23}$$

is the spectral density of relative loss function at frequency zero.

$$\gamma_d(\tau) = E\left[(d_t - \mu)(d_{t-\tau} - \mu)\right] \tag{24}$$

is the autocovariance of d_t at displacement τ, where μ is the population mean loss differential.

If the p-value corresponding to DM is less than the significant level, which normally is 0.01 or 0.05, the original hypothesis is rejected; otherwise, it cannot be rejected.

TABLE 3 | Errors of the test set for different α and N.

n	Error type	$\alpha = 1$	$\alpha = 3$	$\alpha = 5$	$\alpha = 7$	$\alpha = 9$
$N = 10$	NRMSE	33.3898	22.8639	15.8455	7.3671	7.4929
	MAE	7.2234	7.0068	6.5260	5.2611	6.1537
	MAPE	0.0093	0.0083	0.0080	0.0072	0.0081
$N = 15$	NRMSE	34.4198	23.7839	27.4610	1.0415	5.5455
	MAE	7.0454	7.0378	6.5378	6.5242	6.6737
	MAPE	0.0086	0.0087	0.0089	0.0087	0.0097
$N = 20$	NRMSE	38.8445	35.8757	10.7160	4.7947	0.0072
	MAE	7.3071	6.5455	5.8260	5.6311	6.2587
	MAPE	0.0098	0.0083	0.0080	0.0076	0.0081
$N = 25$	NRMSE	39.7045	20.3099	11.6360	5.6547	8.2882
	MAE	7.2102	6.5139	6.5072	5.9937	6.8254
	MAPE	0.0098	0.0094	0.0093	0.0084	0.0082

Experimental Analysis

To determine the values of parameters α and N, we apply $\alpha = 1, 3, 5, 7, 9$ and $N = 10, 15, 20, 25$ into the model to test the performance of the model. In this section, the data from January 1, 2014, to January 30, 2014, in Wuhan are used as training data, and the load data on January 31, 2014, are regarded as test data. Finally, the anti-normalization processing is carried out, and the NRMSE, MAE, and MAPE are calculated.

Table 3 shows the prediction errors for different α and N. We can see that whatever value N takes, three types of errors are obviously higher than others when $\alpha = 1$. When $\alpha = 3$ and

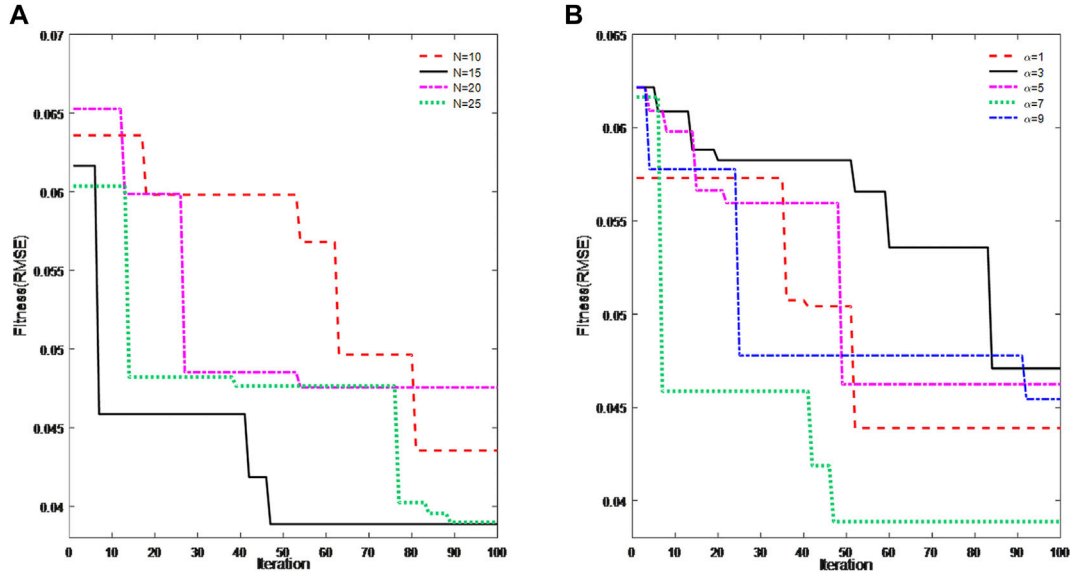


FIGURE 4 | Fitness of variation with the increase of iterations: **(A)** $\alpha = 7$; **(B)** $N = 15$.

TABLE 4 | Forecast results of power load.

Time	January			February			March			April		
	Actual values (MW)	Forecast values (MW)	Relative errors (%)	Actual values (MW)	Forecast values (MW)	Relative errors (%)	Actual values (MW)	Forecast values (MW)	Relative errors (%)	Actual values (MW)	Forecast values (MW)	Relative errors (%)
1:00	677.40	689.79	0.0183	709.82	720.35	0.0148	692.63	690.92	0.0025	658.30	657.68	0.0009
2:00	653.06	672.15	0.0292	694.35	685.74	0.0124	680.72	675.48	0.0077	647.19	645.88	0.0020
3:00	653.51	658.18	0.0071	670.91	672.69	0.0027	673.70	673.74	0.0001	637.70	637.16	0.0009
4:00	642.52	650.65	0.0127	677.20	661.16	0.0237	670.04	670.97	0.0014	630.18	644.47	0.0227
5:00	639.44	641.16	0.0027	662.21	660.71	0.0023	671.82	671.71	0.0002	637.72	643.55	0.0091
6:00	653.05	658.27	0.0080	684.66	687.16	0.0036	698.88	686.72	0.0174	676.44	667.54	0.0132
7:00	690.42	695.86	0.0079	731.70	729.71	0.0027	745.87	747.82	0.0026	722.95	717.03	0.0082
8:00	803.07	812.36	0.0116	814.08	814.16	0.0001	795.26	812.49	0.0217	751.95	762.93	0.0146
9:00	874.77	873.36	0.0016	889.29	889.97	0.0008	833.28	835.70	0.0029	788.23	791.43	0.0041
10:00	853.36	853.67	0.0004	885.28	888.98	0.0042	814.67	809.35	0.0065	777.08	776.24	0.0011
11:00	851.10	850.23	0.0010	884.13	880.51	0.0041	803.69	804.61	0.0011	767.25	769.31	0.0027
12:00	813.58	811.36	0.0027	848.45	846.00	0.0029	783.17	769.67	0.0172	745.26	740.57	0.0063
13:00	809.11	808.87	0.0003	831.73	828.58	0.0038	786.08	771.60	0.0184	730.55	722.46	0.0111
14:00	788.90	799.90	0.0139	818.47	822.24	0.0046	777.79	779.13	0.0017	742.14	731.85	0.0139
15:00	790.56	800.82	0.0130	826.27	826.70	0.0005	780.92	779.61	0.0017	739.81	742.17	0.0032
16:00	797.75	807.82	0.0126	824.00	830.17	0.0075	788.51	784.73	0.0048	753.58	737.04	0.0219
17:00	826.79	825.41	0.0017	856.85	847.88	0.0105	797.58	803.80	0.0078	771.64	770.39	0.0016
18:00	859.40	857.91	0.0017	859.35	865.93	0.0077	787.30	787.15	0.0002	748.37	753.00	0.0062
19:00	899.44	894.48	0.0055	896.94	899.66	0.0030	814.22	807.29	0.0085	756.61	767.41	0.0143
20:00	894.75	904.78	0.0112	909.69	911.58	0.0021	804.24	809.36	0.0064	760.59	770.73	0.0133
21:00	871.66	880.29	0.0099	887.56	883.48	0.0046	779.43	786.43	0.0090	728.23	741.39	0.0181
22:00	828.78	838.28	0.0115	824.59	828.39	0.0046	760.20	767.66	0.0098	707.18	717.41	0.0145
23:00	776.42	788.16	0.0151	760.49	764.49	0.0053	739.43	751.64	0.0165	696.21	707.82	0.0167

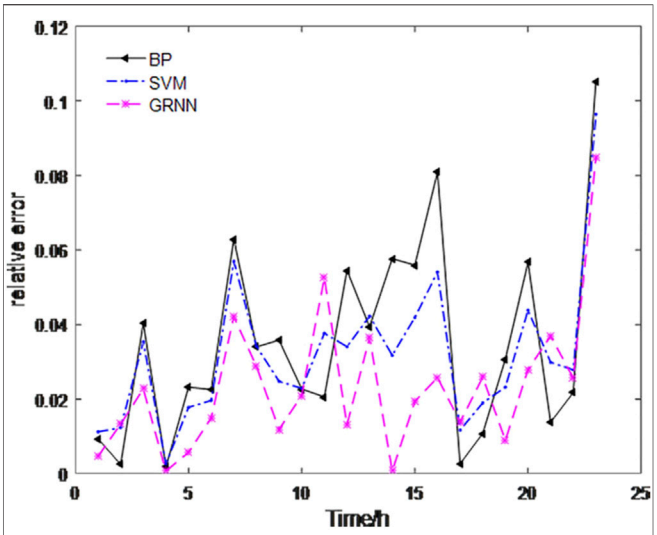


FIGURE 5 | Relative errors of the single models.

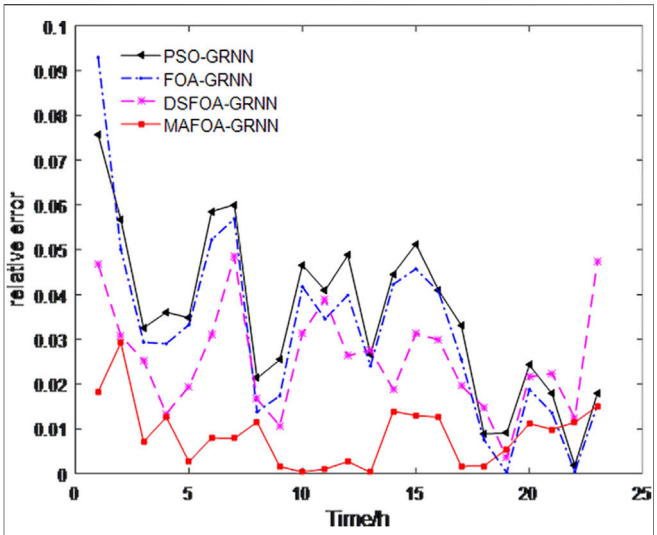


FIGURE 6 | Relative errors of the hybrid models.

$N = 25$, the NRMSE and MAE are small. The NRMSE, MAE, and MAPE are small at the same time when $\alpha = 5$ and $N = 20$, which is consistent with that reported in the FFOA proposed by Yu et al. (2016). When $\alpha = 9$ and $N = 20$, the NRMSE and MAPE are both small. But the smallest values of NRMSE, MAE, and MAPE are obtained when $\alpha = 7$ and $N = 15$, which means the forecasting performance is the best at this moment. So, we can initially claim that when $\alpha = 7$ and $N = 15$, the step size can realize the balance of global search capability and local optimization ability. **Figure 4** shows the fitness curves of MAFOA when $\alpha = 7$ and $N = 15$.

Figure 4 provides the fitness variation with the increase of iteration number. **Figure 4A** shows the convergence situation of different N when $\alpha = 7$, and **Figure 4B** shows the convergence performance of different α when $N = 15$. It can be seen that, under the condition of $\alpha = 7$, when we choose $N = 15$, the algorithm has the minimal fitness value and arrives at its optimal value much more quickly than other conditions; under the condition of $N = 15$, when we choose $\alpha = 7$, the algorithm has the same performance. Accordingly, $\alpha = 7$ and $N = 15$ are perceived as an ideal choice in the step size formula.

After choosing $\alpha = 7$ and $N = 15$, we train the power load data of each month except the last day to predict the load of the last day of each month. **Table 4** shows the prediction results obtained by the MAFOA-GRNN algorithm from January to April 2014, respectively. The relative errors are basically within 2%, and the accuracy is high.

In order to test the forecasting performance of the proposed model, the backpropagation (BP) neural network, support vector machine (SVM), GRNN, PSO-GRNN, FOA-GRNN, and DSFOA-GRNN are regarded as benchmark models to be compared with MAFOA-GRNN in short-term power load forecasting. The PSO was proposed by Kennedy and Eberhart in 1995, which was inspired by the swarm behavior of birds. The FOA proposed by Pan in 2012 was also used in this work. Since PSO and FOA are both classical optimization algorithms that have been widely utilized in research, we have chosen PSO-GRNN and FOA-GRNN as benchmark models. The DSFOA proposed by Hu et al. in 2017 is an improvement algorithm of FOA. With the decreasing step size in mind, the DSFOA performed well in optimizing the spread parameter of GRNN. The flight distance is updated referring to the sigmoid function. So, DSFOA-GRNN has also been compared with our proposed model. Besides, some other basic prediction models are also taken into account, such as the BP neural network and SVM. **Figure 5** shows the relative error curves of the single models on January 31. **Figure 6** shows the relative error curves of the hybrid models on January 31.

It can be seen from **Figure 5** that, in the commonly applied forecasting methods, the GRNN has the best prediction ability. **Figure 6** shows that the proposed method can accurately predict the overall trend of power load, and the fitting effect is very good. From the relative error curves, it can be seen that MAFOA-GRNN can offer a better predicting performance and higher precision than DSFOA-GRNN, FOA-GRNN, and GRNN. In addition, the relative errors of MAFOA-GRNN are more stable, and the majority are below 0.02, which demonstrates that the improved FOA is perceived as an ideal method in optimizing model parameters during GRNN training.

Then, the anti-normalization processing is carried out, and the comparison results of NRMSE, MAE, and MAPE evaluation criteria are shown in **Figures 7–9**. **Table 5** shows the error analysis of the training set and test set.

Obviously, MAFOA-GRNN has the smallest NRMSE, MAE, and MAPE, followed by FOA-GRNN, but the BP neural network has the worst performance. Besides, the prediction error of the training set and test set has no obvious difference, which indicates that MAFOA-GRNN

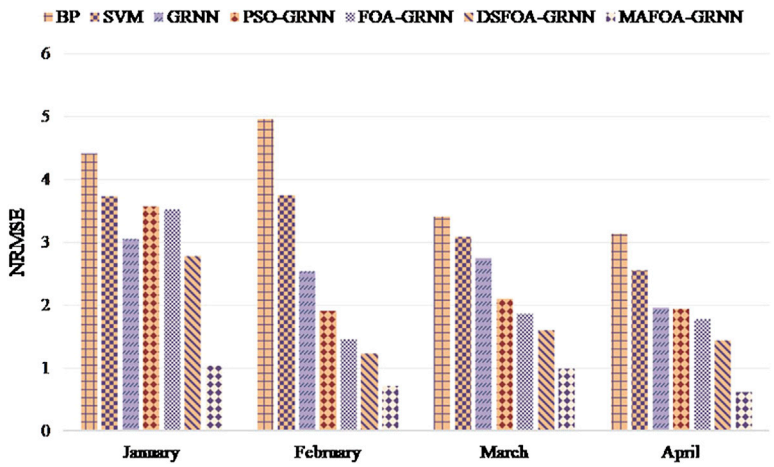


FIGURE 7 | Performance comparison of models in terms of NRMSE criteria.

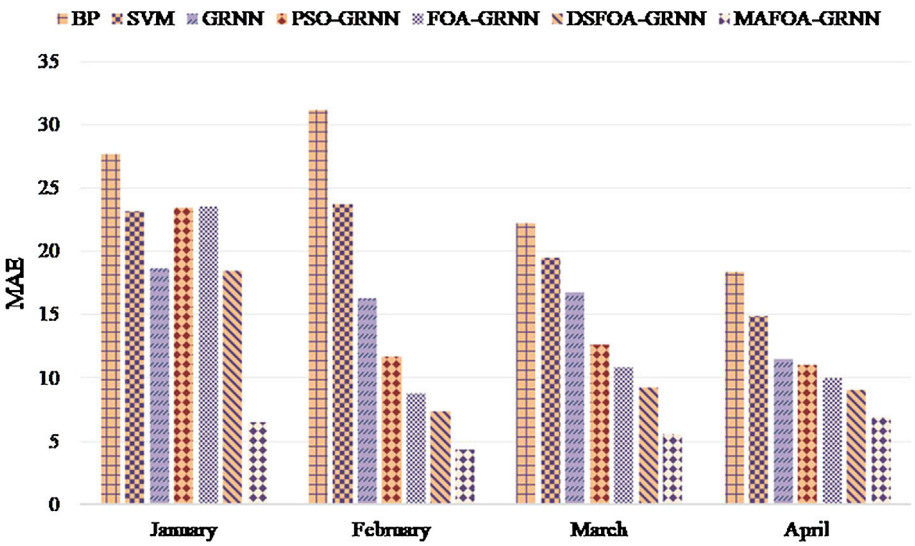


FIGURE 8 | Performance comparison of models in terms of MAE criteria.

has high generalization performance. According to the comparison results, it can be concluded that MAFOA-GRNN outperforms other models in both accuracy and stability. **Table 5** demonstrates the same conclusions as above.

Although the NRMSE, MAE, and MAPE can be used as criteria to obtain model-predicted loss values, it cannot be verified whether the comparison result is statistically significant. To statistically compare the differences between the prediction accuracy of different models, the DM statistics test is carried out in this paper, and the results are shown in **Table 6**. For all the benchmark models, the values of the MAFOA-GRNN model proposed in this paper are below 0.05, which indicates that the predictive ability of the MAFOA-GRNN model is better than that of DSFOA-GRNN, DSFOA-GRNN, GRNN, SVM, and BP neural network under the confidence interval of 95%.

According to the above comparisons, the following three main conclusions can be summarized:

1) The proposed MAFOA-GRNN outperforms the GRNN, which indicates that the MAFOA can optimize the smoothing parameter of GRNN effectively.

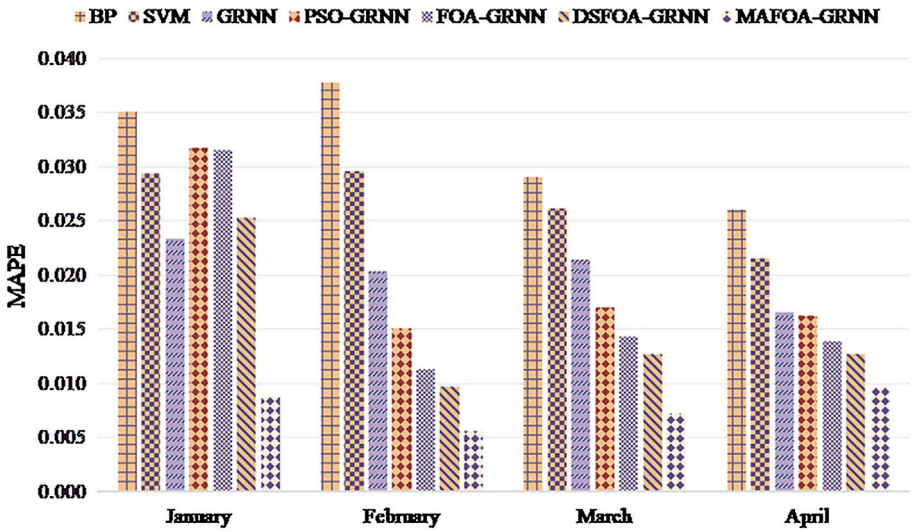


FIGURE 9 | Performance comparison of models in terms of MAPE criteria.

TABLE 5 | Error comparison between MAFOA-GRNN and the benchmark models.

Month	Error type	BP	SVM	GRNN	PSO-GRNN	FOA-GRNN	DSFOA-GRNN	MAFOA-GRNN
January	NRMSE	4.4106	3.7301	3.0477	3.5618	3.5189	2.7762	1.0415
	MAE	27.6808	23.1618	18.6411	23.4856	23.5662	18.4538	6.5242
	MAPE	0.0351	0.0293	0.0234	0.0317	0.0315	0.0253	0.0087
February	NRMSE	4.9501	3.7416	2.5321	1.9097	1.4507	1.2274	0.7050
	MAE	31.1464	23.7110	16.2740	11.7292	8.7336	7.4065	4.3071
	MAPE	0.0378	0.0296	0.0204	0.0151	0.0113	0.0096	0.0056
March	NRMSE	3.4144	3.0825	2.7506	2.1027	1.8578	1.5972	0.9859
	MAE	22.2451	19.4858	16.7246	12.6088	10.8888	9.2868	5.5455
	MAPE	0.0291	0.0262	0.0214	0.0171	0.0143	0.0127	0.0072
April	NRMSE	3.1352	2.5490	1.9609	1.9380	1.7843	1.4344	0.6178
	MAE	18.3732	14.9324	11.4907	10.9833	9.9950	9.0616	6.8816
	MAPE	0.0260	0.0216	0.0166	0.0163	0.0139	0.0128	0.0096

TABLE 6 | DM results of the different models.

Tested model	Benchmark model (January)					
	DSFOA-GRNN	FOA-GRNN	PSO-GRNN	GRNN	SVM	BP
MAFOA-GRNN	2.4083 (0.0036)	4.9502 (0.0120)	−2.5780 (0.0073)	2.8360 (0.0048)	−2.5430 (0.0160)	−3.4426 (0.0012)
DSFOA-GRNN		3.8311 (0.0280)	−3.7578 (0.0260)	4.1183 (0.0000)	−3.5324 (0.0180)	4.3981 (0.1802)
FOA-GRNN			−1.1790 (0.0370)	0.5811 (0.7165)	−3.5824 (0.0000)	−1.5424 (0.0686)
PSO-GRNN				4.4083 (0.0480)	−1.8246 (0.0190)	3.2784 (0.0735)
GRNN					−2.4083 (0.0370)	−2.6555 (0.0072)
SVM						4.1736 (0.0078)

2) The performance of MAFOA-GRNN is better than that of FOA-GRNN, which shows that the multivariate adaptive step can effectively improve the optimization ability of FOA.

3) From January to April 2014, MAFOA-GRNN has reached high prediction accuracy, which shows that the proposed algorithm is a stable and effective forecasting framework.

CONCLUSION

In this paper, we have proposed MAFOA-GRNN and applied it to short-term load forecasting. Firstly, we discussed a number of external factors including weather types and date types as input variables of the GRNN, in order to optimize the structure of NNs. Then, we propose an efficient interval segmentation technique for temperature types and weather types. Finally, we use the MAFOA to obtain the optimal GRNN model instead of the ordinary FOA, which solves the problem of local optimum in the implementation of FOA. The hybrid model proposed in this paper has a higher accuracy than the BP neural network, SVM, GRNN, PSO-GRNN, FOA-GRNN, and DSFOA-GRNN, and the majority of relative errors are below 0.02.

The proposed models can accurately predict the load of the power system, especially in short-term load forecasting. Electric energy cannot be stored in large quantities, and its generation and consumption are almost completed at the same time. Therefore, in order to arrange the work of power plants economically and reasonably, short-term load forecasting is indispensable. Furthermore, the proposed model can also predict other time series by adjusting the input vector and parameters.

In addition to short-term load forecasting, the proposed MAFOA-GRNN can be applied to solve other complex multivariable problems, including solar radiation forecasting, crude oil price forecasting, and wind load forecasting. Furthermore, the factors considered in this article are limited, and the forecasting performance may be better if other valuable factors are taken into consideration. Finally, further research may improve the performance of proposed model such as training the data of weekdays and holidays separately.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

FJ contributed to conception and design of the study. WZ and ZP conducted the data analysis, wrote the original draft, and edited the paper. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China, under grant number 61773401.

ACKNOWLEDGMENTS

The authors are grateful to the reviewers and the editor for their constructive comments and suggestions for this paper.

REFERENCES

Abedinia, O., and Amjady, N. (2016). Short-term Load Forecast of Electrical Power System by Radial Basis Function Neural Network and New Stochastic Search Algorithm. *Int. Trans. Electr. Energ. Syst.* 26 (7), 1511–1525. doi:10.1002/etep.2160

Agarkar, P., Hajare, P., and Bawane, N. (2016). "Optimization of Generalized Regression Neural Networks Using PSO and GA for Non-performer Particles," in Proceeding of the 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, May 2016 (IEEE), 103–107. doi:10.1109/rteict.2016.7807792

Cao, G., and Wu, L. (2016). Support Vector Regression with Fruit Fly Optimization Algorithm for Seasonal Electricity Consumption Forecasting. *Energy* 115, 734–745. doi:10.1016/j.energy.2016.09.065

Das Goswami, A., Mishra, M., and Patra, D. (2017). Investigation of General Regression Neural Network Architecture for Grade Estimation of an Indian Iron Ore deposit. *Arab J. Geosci.* 10 (4) 80. doi:10.1007/s12517-017-2868-5

Diebold, F., and Mariano, R. (1994). Comparing Predictive Accuracy. *J. Business Econ. Stat.* 20 (1), 134–144. doi:10.3386/t0169

Ding, N., Benoit, C., Foggia, G., Bésanger, Y., and Wurtz, F. (2016). Neural Network-Based Model Design for Short-Term Load Forecast in Distribution Systems. *IEEE Trans. Power Syst.* 31 (1), 72–81. doi:10.1109/TPWRS.2015.2390132

Du, P., Wang, J., Yang, W., and Niu, T. (2019). A Novel Hybrid Model for Short-Term Wind Power Forecasting. *Appl. Soft Comput.* 80, 93–106. doi:10.1016/j.asoc.2019.03.035

Dudek, G. (2016). Pattern-based Local Linear Regression Models for Short-Term Load Forecasting. *Electric Power Syst. Res.* 130, 139–147. doi:10.1016/j.epsr.2015.09.001

Friedrich, L., and Afshari, A. (2015). Short-term Forecasting of the Abu Dhabi Electricity Load Using Multiple Weather Variables. *Energ. Proced.* 75, 3014–3026. doi:10.1016/j.egypro.2015.07.616

Gao, Z., and Chen, L. (2015). "Sea Clutter Sequences Regression Prediction Based on PSO-GRNN Method," in Proceeding of the 2015 8th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, Dec. 2015 (IEEE), 72–75. doi:10.1109/iscid.2015.249

Hobbs, B. F., Jitprapaikulsarn, S., Konda, S., Chankong, V., Loparo, K. A., and Maratukulam, D. J. (1999). Analysis of the Value for Unit Commitment of Improved Load Forecasts. *IEEE Trans. Power Syst.* 14 (4), 1342–1348. doi:10.1109/59.801894

Hu, R., Wen, S., Zeng, Z., and Huang, T. (2017). A Short-Term Power Load Forecasting Model Based on the Generalized Regression Neural Network with Decreasing Step Fruit Fly Optimization Algorithm. *Neurocomputing* 221, 24–31. doi:10.1016/j.neucom.2016.09.027

Jiang, F., Yang, H., and Shen, Y. (2014). On the Robustness of Global Exponential Stability for Hybrid Neural Networks with Noise and Delay Perturbations. *Neural Comput. Applic* 24 (7), 1497–1504. doi:10.1007/s00521-013-1374-2

Jiang, P., and Chen, J. (2016). Displacement Prediction of Landslide Based on Generalized Regression Neural Networks with K-fold Cross-Validation. *Neurocomputing* 198, 40–47. doi:10.1016/j.neucom.2015.08.118

Jiang, W., Wu, X., Gong, Y., Yu, W., and Zhong, X. (2020). Holt-Winters Smoothing Enhanced by Fruit Fly Optimization Algorithm to Forecast Monthly Electricity Consumption. *Energy* 193, 116779. doi:10.1016/j.energy.2019.116779

Jianzhou Wang, J., Yang, W., Du, P., and Li, Y. (2018). Research and Application of a Hybrid Forecasting Framework Based on Multi-Objective Optimization for Electrical Power System. *Energy* 148, 59–78. doi:10.1016/j.energy.2018.01.112

Kumar, G., and Malik, H. (2016). Generalized Regression Neural Network Based Wind Speed Prediction Model for Western Region of India. *Proced. Comp. Sci.* 93, 26–32. doi:10.1016/j.procs.2016.07.177

Li, J., Duan, P., Sang, H., Wang, S., Liu, Z., and Duan, P. (2018). An Efficient Optimization Algorithm for Resource-Constrained Steelmaking Scheduling Problems. *IEEE Access* 6, 33883–33894. doi:10.1109/ACCESS.2018.2840512

Liu, D., Zhu, S., and Sun, K. (2019). Global Anti-synchronization of Complex-Valued Memristive Neural Networks with Time Delays. *IEEE Trans. Cybern.* 49 (5), 1735–1747. doi:10.1109/TCYB.2018.2812708

Lu, Y., Zhang, T., Zeng, Z., and Loo, J. (2016). "An Improved RBF Neural Network for Short-Term Load Forecast in Smart Grids," in Proceeding of the 2016 IEEE International Conference on Communication Systems (ICCS), Shenzhen, China, Dec. 2016 (IEEE), 1–6. doi:10.1109/iccs.2016.7833643

Meng, T., and Pan, Q.-K. (2017). An Improved Fruit Fly Optimization Algorithm for Solving the Multidimensional Knapsack Problem. *Appl. Soft Comput.* 50, 79–93. doi:10.1016/j.asoc.2016.11.023

Mitić, M., Vuković, N., Petrović, M., and Miljković, Z. (2015). Chaotic Fruit Fly Optimization Algorithm. *Knowledge-Based Syst.* 89, 446–458. doi:10.1016/j.knosys.2015.08.010

Ozturk, A. U., and Turan, M. E. (2012). Prediction of Effects of Microstructural Phases Using Generalized Regression Neural Network. *Construction Building Mater.* 29, 279–283. doi:10.1016/j.conbuildmat.2011.10.015

Pan, Q.-K., Sang, H.-Y., Duan, J.-H., and Gao, L. (2014). An Improved Fruit Fly Optimization Algorithm for Continuous Function Optimization Problems. *Knowledge-Based Syst.* 62, 69–83. doi:10.1016/j.knosys.2014.02.021

Pan, W.-T. (2012). A New Fruit Fly Optimization Algorithm: Taking the Financial Distress Model as an Example. *Knowledge-Based Syst.* 26, 69–74. doi:10.1016/j.knosys.2011.07.001

Samadianfard, S., Jarhan, S., Salwana, E., Mosavi, A., Shamshirband, S., and Akib, S. (2019). Support Vector Regression Integrated with Fruit Fly Optimization Algorithm for River Flow Forecasting in Lake Urmia Basin. *Water* 11 (9), 1934. doi:10.3390/w11091934

Sammen, S. S., Mohamed, T. A., Ghazali, A. H., El-Shafie, A. H., and Sidek, L. M. (2017). Generalized Regression Neural Network for Prediction of Peak Outflow from Dam Breach. *Water Resour. Manage.* 31 (1), 549–562. doi:10.1007/s11269-016-1547-8

Specht, D. F. (1991). A General Regression Neural Network. *IEEE Trans. Neural Netw.* 2 (6), 568–576. doi:10.1109/72.97934

Taghi Sattari, M., Feizi, H., Samadianfard, S., Falsafian, K., and Salwana, E. (2021). Estimation of Monthly and Seasonal Precipitation: A Comparative Study Using Data-Driven Methods versus Hybrid Approach. *Measurement* 173, 108512. doi:10.1016/j.measurement.2020.108512

Talele, K., Shirsat, A., Uplenchwar, T., and Tuckley, K. (2016). "Facial Expression Recognition Using General Regression Neural Network," in Proceeding of the 2016 IEEE Bombay Section Symposium (IBSS), Baramati, India, Dec. 2016 (IEEE), 1–6. doi:10.1109/ibss.2016.7940203

Wang, L., Shi, Y., and Liu, S. (2015). An Improved Fruit Fly Optimization Algorithm and its Application to Joint Replenishment Problems. *Expert Syst. Appl.* 42 (9), 4310–4323. doi:10.1016/j.eswa.2015.01.048

Wang, J., Yang, W., Du, P., and Niu, T. (2020). Outlier-robust Hybrid Electricity price Forecasting Model for Electricity Market Management. *J. Clean. Prod.* 249, 119318. doi:10.1016/j.jclepro.2019.119318

Xuan, Y., Si, W., Zhu, J., Sun, Z., Zhao, J., Xu, M., et al. (2021). Multi-Model Fusion Short-Term Load Forecasting Based on Random Forest Feature Selection and Hybrid Neural Network. *Ieee Access* 9, 69002–69009. doi:10.1109/access.2021.3051337

Yang, W., Wang, J., and Wang, R. (2017). Research and Application of a Novel Hybrid Model Based on Data Selection and Artificial Intelligence Algorithm for Short Term Load Forecasting. *Entropy* 19 (2), 52. doi:10.3390/e19020052

Yang, W., Sun, S., Hao, Y., and Wang, S. (2022). A Novel Machine Learning-Based Electricity price Forecasting Model Based on Optimal Model Selection Strategy. *Energy* 238, 121989. doi:10.1016/j.energy.2021.121989

Yu, Y., Li, Y., Li, J., and Gu, X. (2016). Self-adaptive Step Fruit Fly Algorithm Optimized Support Vector Regression Model for Dynamic Response Prediction of Magnetorheological Elastomer Base Isolator. *Neurocomputing* 211, 41–52. doi:10.1016/j.neucom.2016.02.074

Zhang, Z., and Hong, W.-C. (2019). Electric Load Forecasting by Complete Ensemble Empirical Mode Decomposition Adaptive Noise and Support Vector Regression with Quantum-Based Dragonfly Algorithm. *Nonlinear Dyn.* 98 (2), 1107–1136. doi:10.1007/s11071-019-05252-7

Zhao, M., Ji, S., and Wei, Z. (2020). Risk Prediction and Risk Factor Analysis of Urban Logistics to Public Security Based on PSO-GRNN Algorithm. *Plos One* 15 (10), e0238443. doi:10.1371/journal.pone.0238443

Zhen Wang, Z., Yin, Q., Guo, P., and Zheng, X. (2018). "Flux Extraction Based on General Regression Neural Network for Two-Dimensional Spectral Image," in *International Conference on Human-Computer Interaction* (Springer), 219–226. doi:10.1007/978-3-319-92270-6_30

Zheng, X.-l., and Wang, L. (2016). A Knowledge-Guided Fruit Fly Optimization Algorithm for Dual Resource Constrained Flexible Job-Shop Scheduling Problem. *Int. J. Prod. Res.* 54 (18), 5554–5566. doi:10.1080/00207543.2016.1170226

Zhu, S., Lian, X., Wei, L., Che, J., Shen, X., Yang, L., et al. (2018). PM2.5 Forecasting Using SVR with PSOGSA Algorithm Based on CEEMD, GRNN and GCA Considering Meteorological Factors. *Atmos. Environ.* 183, 20–32. doi:10.1016/j.atmosenv.2018.04.004

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.

Copyright © 2022 Jiang, Zhang and Peng. 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 Environmental Science

ORIGINAL RESEARCH
published: 06 April 2022
doi: 10.3389/fenvs.2022.883711

Forecasting Energy Consumption Based on SVR and Markov Model: A Case Study of China

Zhaosu Meng[1,2]*, *Huike Sun*[1] and *Xi Wang*[1]

[1]School of Economics, Ocean University of China, Qingdao, China, [2]Institute of Marine Development of Ocean University of China, Qingdao, China

Forecasting energy demand in emerging nations is a critical policy tool utilized by decision makers worldwide. However, as estimated economic and demographic characteristics frequently diverge from realizations, precise forecast results are difficult to get due to the economic system's intrinsic complexity. This work proposed a machine learning model for estimating energy consumption in China using the support vector regression model (SVR). Additionally, Markov Chain (MC) is employed to forecast and analyze the evolving energy consumption structure. The results demonstrate that SVR model is more accurate (98.4%) than the linear model (Moving Average model), the nonlinear model (Grey model), and past research in predicting energy usage. Under the current rate of energy consumption, China's total energy consumption will break through six billion in the next 4 years. Furthermore, it is expected that China's energy consumption structure will be more rational in 2025, with increased non-fossil energy consumption and decreased coal consumption, while natural gas consumption continues to grow at a low rate. It provides scientific basis for the implementation of carbon emission peak action, energy security and energy development plan during the 14th Five-Year Plan period.

Keywords: forecast, energy consumption, machine learning, sustainable development, energy structure

OPEN ACCESS

Edited by:
Wendong Yang,
Shandong University of Finance and Economics, China

Reviewed by:
Kevin Li,
University of Windsor, Canada
Lifeng Wu,
Hebei University of Engineering, China

***Correspondence:**
Zhaosu Meng
zhaosu.meng@ouc.edu.cn

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: *25 February 2022*
Accepted: *21 March 2022*
Published: *06 April 2022*

Citation:
Meng Z, Sun H and Wang X (2022) Forecasting Energy Consumption Based on SVR and Markov Model: A Case Study of China. Front. Environ. Sci. 10:883711. doi: 10.3389/fenvs.2022.883711

1 INTRODUCTION

One of the most essential policy instruments utilized by decision makers worldwide is energy consumption predictions. This is especially true in rising energy markets like China. According to the National Bureau of Statistics, in 2021, China's gross domestic product (GDP) was RMB 114.4 trillion ($17.7 trillion), up around RMB 13 trillion (United States $3 trillion) from 2020[1]. With an economy anticipated to develop at a rate of 6–8% for decades, China's influence in the global energy market is growing.

Between 1990 and 2020, China's overall energy output and consumption increased steadily (**Figure 1**). Energy production has gradually increased from about 1.26 billion tons of ordinary coal to 4.08 billion tons, but it is still unable to satisfy energy demand, and the gap between the two is widening. China's total energy imports in 2020 is 1.20 billion tons, and this figure will rise further.[1] As the biggest energy consumer of the world, China's underlying demand and supply imbalances will have a significant impact on global energy markets. Long-term estimates are also required to determine the extent to which future trade and investment plans are required to secure China's energy security. Furthermore, in the General Debate of the 75th Session of the United Nations

[1]China Statistical Yearbook 2021, http://www.stats.gov.cn/tjsj/ndsj/

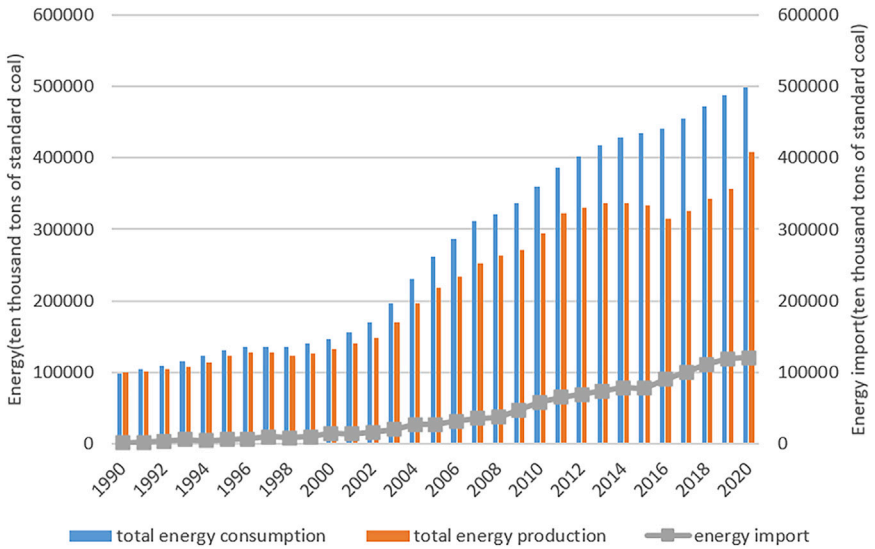


FIGURE 1 | Total energy consumption and production in China from 1999 to 2020. Data source: China Statistical Yearbook 2021.

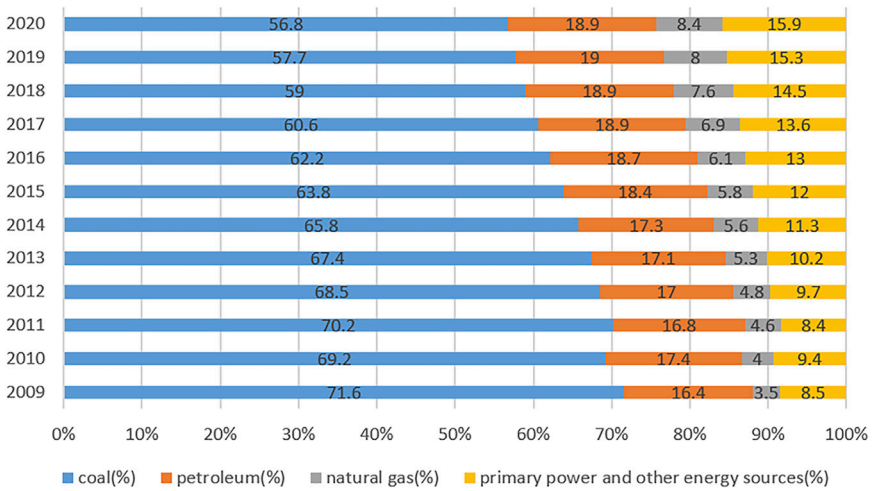


FIGURE 2 | Proportion of energy consumption in China from 2009 to 2020. Data source: China Statistical Yearbook 2021.

General Assembly in 2020, China has pledged to peaking carbon emissions by 2030 and reaching carbon neutrality by 2060, while also aiming to double the size of the Chinese economy by 2035. As a result, forecasting energy use in China is crucial.

China's energy use is imbalanced (**Figure 2**). This is shown by coal's considerable domestic market share in China, despite its waning dominance in recent years (dropped to 56.8% in 2020)[1]. Natural gas is also insignificant. As a result of this predicament, the economic cost of environmental damage associated with excessive coal consumption has been substantial. As a result, the energy industry and policymakers are under intense pressure to modify the structure of energy use and shift away from coal and toward cleaner options such as natural gas and hydropower.

As illustrated in **Figure 3**, in 2019, the total urban consumption was 26,665 ten thousand tons of standard coal, 3.89 times that of 1997 (6,845.89 ten thousand tons of standard coal)[2]. The total energy consumption of rural residents is also growing rapidly, but not as fast as it is in the urban area. Since 2000, the gap between urban and rural enlarged. Therefore, with rapid urbanization, the energy consumption in China shows new characteristics. Demand for coal, oil, natural gas, and other fossil fuels is expanding at a rapid pace, while consumption of coal is declining. However, the energy gap continues to widen.

[2]China Statistical Yearbook 2021, http://www.stats.gov.cn/tjsj/ndsj/

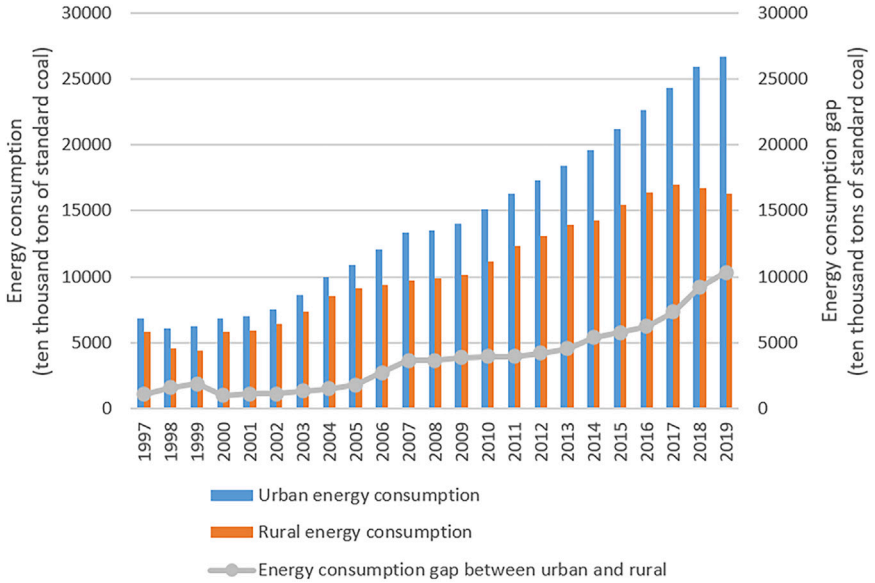


FIGURE 3 | Changes in household energy consumption from 1997 to 2019. Data source: China Energy Statistical Yearbook.

While China's energy conservation and emission reduction policies have had some initial success, the future situation remains unpredictable. Conducting a thorough examination of future energy forecasts will not only aid in comprehending the future energy situation, but also in providing scientific support for overall energy planning and policy creation. To meet rising consumer demand and achieve green and sustainable growth, energy consumption structure should be optimized, with less coal consumption, and more clean energy. And better data interpretation and forecasting techniques are required.

Due to their applicability, machine learning models are being utilized in a wide number of sectors, and their operation is similar to that of a function that best maps input data to output. Using machine learning methods, energy consumption forecasts can be created with high accuracy. As a result, governments can use them to undertake energy-saving efforts. They can be used to forecast future energy usage, such as electricity or natural gas.

The goal of this research is to develop a unique forecasting system that utilizes nonlinear machine learning models to optimize linear time series in order to discover energy consumption patterns. Then, we assess the model prediction findings in conjunction with the actual scenario and offer countermeasures and suggestions in order to fulfill the goal of providing a foundation for policymakers to create policies more effectively. Numerous optimization techniques, including the SVR model, the linear model MA (3), and the nonlinear model GM (1, 1) are examined and compared for usage in the proposed forecasting system. By anticipating primary energy consumption and its structure, it will assist in the process of energy decarbonization, minimize the effects of climate change, and rein in China's future climate initiative. It is beneficial for policymakers to increase energy efficiency through energy policies and other measures, while reducing greenhouse gas emissions enables the energy structure to be adjusted to achieve co-benefits such as the efficient development of a green and low-carbon economy and climate change adaptation.

The remainder of this study is structured in the following manner. The second section covers the preceding research literature in two areas: variables influencing energy use and energy prediction models. The third section describes the influencing factors on energy consumption that were chosen for this study, as well as the methods, models, variables, and data sources employed in this study. The fourth section assesses SVR model's validity and prediction accuracy before estimating China's energy consumption for the time covered by the 14th Five-Year Plan. Then, we use Markov Chain to predict the change of energy structure during the 14th Five-Year Plan period. The fifth section discusses the findings of the investigation, as well as the accompanying countermeasures and recommendations.

2 LITERATURE REVIEW

Studying the relationship between numerous drivers and energy consumption is not only required for model predictions, but it also helps readers understand the mechanisms by which these factors affect energy consumption. Some scholars believe that economic development has a substantial impact on energy consumption intensity (Asafu-Adjaye, 2000; Stern, 2000; Soytas and Sari, 2003; Ang, 2004; Aboagye, 2017; Marques et al., 2019). According to Birol and Keppler (2000), technological advancements will not only reduce energy consumption intensity, but will also boost energy consumption per unit of GDP. Other researchers believe that improved technology can promote a decline in energy intensity (Li and Lin, 2014; Huang et al., 2017). Hunt and Ninomiya (2005)

discovered a negative relationship between energy prices and demand. According to Inglesi (2010), there is a long-term link between energy usage, prices, and economic growth. He et al. (2011) believed that the secondary industry was not conducive to improving energy efficiency, while the development of the tertiary industry could effectively encourage the decline of energy consumption intensity. Achour and Belloumi (2016) believe that in addition to economic growth and energy prices, population size also has a positive impact on energy consumption. Some scholars studied from the perspective of industrial structure. Lin and Zhu (2017) conducted additional research and found that industrial structure upgrading is adversely connected with energy consumption intensity. Later, from a micro perspective, some scholars believe that household variables such as family size and education level of family members have a significant impact on energy use (Zou and Luo, 2019).

Machine Learning (ML) is the basic approach to achieve artificial intelligence. In recent years, more scholars have paid attention to ML with its advantages of recognizing trends and patterns from data, not requiring human intervention, constantly optimizing cognition according to data changes, and processing various high-dimensional data. Hao et al. (2021) developed an economic loss analysis system to quantitatively evaluate the losses caused by haze pollution on tourism. Yang et al. (2022) developed an improved electricity price prediction model, which has the advantages of adaptive data preprocessing, etc.

Machine Learning has experienced a transition from shallow learning to deep learning. Rumelhart et al. (1986) proposed the idea of learning representation based on error back propagation, which enabled artificial neural network (ANN) to be trained. Subsequently, Support Vector Machine (SVM), Decision tree (DT) and other models that are easier to obtain global optimal solutions have been widely concerned and applied. With the progress of the data accumulation, the computing power and the improvement to the algorithm, deep and complex training to become possible. At present, both shallow learning model and deep learning model are widely used in various fields. To predict specific problems, it is necessary to select appropriate machine learning algorithms from the perspectives of data category, sample size and model characteristics.

Numerous methods for forecasting energy consumption have been proposed in the literature, such as the energy demand model (EDM), the Autoregressive Integrated Moving Average Model (ARIMA), the particle swarm optimization technique (PSO), the artificial neural network model (ANN), and the grey model (GM). Gori and Takanen (2004) used the modified EDM model to predict the energy consumption of Italian industry, household and service. Ediger and Akar (2007) used the ARIMA and seasonal ARIMA methods to anticipate Turkey's energy demand from 2005 through 2020. Some scholars forecasted Turkey's energy consumption using PSO methodologies (Uenler, 2008; Kran et al., 2012; Yakut and Özkan, 2020). Zong and Roper (2009) suggested an ANN model for estimating Korea's energy demand. Meng et al. (2020) estimate China's energy intensity using an upgraded DVCGM (1, N) model, and found that the improved DVCGM (1, N) model could reflect the lag effect of government policies. Zhu et al. (2020) developed an adaptive gray-scale weighted model to forecast Jiangsu's electricity usage. Chen et al. (2021) predicted energy consumption in the Beijing-Tianjin-Hebei region using the FGM (1, 1) model. At the same time, some scholars predict the structure of energy consumption. Xie et al. (2015) forecasted total energy production and consumption using an improved univariate discrete grey forecasting model and suggested a new Markov technique based on a quadratic programming model to predict the structure of energy production and consumption. Ren et al. (2017) used Markov model to predict the energy consumption structure of Beijing.

The benefits and drawbacks of the most commonly used energy forecast methods of the literature is outlined in **Table 1**. The energy consumption system is a nonlinear system with many influencing factors. Because most traditional prediction approaches lack a learning mechanism, it is difficult to describe the nonlinear relationship in the energy system, resulting in low forecast accuracy. Some new computer forecasting approaches are exceedingly complicated and subjective, jeopardizing the accuracy of energy consumption forecasting. Time series models are problematic due to a paucity of data on China's energy consumption. PSO model is easy to fall into local optimal solution, so it is difficult to obtain global optimal solution. If an ANN model is used to forecast, the performance of the prediction will be inconsistent due to insufficient sample training and over-learning. The GM model's prediction is based on exponential prediction. This approach works best with data that grows exponentially, which makes it unsuitable for calculating China's energy consumption. Therefore, we propose a Support Vector Regression (SVR) model to anticipate China's energy consumption during the 14th Five-Year Plan period, while the widely used Markov Chain (MC) is used to predict the energy structure.

3 METHODOLOGY

3.1 Model

3.1.1 Support Vector Regression

SVMs are supervised machine learning models used to tackle classification and regression problems. SVMs are composed of two components: a kernel and an optimizer method. The kernel converts non-linear data to a high-dimensional space and then linearly separates the data. Optimizer algorithm tackles the optimization problem, which is computationally costly. SVM usually outperforms other machine learning approaches in terms of generalization.

This section describes the data collection and the process of developing a prediction model using many variables. To begin, we identify the elements that influence energy use and preprocess the data. The SVR model is then used to create training and test sets randomly, and the accuracy is compared to that of the GM and MA models. Then, we make predictions using the SVR model. The framework for projecting energy use is displayed in **Figure 4**.

TABLE 1 | Comparison of prediction methods.

Model	Pros	Cons
ARIMA	· Good in short-term forecasting · It can show how data develops and changes	·Not suitable for long-term forecasting
PSO	· The calculation process is simple · Obtain the best prediction result through continuous iteration	·Easy to fall into local optimal solution, difficult to obtain global optimal solution
ANN	· Can process complex data · Simple modeling · Any nonlinear sequence can be approximated with any accuracy	·Easy to fall into local convergence ·Learning mechanism is difficult to acquire ·Parameters need to be adjusted continuously during the modeling process ·Time consuming
GM	· Suitable for dynamic process analysis · Dot demanding for sample data	·Modeling bias exists in the model itself ·The prediction accuracy will decrease with the increase of time span
SVR	· Applicable to small sample, nonlinear, high-dimensional, local minimum and other problems · Good prediction on limited data samples · Easy to use and doesn't have any "over-learning" faults · Avoids the influence of subjective factors, but also integrates more relevant factors	·Outliers are difficult to excavate

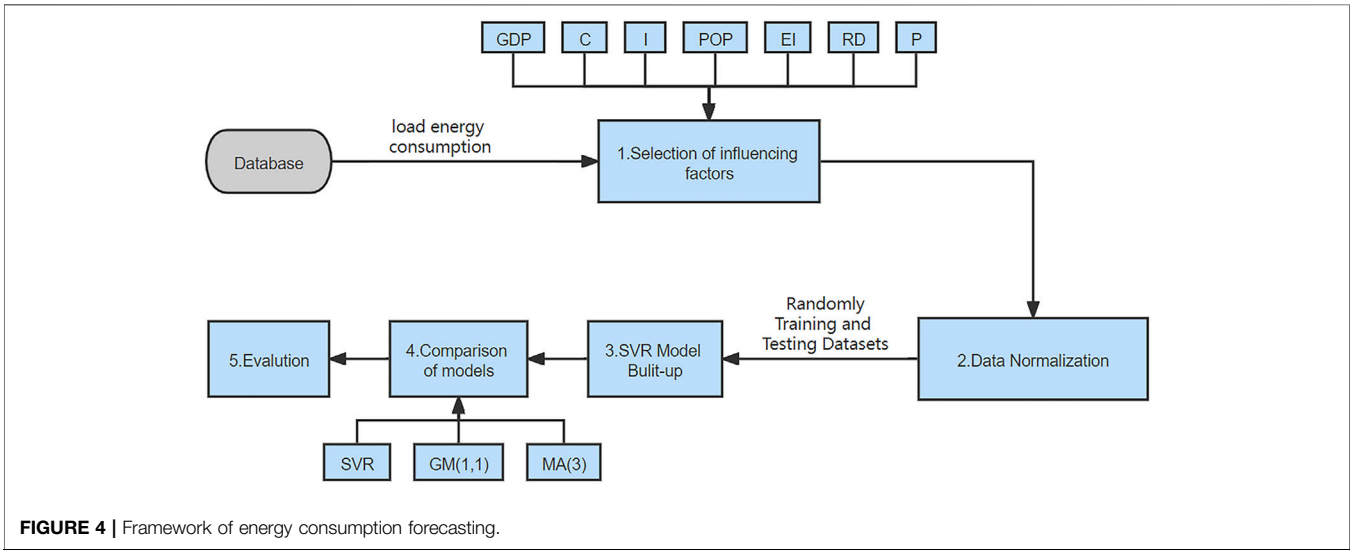


FIGURE 4 | Framework of energy consumption forecasting.

Based on the basic idea of support vector machine (SVM) model prediction (Cortes and Vapnik, 1995), input various energy demand influencing factors x_1, x_2..., x_d maps to a higher dimensional eigenspace $[\varphi(x_1), \varphi(x_2) \ldots, \varphi(x_n)]$. According to statistical learning theory, the original nonlinear model can be transformed into a linear regression model of feature space, as shown in **Eqn 1**

$$f\left(x_j\right) = \omega^T \varphi\left(x_j\right) + b \tag{1}$$

Wherein, $\varphi(x_j)$ is a kernel function, ω and b are the parameters to be identified in the model. The parameters to be identified in **Eqn 1** are processed using the idea of reducing structural risk, as shown in **Eqn 2**.

$$R_{emp}(f) = \sum_{i=1}^{s} C(e_i) + \lambda \parallel \omega \parallel^2 \tag{2}$$

Where, $R_{emp}(f)$ is empirical risk, $\parallel \omega \parallel^2$ is the confidence risk, $C(e_i)$ is the loss function. Further, according to the principle of SVM, solving **Eqn 2** is equivalent to solving the optimization problem in **Eqn 3**

$$\begin{aligned} &minL = \frac{1}{2}\omega^T \omega + C\sum_{i=1}^{s} \varepsilon_i^* + \varepsilon_i \\ &s.t. \begin{cases} y - (\omega, \varphi(X_t)) - b \le \varepsilon + \varepsilon_i^* \\ (\omega, \varphi(X_t)) + b - y \le \varepsilon + \varepsilon_i \\ \varepsilon_i^*, \varepsilon_i \ge 0 \end{cases} \end{aligned} \tag{3}$$

In order to facilitate the solution, **Eqn 3** is often transformed into a duality problem, then the nonlinear function $f(x)$ can be obtained.

$$f(x) = \sum_{i=1}^{l} (a_i - a_i^*) K(X_t, X) + b \tag{4}$$

Where, a_i and a_i^* are support vector parameters, $K(X_t, X)$ is the inner product. According to Mercer conditions, define kernel function and select radial basis kernel function (RBF)

$$K(X_t, X) = exp\left\{-\frac{\| x_j - x_v\|^2}{\sigma^2}\right\} \quad (5)$$

Substituting **Eqn 5** into **Eqs 4**, **6** can be obtained through equivalent transformation.

$$f(x) = \sum_{i=1}^{l} a_j exp\left\{-\frac{\| x_j - x_v\|^2}{\sigma^2}\right\} + b \quad (6)$$

Where, a_j is the parameter value corresponding to the support vector, x_j is the input data vector of training year. x_v is the input data vector for predicting years, *f(x)* is the set of output vectors. According to **Eqn. 6**, energy demand prediction parameters a_j and b are obtained, and the energy demand prediction model is estimated.

3.1.2 Markov Chain

Because the direction in the energy consumption structure is policy-driven and the process is gradual, the Markov Chain hypothesis is used (Niu et al., 2004). That is, in the process of event change, the state at m is determined by the state at m-*1*, and is not correlated with the state at m-i ($i \geq 2$). Conversely, to predict the state at $m + i$, it is necessary to calculate the state at m+*1* from the state at m, repeat the process and get the result.

We propose a time set X_t with n states, at time m, X_m is set as a, X_{m+1} is set as b. Then the conditional probability (p_{ab}) of the transition from the state at m to m+*1* is as follows:

$$p_{ab} = \{X_{m+1} = b | X_m = a\} \quad (7)$$

and the formula of the transition matrix p can be derived in **Eqn (8)**.

$$P = p_{ab} = \begin{bmatrix} p_{11} & \cdots & p_{1b} \\ \vdots & \ddots & \vdots \\ p_{a1} & \cdots & p_{ab} \end{bmatrix} \quad (8)$$

After the transfer matrix p is obtained, the energy consumption structure at m is known as *Per (m)*, then the energy consumption structure at $m + i$ is calculated as follows:

$$\begin{aligned} &Per(m) = \{Per_c(m), Per_o(m), Per_g(m), Per_e(m)\} \\ &Per(m+i) = Per(m)*P^i \end{aligned} \quad (9)$$

in which, $Per_c(m)$, $Per_o(m)$, $Per_g(m)$, $Per_e(m)$ respectively represent the proportions of coal, crude oil, natural gas and clean energy (hydro power, wind power, etc.) in the total primary energy consumption.

Energy consumers and producers may change their option on energy consumption. For example, if some energy customers select natural gas over coal this year, the proportion of coal consumption will fall and the proportion of natural gas will rise. The energy structure change is the calculated transfer probability from coal to natural gas, and the probability of each energy structures transferring to each other constitutes the transition probability matrix. The retention probability $p_{c\to c}(m)$ is calculated. The energy structure changes from m to $m + 1$, and the retention probability is the share at $m + 1$ divided by the share at m. The detailed calculation formula is shown in **10–12**.

Probability of coal switching to oil (**Eqn 10**):

$$p_{c\to o}(m) = \frac{[1 - p_{c\to c}(m)] * [Per_o(m+1) - Per_o(m)]}{[Per_o(m+1) - Per_o(m)] + [Per_g(m+1) - Per_g(m)] + [Per_e(m+1) - Per_e(m)]} \quad (10)$$

Probability of coal switching to natural gas (**Eqn 11**):

$$p_{c\to g}(m) = \frac{[1 - p_{c\to c}(m)] * [Per_g(m+1) - Per_g(m)]}{[Per_o(m+1) - Per_o(m)] + [Per_g(m+1) - Per_g(m)] + [Per_e(m+1) - Per_e(m)]} \quad (11)$$

Probability of coal turning to water and wind power (**Eqn 12**):

$$p_{c\to e}(m) = \frac{[1 - p_{c\to c}(m)] * [Per_e(m+1) - Per_e(m)]}{[Per_o(m+1) - Per_o(m)] + [Per_g(m+1) - Per_g(m)] + [Per_e(m+1) - Per_e(m)]} \quad (12)$$

Where, $Per_i(m)$ is the proportion of i energy in the total amount at moment m.

Suppose that the one-step transition probability matrix of China's primary energy consumption structure from moment m to moment $m + 1$ is:

$$P(m) = \begin{bmatrix} p_{c\to c}(m) & p_{c\to o}(m) & p_{c\to g}(m) & p_{c\to e}(m) \\ p_{o\to c}(m) & p_{o\to o}(m) & p_{o\to g}(m) & p_{o\to e}(m) \\ p_{g\to c}(m) & p_{g\to o}(m) & p_{g\to g}(m) & p_{g\to e}(m) \\ p_{e\to c}(m) & p_{e\to o}(m) & p_{e\to g}(m) & p_{e\to e}(m) \end{bmatrix} \quad (13)$$

3.2 Influencing Factors Preparation

A country's energy consumption (*EC*) is a nonlinear complicated system that is influenced by a large number of variables. GDP, energy price, secondary industry's fraction of added value, energy consumption intensity, technical progress, resident consumption level, and population size are used as dependent variables on energy consumption (Chen and Zhu, 2013; Marques et al., 2019), which can be expressed as:

$$EC = f(GDP, P, I, EI, RD, C, POP) \quad (14)$$

Where: *EC* denotes total energy consumption, *GDP* denotes gross domestic product, *p* is for energy price, *I* denotes the proportion of secondary industry, *EI* represents energy consumption intensity, *RD* is for technological progress, *C* represents household consumption level, and *POP* denotes total population size.

To avoid the impact of price considerations, we recalculate GDP at 1990 prices and determine energy consumption intensity per unit GDP. China's GDP increased by 9% yearly from 1990 to 2020, rising from 1,887.3 billion yuan to 10,1598.6 billion yuan[3]. Economic growth results in increased energy consumption, which continues to rise year after year. As a result, economic

[3]China Statistical Yearbook 2021, http://www.stats.gov.cn/tjsj/ndsj/

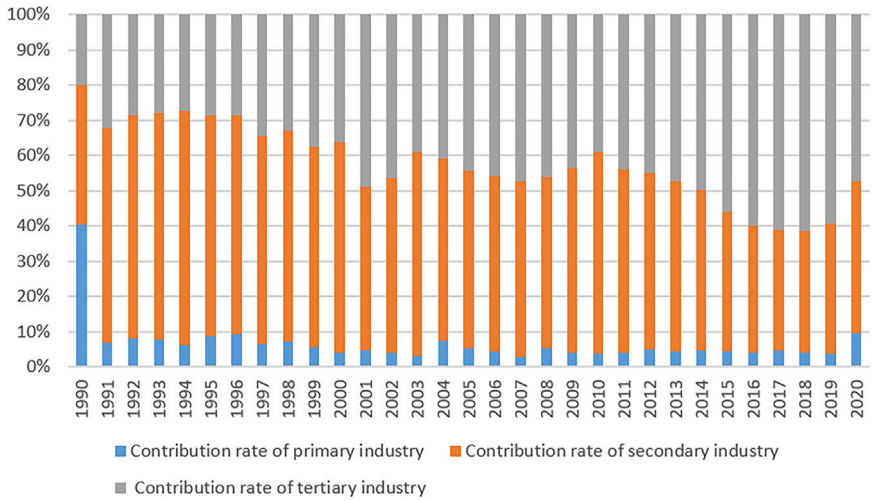


FIGURE 5 | China's industrial structure, 1990–2020. Data source: China Statistical Yearbook 2021.

development is the most important determinant of energy use consumption.

Population has a direct impact on energy consumption. By the end of 2019, China's overall population exceeded 1.4 billion.[3] Despite the fact that China's annual net increase has decreased below 10 million, showing that the country's population is stabilizing, China's energy consumption is enormous. The natural population growth rate will remain at about 4‰, which will continue to drive high demand for energy.[4] Additionally, there is a major discrepancy in the levels and ways of energy consumption between urban and rural inhabitants. With China's urbanization process accelerating, the growth in urban population necessarily results in an increase in overall energy demand.

The secondary industry has always been the largest energy consumption industry in China. The primary industry consumes less energy and the tertiary industry has relatively high energy utilization efficiency. The increment of energy consumption brought by the vigorous development of the tertiary industry is not obvious. As the energy consumption of the secondary industry is much higher than that of the primary and tertiary industries, with the gradual adjustment of China's industrial structure in the future, especially the adjustment of the proportion of the secondary industry, energy consumption will be affected.

It is shown that the output value of secondary industries has remained dominant during the last few decades (**Figure 5**). China is aggressively establishing a new green and low-carbon economy, promoting reduced energy consumption in industrial industries, owing to industrial restructuring and upgrading and the strict limits of the "carbon peak and carbon neutrality" targets. China's economy will enter a new era during the 14th Five-Year Plan period. Under the "Belt and Road Initiative", industrial transformation and upgrading, the proportion of primary industry will continue to shrink, as will the proportion of secondary industry, while the proportion of tertiary industry will continue to expand, and by 2050, will be the supporting industry of China's national economic growth. Accordingly, the structure of energy consumption will change as well.

China's total energy processing and conversion efficiency is expected to reach 73.3% by 2019.[5] We believe that the impact of technical innovation on energy consumption is mostly due to advancements in scientific and technological levels, notably the advancement of energy-saving technology, which increases the rate of energy conservation and decreases energy consumption. Furthermore, technical advancement allows for the new energy adaptation, resulting in a shift in energy consumption structure and, as a result, a shift in China's energy consumption. Technological progress to promote energy efficiency, energy saving, and emissions reduction significantly promote the role, is the key to achieve energy conservation and emissions reduction. As a result, we choose independent research and development as a metric for China's technical advancement.

Household consumption has changed dramatically as people's living standards have improved. People seek enhanced consuming enjoyment, thereby increasing the demand for direct energy consumption, especially for high-quality energy such as electricity, gas fuel and new energy. Urbanization has enhanced people's living standards. Compared with things in the 1990's, in 2019, the rural consumption increased by 504%, while urban consumption increased by 793%. The rise in living standards has increased household energy consumption, from 139 kg standard coal in 1990 to 442 kg standard coal in 2019.[5] The future will see increased per capita energy consumption due to improved quality of life, increased family automobile ownership, and increased usage of air conditioning and other household equipment.

[4]China Statistical Yearbook 2021, http://www.stats.gov.cn/tjsj/ndsj/

[5]China Statistical Yearbook 2021, http://www.stats.gov.cn/tjsj/ndsj/

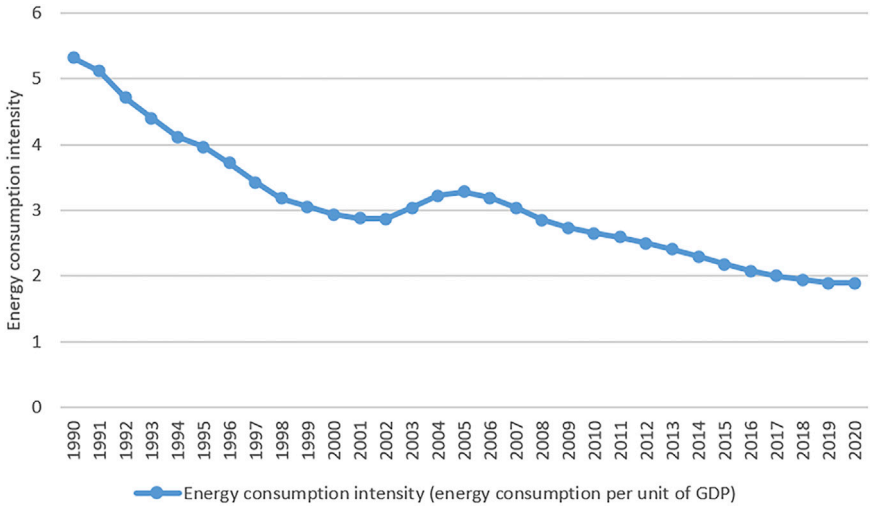


FIGURE 6 | change of China's energy consumption intensity from 1990 to 2020. Data source: China Statistical Yearbook 2021.

Energy consumption intensity, or energy intensity for short is first proposed by Patterson et al. (1996). This macro index mainly measures the energy consumption per unit output value and can better reflect the energy utilization efficiency level of a country. **Figure 6** illustrates China's energy intensity from 1990–2000 at 1990 prices. Throughout the last 2 decades, there is a rising energy consumption accompanied by a remarkable drop in energy intensity of use. China's energy intensity reduced from 5.3 tons of standard coal per 10,000 yuan to 1.88 tons, indicating that China's energy efficiency is improving.[5] By 2020, China's energy intensity had decreased to half of its levels in 1990. The main causes have been recognized as technical and structural changes. The 14th Five-Year Plan calls for promote high-quality development, improve the implementation of the dual management of energy consumption intensity and total amount and organize the implementation of key energy saving and emission reduction projects, and promote energy efficiency.

Price has a direct factor affecting consumption. We may deduce from supply and demand theory that high prices lead to lower demand and consumption, whereas low prices lead to increased consumption. Similarly, energy costs have an impact on energy use. Other variables, such as energy-saving and emission-reduction programs, automobile purchase limitation rules, and so on, will also affect energy consumption, but these aspects are difficult to measure or lack relevant statistical data, therefore they are not taken into account for the time being. After selecting the influencing factors, we preprocess the data for the SVR model prediction.

3.3 Data Preprocessing

All data in this paper are from China Statistical Yearbook, China Energy Statistical Yearbook and China Science and Technology Statistical Yearbook from 1990 to 2021. The influencing factors introduced above are taken as independent variables of energy consumption. Variable Settings and their meanings are shown in **Table 2** below. The indicators of GDP, household consumption level and energy consumption intensity have been converted according to the constant prices in 1990, and energy price p is represented by fuel and power purchasing price index.

Furthermore, the dimensions of the seven input and one output indices examined in this paper's analysis of energy consumption are inconsistent, and the data varies substantially in magnitude, which might result in significant variations in prediction results. As a result, every index data must be normalized in advance, that is, all index data must be converted to between [0,1] using **Eqn 15**.

$$x = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{15}$$

Next, we introduce the data set and how the prediction model is built with several variants and data exploratory analysis. We collect energy consumption data from statistical yearbooks from 1990 to 2020, randomly select 22 years of sample data as a training set and the remaining years data as a test set to evaluate the model's prediction accuracy, and then forecast China's energy consumption during the 14th Five-Year Plan period. This study examined the trend in China's energy use and makes recommendations for China's future energy development.

4 EMPIRICAL ANALYSIS

4.1 Data Training and Accuracy Analysis of Prediction Model

The input to the SVMs model is set to China's energy consumption influencing factors, while the output is set to energy consumption. Following normalization of the predicted and actual values of the linear fitting, the SVR model chose 22 years of historical data at random as the training sample for modeling and simulation, as shown in **Figure 7** (a). This indicates a high accuracy of the model (R^2 = 0.999). The model's validity is validated using the same test set data; **Figure 7** (b) shows a

TABLE 2 | Variables and their meanings.

	Name	Variable Meaning
The dependent variable	Energy consumption (*EC*)	The sum of all kinds of energy consumed in China for production and living
The independent variables	Gross Domestic Product (*GDP*)	GDP calculated at 1990 constant prices
	Residents' consumption level (*C*)	The average consumption of all Chinese residents in material products and services
	Energy price (*p*)	Expressed by fuel power category purchase price index
	Population size (*POP*)	Resident population of China
	Technical Progress (*RD*)	Internal expenditure
	Energy consumption intensity (*EI*)	China's energy consumption per unit of GDP.
	The proportion of secondary industry (*I*)	The proportion of added value of China's secondary industry in GDP.

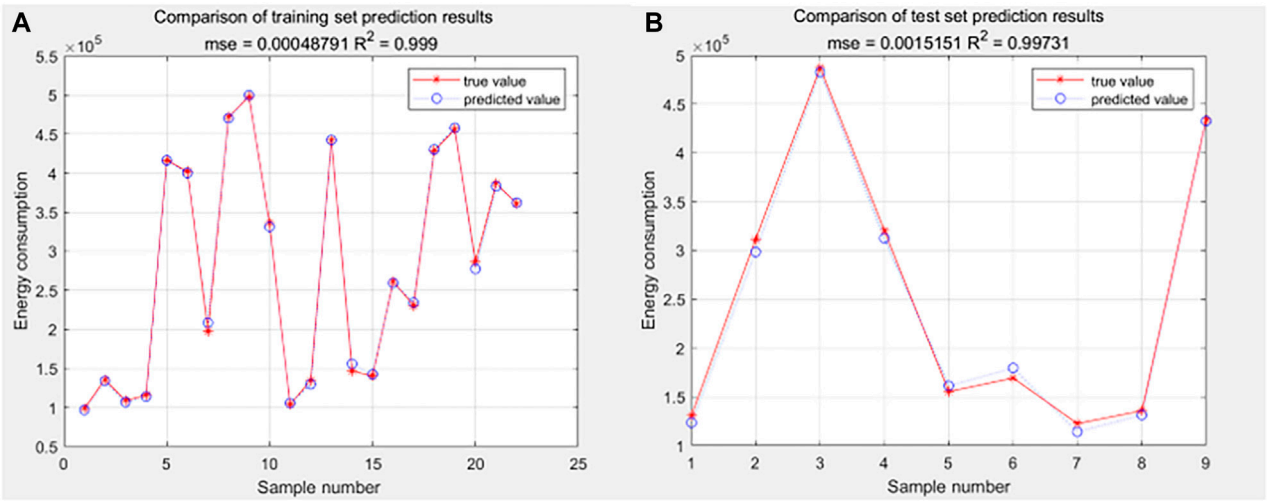


FIGURE 7 | Real and predicted values of China's energy consumption in the training (a) and testing set (b).

nine-year test set showing the connection between raw data and predicted data. As can be observed, the model projected values and the real value are reasonably comparable (R^2 = 0.997), implying that it can be used to forecast China's energy consumption.

Then, we compare actual and predicted values from 2013 through 2020 (**Table 3**). The prediction results are pretty close to the actual energy consumption figure. The model's validity is established, and it is demonstrated that the SVR model has an excellent forecasting effect on China's energy consumption and a high degree of forecasting ability or generalization, making it suitable for predicting China's future energy consumption changes.

Two competing models, namely GM (1, 1) and MA (3) are employed to test the accuracy of the SVR model on energy consumption forecasting. We chose these models for comparison because they are the most commonly used in energy prediction research, have a high level of representativeness, and can be applied to a wide range of situations. Consistent with the steps of the SVR model, the energy consumption data from 1990 to 2012 were used. Then, based on GM (1, 1) model and MA (3) model, the energy consumption of China from 2013 to 2020 is predicted.

TABLE 3 | Prediction results of China's energy consumption based on SVR model.

year	Actual value (ten thousand tons of standard coal)	Predicted value (ten thousand tons of standard coal)	Prediction error (%)
2013	416,913	416,317.12	−0.14
2014	428,334	430,307.75	0.46
2015	434,113	433,765.65	−0.08
2016	441,492	443,472.43	0.45
2017	455,827	457,781.11	0.43
2018	471,925	471,483.83	−0.09
2019	487,488	485,610.61	−0.39
2020	498,000	500,432.54	0.49

Taking into account the predictions from the SVR, GM (1, 1), and MA (3) models, **Table 4** and **Figure 8** illustrate the accuracy of the fitting value in contrast to the original value. The results reveal that the GM (1, 1) model exhibits a substantial prediction error, the majority of which exceeds 5%. Additionally, while the MA (3) model outperforms the GM (1, 1) model in terms of prediction performance, its prediction accuracy is not as steady as the SVR model, and prediction results vary greatly. In general, the

TABLE 4 | Simulation of energy consumption in China by SVR, GM (1, 1) and MA (3) model (Unit: ten thousand tons of standard coal).

year	Original value	SVR Predicted value	Prediction error (%)	GM (1,1) Model predicted value	Prediction error (%)	MA (3) Model predicted value	Prediction error (%)
2013	416,913	416,317.12	−0.14	387,374.40	−7.09	417,965.05	0.252
2014	428,334	430,307.75	0.46	409,238.80	−4.46	432,031.39	0.863
2015	434,113	433,765.65	−0.08	432,337.28	−0.41	432,965.35	−0.264
2016	441,492	443,472.43	0.45	456,739.49	3.45	442,998.62	0.341
2017	455,827	457,781.11	0.43	482,519.03	5.86	451,203.35	−1.014
2018	471,925	471,483.83	−0.09	509,753.63	8.02	471,942.73	0.004
2019	487,488	485,610.61	−0.39	538,525.42	10.47	487,506.31	0.004
2020	498,000	500,432.54	0.49	568,921.17	14.24	498,018.71	0.004

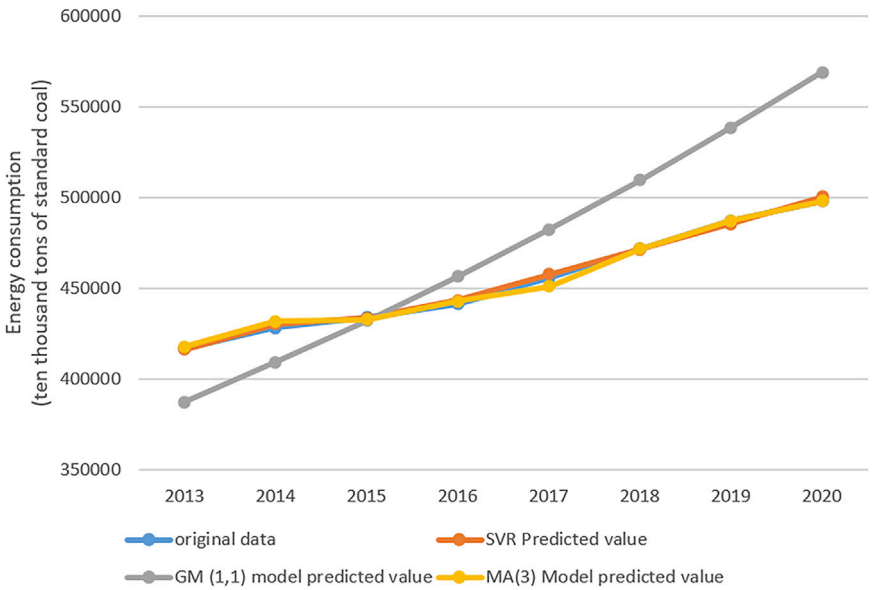


FIGURE 8 | Estimated values of energy consumption.

SVR model surpasses the other two commonly used forecasting models when it comes to projecting China's energy consumption. As a result, the SVR model can be used to anticipate China's future energy consumption throughout the period covered by the "14th Five-Year Plan".

4.2 Predicted Value of Influencing Factors

According to the Chinese government network, the National Bureau of Statistics, the energy bureau of the relevant data, 14th Five-Year energy development planning and the relevant data such as the government statistical bulletin, through the scenario analysis, this article selects the China 2021–2025, the total energy consumption of seven influencing factors of value analysis and estimate (**Table 5**). The following values of the influencing factors setting are mainly based on the "*Outline of the 14th Five-Year Plan (2021–2025) for National Economic and Social Development and Vision 2035*".[6]

(1) *GDP*: GDP increased by 6% year on year in 2019 in terms of comparable prices. The GDP for 2019 is 25,783.73 trillion yuan based on 1990 prices. According to the outline of the 14th Five-Year Plan, new economic development progress has been made, growth potential has been fully realized, and average annual GDP growth has been kept within a reasonable range. China's GDP is expected to grow at an annual rate of around 7% during the 14th Five-Year Plan period. The national economy is predicted to sustain a modest growth range of 6–8% over the 14th Five-Year Plan period, thanks to the expansion of the national economy and the expansion of the macroeconomic base. Then we set the annual GDP growth rate for the 14th Five-Year Plan at 6, 6.5, 7, 7.5, and 8%, respectively.

(2) Proportion of added value of secondary industry (*I*): In 2019, the tertiary industry contributed 59.4% of added value, and this figure is expected to exceed 60% by 2025. However, as a result of the epidemic's impact, the service industry has been severely restricted in recent years, with the secondary industry

[6]http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm

TABLE 5 | Indicators for the 14th Five-Year Plan period.

Variables	Estimated annual growth rate from 2021 to 2025
GDP	6%, 6.5, 7%, 7.5, and 8%
I	38, 37, 36, 36, and 36%
RD	An average annual rate of 7%
C	6, 6.5, 7, 7.5, and 8%
POP	1.418 billion, 1.423 billion, 1.429 billion, 1.435 billion, and 1.441 billion
EI	1.84, 1.79, 1.74, 1.69, and 1.65 tons of standard coal per 10,000 yuan
p	Price index is fixed at 106

accounting for 36.8% of added value in 2019. The proportion of secondary industry and tertiary industry are expected to wane and wax during the 14th Five-Year Plan period, and the tertiary industry will become the supporting industry of China's national economic growth by 2050, and the energy consumption structure will also change accordingly. The secondary industry's added value is expected to fluctuate between 35 and 38% during the "14th Five-Year Plan." We set the added value of secondary industry at 38, 37, 36, 36, and 36% in 2021–2025, respectively.

(3) Technological progress (*RD*): Technological progress plays a critical role in promoting energy efficiency, energy conservation, and emissions reduction. Independent research and development (R&D) is frequently viewed as the primary means of a country's technological progress, research and development expense. The more money we invest in research and development, the more we can foster independent innovation and technical progress, thereby increasing energy efficiency, lowering energy intensity, and lowering energy consumption. Therefore, we select independent research and development to measure China's technological progress. The 14th Five-Year Plan outlines a major increase in China's capability for innovation, with overall R&D investment increasing at an annual pace of more than 7% on average. As a result, we project that between 2021 and 2025, China's internal R&D investment would expand at an average annual rate of 7%.

(4) Household consumption level (*C*): The standard of life has risen, and people's well-being has reached a new level. Per capita disposable income has mostly maintained pace with GDP growth, as has consumption. As a result, we anticipate consumption to increase by 6, 6.5, 7, 7.5, and 8% throughout the 14th Five-Year Plan period.

(5) Population size (*POP*): In 2019, China's population surpassed 1.4 billion and is projected to reach 1.412 billion by 2020. Given the low level of desire for children, China proposed and pushed the "three-child" policy. However, a number of studies have demonstrated that implementing the "three-child" policy will have little effect on China's natural population growth rate. The natural population growth rate is not expected to surge, but due to the effect of inertia of the huge population base and growth, the size of the population will expand further. Therefore, if the average annual growth rate from 2010 to 2020 remains at 4% till 2025, China's population will be 1.418 billion, 1.423 billion, 1.429 billion, 1.435 billion, and 1.441 billion in 2021 and 2025, respectively.

(6) Energy consumption intensity (*EI*): The 14th Five-Year Plan prioritizes the development of a clean, low-carbon, safe, and efficient energy system, as well as the enhancement of energy security. Increasing the share of non-fossil energy in total energy consumption to around 20% and maintaining a healthy balance of overall energy consumption and energy intensity. The 14th Five-Year Plan also asks for sticking to a new vision of development, ensuring sustainable and sound economic growth while considerably enhancing development's quality and efficiency, and reducing energy consumption intensity by 13.5% over the last 5 years. Based on a total reduction of 13.5% in energy consumption per unit GDP over the next 5 years, or a 2.7% annual reduction rate, energy consumption per unit GDP in 2021–2025 will be 1.84, 1.79, 1.74, 1.69, and 1.65 tons of standard coal per 10,000 yuan.

(7) Energy prices (*P*): We anticipate reasonably constant energy prices, with an average annual growth rate of 6% from 2010 to 2020. China's 2021–2025 gasoline and electricity purchase price index is fixed at 106 in comparison to the corresponding price in 1990.

4.3 Energy Consumption and Energy Structure Forecast

We forecast China's energy usage between 2021 and 2025 using a trained SVR model (**Figure 9**). The accuracy of the prediction is greater than 98.4%. China's energy consumption is projected to continue growing at an average annual rate of 7%, surpassing 6 billion tons of standard coal by 2024. As China's industrialization and urbanization intensify, the country's energy consumption will continue to grow rapidly, putting policymakers under increased pressure to strike a balance between energy use and economic development.

In accordance with "*Outline of the 14th Five-Year Plan (2021–2025) for National Economic and Social Development and Vision 2035*" [7], China's economic and social development must adopt a green transition aimed at carbon reduction. The expanding energy consumption will exacerbate the imbalance and raise the need for energy imports. Energy technology is still in its infancy at the early stages of new energy development and change. All of these circumstances are likely to jeopardize China's energy security, which should be taken into account.

China's energy consumption structure remains irrational at the moment. The share of low-carbon energy sources such as oil, natural gas, and non-fossil energy consumption is significantly lower than the average for industrialized countries. As China's economic development enters a new normal, improving the energy structure will enable the country to progressively wean itself off coal, improve economic quality, and achieve sustainable economic growth. As a result, we use Markov Chain (MC) to predict the energy consumption structure.

[7]http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm

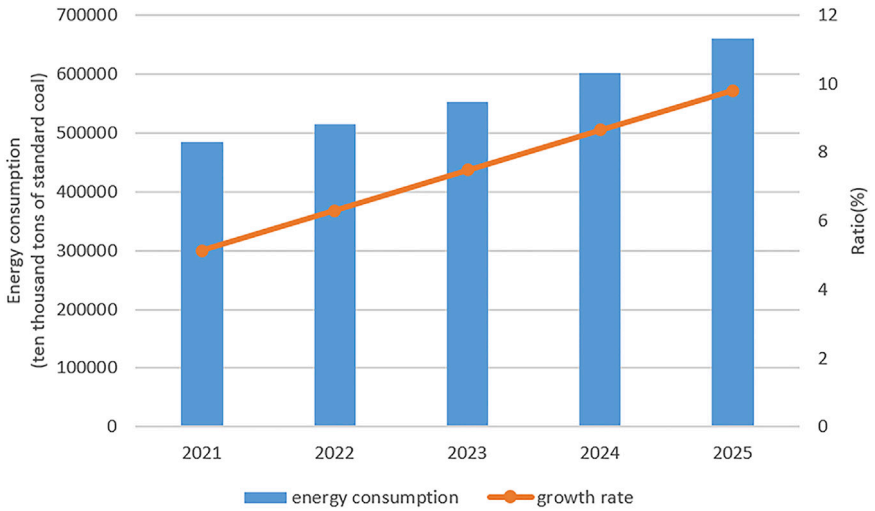


FIGURE 9 | Forecast value of China's energy consumption during the 14th Five-Year Plan period.

Only the proportion of coal declined between 2012 and 2019, while the proportions of crude oil, natural gas, and clean energy grew. It can be seen that the maintenance rate for crude oil, natural gas, and clean energy is 1, however the drop in coal's share is decomposed into the other three. In 2020, the proportion of coal and crude oil has reduced, the retention rate of natural gas and clean energy has increased to 1, and the decline in coal and crude oil proportion has been decomposed into an increase in the other two items. The one-step transition matrix for 2012–2020 computed using **Eqs 9–11** can be found in Appendix A1. The average transfer matrix for 2012–2020, according to the transfer matrix every 2 years, is as follows:

$$\bar{P} = \begin{bmatrix} 0.9769 & 0.0039 & 0.0071 & 0.0121 \\ 0 & 0.9993 & 0.0002 & 0.0004 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

According to the average transfer matrix, the amount of natural gas and clean energy in China's primary energy structure has continuously increased (the transfer proportion on the main diagonal is 1), which is compatible with the country's clean energy policy. The coal-to-oil ratio has decreased. The proportion of coal energy structure declined the most quickly, with the lowest retention rate (97.69%). Petroleum energy is retained at a rate of 99.93%.

Based on the energy consumption structure in 2020, we set the initial state, in which Per (0) = (0.568, 0.189, 0.084, 0.159). The energy consumption proportion in 2021 is anticipated to be the energy consumption proportion matrix in 2020 multiplied by the average state transfer matrix using the average transfer matrix formula. The energy consumption ratio in 2022 is calculated by multiplying the energy consumption ratio in 2020 by the square of the average state transition matrix, and so on. The prediction results of 2021–2025 energy consumption structure are shown in **Table 6**. The results show that coal consumption has been declining, while natural gas is growing slowly (Wang and Wang, 2019; Zhao and Liu, 2019).

TABLE 6 | Forecast of Energy consumption structure.

Year	Coal (%)	Oil (%)	Gas (%)	Electricity (%)
2021	55.5	19.1	8.8	16.6
2022	54.2	19.3	9.2	17.3
2023	53.0	19.5	9.6	17.9
2024	51.7	19.7	10.0	18.6
2025	50.5	19.9	10.3	19.2

Our estimation on the structure of energy consumption suggests that coal consumption is decreasing steadily, and by 2025, coal is expected to account for around 50.5% of overall energy consumption. According to the *Guiding Opinions on the High-Quality Development of the Coal Industry during the 14th Five-Year Plan*, national coal consumption will be managed at around 4.2 billion tons by the end of the "14th Five-Year Plan," and this binding target will be met.[8] Oil consumption is rising year by year and is predicted to account for 19.9% of total energy consumption by 2025, which is in line with the *Development Plan of China's Petroleum and Petrochemical Equipment Manufacturing Industry in the 14th Five-year* Plan's requirements for total oil consumption in my country.[9] However, it is worth noticing that, according to *The 14th Five-Year Comprehensive Work Plan for Energy Conservation and Emission reduction*, natural gas should account for around 15% of China's energy mix by 2025.[10] Compared with our forecast figure of 10.3% in 2025, there is still a gap between natural gas consumption and the target. According to *the Comprehensive Work Plan*[10], non-fossil energy, such as electricity consumption, must account for more than 20% of total energy consumption.

[8]http://www.coalchina.org.cn/index.php?m=content&c=index&a=show&catid=9&id=129818

[9]https://m.cpei.org.cn/index.html

[10]http://www.gov.cn/zhengce/content/2022-01/24/content_5670202.htm

The forecast figure for 2025 is 19.2%, which has yet to be achieved, and the target will be met in 2030.

5 CONCLUSIONS AND SUGGESTIONS

This study anticipates China's overall energy consumption throughout the 14th Five-Year Plan period by proposing a machine learning model and predicts the evolving energy consumption structure. In comparison to commonly used forecasting models, i.e. the GM (1, 1) and MA (3), the SVR model is more accurate in forecasting China's energy consumption. The overall energy consumption is projected to continue growing at an average pace of 7%, surpassing 6 billion tons of standard coal by 2024. In terms of energy structure, it is expected that China's energy consumption structure will be more rational in 2025, with increased non-fossil energy consumption and decreased coal consumption, while natural gas consumption continues to grow at a low rate. The growing disparity between energy consumption and production will undermine China's energy security. We are compelled to make proactive adjustments to our energy strategy and structure. The following policy implications are made based on the findings:

(1) When energy supply expansion is limited, ensuring national security requires scientific regulation of the energy consumption elasticity system. We need to reform our energy consumption structure, speed up industrial restructuring, and recommence our energy conservation efforts. It is very important to increase investment in clean energy, improve the efficiency of conventional energy sources (such as coal), and advocate for energy conservation and emission reduction.
(2) There is more intense competition between provinces and cities, urban and rural areas, and industry and civilians in terms of energy consumption. To avert the deterioration, regional total energy consumption management strategies should be attempted. A relationship between energy consumption and economic performance should be established so that when economic performance exceeds the task, overall regional energy consumption exceeds the regional energy development goal.
(3) In terms of energy consumption, some policies that have worked well, such as lax coal management, clean oil promotion, and electric energy replacement programs, must be maintained. While retaining the proportion of oil, the amount of coal should be gradually reduced and transferred to natural gas and other clean energy sources.
(4) The long-term sustainable development of natural gas should be prioritized. On the supply side, we should enhance the efficiency of existing gas resources and vigorously boost offshore gas and natural gas hydrate development. While total output continues to increase, it is critical to monitor market prices and analyze and establish an acceptable natural gas pricing system to ensure steady growth in natural gas consumption demand.

A few caveats are necessary. Investigating the disparities in energy consumption between urban and rural locations would be an intriguing side study. Additionally, policies aimed at energy conservation and emission reduction have a close connection with energy consumption. These will be studied in greater detail in our subsequent investigation.

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

ZM was instrumental in the conceptualization and design of the study. HS was responsible for material preparation, data collecting, and analysis. XW contributed to the literature. HS wrote the initial draft of the manuscript, and all writers provided feedback on prior versions. The final manuscript was read and approved by all writers.

FUNDING

This research is funded by the National Social Science Foundation of China (NO.19BJY151). The National Social Science Foundation of China provided the necessary financial support for the research.

ACKNOWLEDGMENTS

Funding from the National Social Science Foundation of China (NO.19BJY151) is gratefully acknowledged.

REFERENCES

Aboagye, S. (2017). The Policy Implications of the Relationship between Energy Consumption, Energy Intensity and Economic Growth in ghana. *OPEC Energ. Rev* 41 (4), 344–363. doi:10.1111/opec.12114

Achour, H., and Belloumi, M. (2016). Decomposing the Influencing Factors of Energy Consumption in Tunisian Transportation Sector Using the Lmdi Method. *Transport Policy* 52, 64–71. doi:10.1016/j.tranpol.2016.07.008

Ang, B. W. (2004). Decomposition Analysis for Policymaking in Energy: Which Is the Preferred Method? *Energy Policy* 32 (9), 1131–1139. doi:10.1016/S0301-4215(03)00076-4

Asafu-Adjaye, J. (2000). The Relationship between Energy Consumption, Energy Prices and Economic Growth: Time Series Evidence from Asian Developing Countries. *Energ. Econ.* 22 (6), 615–625. doi:10.1016/S0140-9883(00)00050-5

Birol, F., and Keppler, J. H. (2000). Prices, Technology Development and the Rebound Effect. *Energy Policy* 28 (6/7), 457–469. doi:10.1016/S0301-4215(00)00020-3

Chen, H., Tong, Y., Wu, L., and Engineering, M. (2021). Forecast of Energy Consumption Based on Fgm (1, 1) Model. *Math. Probl. Eng.* 2021, 1–11. doi:10.1155/2021/6617200

Chen, W., and Zhu, H. (2013). Energy Demand Forecast in China Based on Particle Swarm Optimization Algorithm. *China Popul. Resour. Environ.* 23 (03), 39–43.

Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. *Mach Learn.* 20 (3), 273–297. doi:10.1007/bf00994018

Ediger, V. Ş., and Akar, S. (2007). Arima Forecasting of Primary Energy Demand by Fuel in Turkey. *Energy Policy* 35 (3), 1701–1708. doi:10.1016/j.enpol.2006.05.009

Gori, F., and Takanen, C. (2004). Forecast of Energy Consumption of Industry and Household & Services in italy. *Int. J. Heat Technol.* 22 (2), 115–121.

Hao, Y., Niu, X., and Wang, J. (2021). Impacts of Haze Pollution on China's Tourism Industry: A System of Economic Loss Analysis. *J. Environ. Manage.* 295, 113051. doi:10.1016/j.jenvman.2021.113051

He, T., Li, Z., and Lin, H. (2011). On the Relationship between Energy Intensity and Industrial Structure in China. *Energ. Proced.* 5 (1), 2499–2503. doi:10.1016/j.egypro.2011.03.429

Huang, J., Du, D., and Hao, Y. (2017). The Driving Forces of the Change in China's Energy Intensity: An Empirical Research Using Dea-Malmquist and Spatial Panel Estimations. *Econ. Model.* 65, 41–50. doi:10.1016/j.econmod.2017.04.027

Hunt, L. C., and Ninomiya, Y. (2005). Primary Energy Demand in Japan: an Empirical Analysis of Long-Term Trends and Future Co2 Emissions. *Energy Policy* 33 (11), 1409–1424. doi:10.1016/j.enpol.2003.12.019

Inglesi, R. (2010). Aggregate Electricity Demand in South Africa: Conditional Forecasts to 2030. *Appl. Energ.* 87 (1), 197–204. doi:10.1016/j.apenergy.2009.08.017

Kran, M. S., Zceylan, E., Gündüz, M., and Paksoy, T. (2012). A Novel Hybrid Approach Based on Particle Swarm Optimization and Ant Colony Algorithm to Forecast Energy Demand of Turkey. *Energ. Convers. Manage.* 53 (1), 75–83. doi:10.1016/j.enconman.2011.08.004

Li, K., and Lin, B. (2014). The Nonlinear Impacts of Industrial Structure on China's Energy Intensity. *Energy* 69, 258–265. doi:10.1016/j.energy.2014.02.106

Lin, B., and Zhu, J. (2017). Energy and Carbon Intensity in China during the Urbanization and Industrialization Process: A Panel Var Approach. *J. Clean. Prod.* 168, 780–790. doi:10.1016/j.jclepro.2017.09.013

Marques, A. C., Fuinhas, J. A., and Tomás, C. (2019). Energy Efficiency and Sustainable Growth in Industrial Sectors in european union Countries: A Nonlinear Ardl Approach. *J. Clean. Prod.* 239, 118045.1–118045.12. doi:10.1016/j.jclepro.2019.118045

Meng, Z., Liu, X., Yin, K., Li, X., and Guo, X. (2020). Forecasting China's Energy Intensity by Using an Improved DVCGM (1, N) Model Considering the Hysteresis Effect. *Grey Syst. Theor. Appl.* 11 (3), 372–393. doi:10.1108/GS-02-2020-0022

Niu, D., Sun, W., and Zhao, L. (2004). Markov Energy Structure Prediction Model Based on Transfer Matrix Recognition. *J. North China Electric Power Univ. (Natural Sci. Edition)* (03), 59–61.

Patterson, M. G., Makarov, V. I., and Khmelinskii, I. V. (1996). What Is Energy Efficiency? - Concepts, Indicators and Methodological Issues. *Energy Policy* 24 (5), 377–390. doi:10.1016/0301-4215(96)00017-1

Ren, J., Xia, J., and Yin, Y. (2017). Prediction of Primary Energy Consumption Based on Improved BP Neural Network and Markov Model: A Case Study of Beijing. *Ecol. Economy* 33 (11), 6.

Rumelhart, D. E., Hinton, G. E., and Williams., R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature* 323, 533–536. doi:10.1038/323533a0

Soytas, U., and Sari, R. (2003). Energy Consumption and Gdp: Causality Relationship in G-7 Countries and Emerging Markets. *Energ. Econ.* 25, 33–37. doi:10.1016/S0140-9883(02)00009-9

Stern, D. I. (2000). A Multivariate Cointegration Analysis of the Role of Energy in the Us Macroeconomy. *Energ. Econ.* 22 (2), 267–283. doi:10.1016/s0140-9883(99)00028-6

Uenler, A. (2008). Improvement of Energy Demand Forecasts Using Swarm Intelligence: The Case of turkey with Projections to 2025. *Energy Policy* 36 (6), 1937–1944. doi:10.1016/j.enpol.2008.02.018

Wang, Y., and Wang, Y. (2019). Feasibility and Optimal Pathway of China's Double Targets for Carbon Reduction—The Perspective of Energy Structure Optimization. *China Environ. Sci.* 39 (10), 4444–4455. doi:10.19674/j.cnki.issn1000-6923.2019.0518

Xie, N.-M., Yuan, C.-Q., and Yang, Y.-J. (2015). Forecasting China's Energy Demand and Self-Sufficiency Rate by Grey Forecasting Model and Markov Model. *Int. J. Electr. Power Energ. Syst.* 66, 1–8. doi:10.1016/j.ijepes.2014.10.028

Yakut, E., and Özkan, E. (2020). Modeling of Energy Consumption Forecast with Economic Indicators Using Particle Swarm Optimization and Genetic Algorithm: An Application in Turkey between 1979 and 2050. *Alphanumeric J.* 8, 59–78. doi:10.17093/alphanumeric.747427

Yang, W., Sun, S., Hao, Y., and Wang, S. (2022). A Novel Machine Learning-Based Electricity price Forecasting Model Based on Optimal Model Selection Strategy. *Energy* 238, 121989. doi:10.1016/j.energy.2021.121989

Zhao, Z., and Liu, Q. (2019). China's Energy Strategy Planning Based on Prediction of Energy Consumption, Production and Structure. *Resour. Industries* 21 (06), 1–8. doi:10.13776/j.cnki.resourcesindustries.20191206.007

Zhu, X., Dang, Y., and Ding, S. (2020). Using a Self-Adaptive Grey Fractional Weighted Model to Forecast Jiangsu's Electricity Consumption in china. *Energy* 190, 116417. doi:10.1016/j.energy.2019.116417

Zong, W. G., and Roper, W. E. (2009). Energy Demand Estimation of south korea Using Artificial Neural Network. *Energy Policy* 37 (10), 4049–4054. doi:10.1016/j.enpol.2009.04.049

Zou, B., and Luo, B. (2019). Rural Household Energy Consumption Characteristics and Determinants in China. *Energy* 182 (1), 814–823. doi:10.1016/j.energy.2019.06.048

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.

Copyright © 2022 Meng, Sun and Wang. 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.

APPENDIX

APPENDIX A1 THE ONE-STEP TRANSITION MATRIX FOR 2012–2020.

$$P_{2012-2013} = \begin{bmatrix} 0.9839 & 0.0015 & 0.0073 & 0.0073 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P_{2013-2014} = \begin{bmatrix} 0.9763 & 0.0030 & 0.0044 & 0.0163 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P_{2014-2015} = \begin{bmatrix} 0.9696 & 0.0167 & 0.0030 & 0.0106 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P_{2015-2016} = \begin{bmatrix} 0.9749 & 0.0047 & 0.0047 & 0.0157 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P_{2016-2017} = \begin{bmatrix} 0.9743 & 0.0032 & 0.0129 & 0.0096 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P_{2017-2018} = \begin{bmatrix} 0.9736 & 0 & 0.0116 & 0.0149 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P_{2018-2019} = \begin{bmatrix} 0.9780 & 0.0017 & 0.0068 & 0.0135 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P_{2019-2020} = \begin{bmatrix} 0.9844 & 0 & 0.0057 & 0.0085 \\ 0 & 0.9947 & 0.0019 & 0.0029 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

frontiers | Frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fevo.2022.855606

Prediction of Air Pollution Interval Based on Data Preprocessing and Multi-Objective Dragonfly Optimization Algorithm

*Jiyang Wang, Jingrui Li and Zhiwu Li**

Institute of Systems Engineering, Macao University of Science and Technology, Macao, China

With the rapid development of global industrialization and urbanization, as well as the continuous expansion of the population, large amounts of industrial exhaust gases and automobile exhaust are released. To better sound an early warning of air pollution, researchers have proposed many pollution prediction methods. However, the traditional point prediction methods cannot effectively analyze the volatility and uncertainty of pollution. To fill this gap, we propose a combined prediction system based on fuzzy granulation, multi-objective dragonfly optimization algorithm and probability interval, which can effectively analyze the volatility and uncertainty of pollution. Experimental results show that the combined prediction system can not only effectively predict the changing trend of pollution data and analyze local characteristics but also provide strong technical support for the early warning of air pollution.

Keywords: atmospheric contamination prediction, temporal convolution network, fuzzy information granulation, multi-objective dragonfly optimization algorithm, interval prediction

OPEN ACCESS

Edited by:
Wendong Yang,
Shandong University of Finance and Economics, China

Reviewed by:
He Jiang,
Jiangxi University of Finance and Economics, China
Feng Liu,
University of Technology Sydney, Australia

***Correspondence:**
Zhiwu Li
zwli@must.edu.mo

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Ecology and Evolution

Received: *15 January 2022*
Accepted: *07 March 2022*
Published: *25 April 2022*

Citation:
Wang J, Li J and Li Z (2022) Prediction of Air Pollution Interval Based on Data Preprocessing and Multi-Objective Dragonfly Optimization Algorithm. Front. Ecol. Evol. 10:855606. doi: 10.3389/fevo.2022.855606

INTRODUCTION

With the continuous development of the economy and the rising living standards, the deterioration of the environment, land desertification, greenhouse effect, and other problems have begun to plague us. In addition, the United States and other developed countries have classified indoor air pollution into the five environmental factors that endanger human health. The Health Effects Institute from the US released "State of Global Air (2020)," which indicates that, at least 6.7 million people worldwide, will die from chronic exposure to air pollution in 2019 (State of Global Air, 2020).

Up to now, many studies have been conducted to study the problem of air pollution. Recently, to accurately measure the quality of air, particulate matter (PM) has become a significant and common index to be monitored (Beaulant et al., 2008). $PM_{2.5}$ is one type of PM, which means that the particulate matter in the ambient air has an aerodynamic equivalent diameter less than or equal to 2.5 μm (van Donkelaar et al., 2006). It can be suspended in the air for a long period, and the higher the concentration of its content in the air, the more serious the air contamination. Compared with coarser atmospheric particles, $PM_{2.5}$ has the following features: small particle size, large area, strong activity, easy adhesion, and long residence time in the atmosphere; thus, it has a greater impact on human health and the quality of the atmospheric environment (Sun and Li, 2020).

As a result, $PM_{2.5}$ has become a worldwide problem to be solved, and many institutions have established various methods to accurately monitor $PM_{2.5}$ concentrations (McKeen et al., 2007; Borrego et al., 2011; Air Quality Expert Group, 2012; Bergen et al., 2013; Wakamatsu et al., 2013).

Bai et al. (2019) proposed the DL-SSAE method as an autoencoder model to consider the advantages of seasonal analysis and deep feature learning to predict the hourly $PM_{2.5}$ concentrations. Irina et al. developed the Community Multiscale Air Quality method with five different data preprocessing strategies to analyze the concentrations of $PM_{2.5}$, and they found that the Kalman filter correction could compute the most precise results (Djalalova et al., 2015). In addition, Samia et al. (2012) combined autoregressive integrated moving average (ARIMA) and Ann to enhance predicting performance, and the results show that the proposed hybrid system could be used to efficiently forecast and provide useful air quality information. In addition, multiple linear regression have been utilized to forecast $PM_{2.5}$ or PM_{10} concentrations in the air to make decisions related to traffic restrictions in the future or support the control of air quality (Akyüz and Çabuk, 2009; Genc et al., 2010). Moreover, Banik et al. (2020) employed long short-term memory (LSTM) to analyze wind speed in various seasons, and they concluded that LSTM performs better than Elman and non-linear auto-regressive models. Osowski and Garanty (2007) used support vector machine (SVM) to decompose the original data and to predict the air quality of Poland based on wavelet representation. Another common method is the gray model (GM), which was employed by Pai et al. After comparing with other models for predicting the performance of $PM_{2.5}$ and PM_{10} concentrations of Taipei, they demonstrated that GM (1, 1) could be a useful early warning system for nearby citizens (Pai et al., 2013).

Additionally, the temporal convolutional network (TCN) is widely used to achieve more accurate performance. For instance, Zhu et al. (2020) solved the problem of long-term dependencies and performance degradation of a deep convolutional model by TCN, which shows that the power system with TCN performs better and more stably compared with others. Li et al. (2018) predicted oil consumption with various parameters according to TCN and found that the proposed model could obtain more satisfying results and help make decisions for the energy market. Wei designs a convolutional spiking neural network to deal with temporal datasets, which corrects and optimizes the historical performance, and more accurately forecasts wind speed. Also, this method could quantify the differences in predicting the performances that resulted from uncertainties (Wei et al., 2021). Chen et al. (2020) established a structure with the convolutional neural network (CNN) to forecast associated sequences and to handle more complex seasonal problems, which helps make useful decisions to assess power generation by providing more evidence. Yang W. et al. (2020) combined empirical mode decomposition (EMD) and **TCN** to forecast the remaining useful life and reduce the cost during the operation. Tian and Wang (2021) applied the temporal convolution networks with the quantile regression (TCNQR) method to judge the period of health and operation. In this study, we used TCN as one of the forecasting tools to obtain the results of air quality.

The recently developed approaches mainly belong to point forecasting, which includes some disadvantages and limitations (Wang J. et al., 2021). For example, Wang et al. (2022b) have pointed out that the point predicting approaches produce an unavoidable error during the operation, which might result in immense risks for an electric power system since it only depends on the accurate results. In addition, a considerable amount of time and high cost will be wasted if precise information cannot be provided, which is also a loss to the entire power system.

Unlike point prediction, which gives a "specific numerical prediction," the interval forecast aims for a future period and gives an interval in which the predicted values are likely to occur, with a prediction interval corresponding to the expected probability. The interval forecast gives more prediction information than the point forecast, which means that we can get the value of the point forecast within a certain interval based on a certain probability, thus more scientifically characterizing the uncertainty of the model forecast.

As for data preprocessing, information granulation (IG) is a technique for studying the formation and representation of information grains and for information pre-processing. Fuzzy information granulation (FIG) is one type of IG first proposed by Zadeh (1997) to discuss how to deal with fuzzy datasets. FIG has been employed to acquire original data of fluctuating traffic and construct a traffic flow, predicting the approach with interval forecasting (Guo et al., 2018). Zhang and Na (2018) applied FIG to transform the historical agricultural price into FIG particles, and the forecasting results show that the proposed price predicting system model performs more efficiently with better accuracy. FIG could also be used in the power system. For example, the authors utilize FIG to remove the variability of the historical series of wind and solar energy, and the experimental results demonstrate that the developed approach performs efficiently and could help decision-makers stabilize the energy system (He et al., 2019b). Additionally, to forecast the actual streamflow data, FIG is combined with support vector regression (SVR) to provide more precise computation and eliminate the fluctuation of the streamflow, which means that the proposed model has a more accurate prediction interval of the hydrologic system (He et al., 2019a).

According to the existing research about $PM_{2.5}$ concentrations and forecasting, we found that the majority of the models are combined models. Compared with the traditional single model, the combined models avoid the error of individual approaches and yield more accurate results. Therefore, more researchers have adopted combined models for prediction. For example, Wang S. et al. (2021) applied a novel wind power combined predicting system to obtain more precise performance, which supports further research in wind generation. Wang et al. (2022a) used four foundation models and optimized the weight coefficient using a multi-objective water cycle algorithm (MOWCA) to predict hourly $PM_{2.5}$ concentrations. Details of a single model and combined models are summarized in **Table 1**. In **Table 2** for detailed nomenclature in the article.

Based on the analyses above, this study employs a novel combined predicting system to monitor $PM_{2.5}$ and PM_{10}. It integrates FIG, TCN, ARIMA, and LSTM to forecast $PM_{2.5}$ and PM_{10} concentrations, then uses a weight generation structure to compute each coefficient, and finally combines the single approaches to achieve a better result of the experiments.

The primary contributions and innovations of this study are shown as follows.

TABLE 1 | Summary of predicting approach types.

Category	References	Advantages	Disadvantages
Physical model	Xiao et al., 2019	Physical model concentrates on long-term predicting data series where it could perform better	Physical models have difficulties in predicting data series in short term, and they need to collect sufficient data as an initial dataset
Statistical approach			
Auto-regressive moving average (ARma)	Wang et al., 2012; Wang and Hu, 2015	Statistical approaches achieve satisfying performance in dealing with linear data	The assumption of statistical approach is tough to realize, and they cannot operate well in non-linear patterns
Auto-regressive integrated moving average (ARIMA)			
Grey method (GM)	Ding, 2019		
Artificial intelligence			
Artificial neural network (Ann)	Wang et al., 2016		
Support vector machine (SVM)	Zhang et al., 2019		
Fuzzy logic (FL)	Sfetsos, 2000		
Back propagation neural network (BPnn)	Bin et al., 2014	They are suitable to handle non-linear data series and could obtain better results and stability.	Artificial intelligence approaches cannot perform well due to the over-fitting problem, and single approaches have some limitations.
General regression neural network (GRnn)	Majumder and Maity, 2018		
Long short-term memory (LSTM)	Banik et al., 2020		
Particle swarm optimization (PSO)	Liu W. et al., 2019		
Combined model			
WPD-PSO-BPnn	Liu H. et al., 2019		
FT-CS	Yang H. et al., 2020	Combined models consider the strengths and weaknesses of single approaches to achieve more precise results	The running period of the combined models is longer than other single models, and the weight coefficient needs to be considered
VMD-BEGA-LSTM	Mencar and Fanelli, 2008		
FIG-SVR	He et al., 2019a		

(1) *To avoid the limitations of point predicting methods, this study proposes a useful interval predicting approach.* This technique could deal with the fluctuation associated with the $PM_{2.5}$ and PM_{10} concentrations by quantifying the information of the original dataset. The performance of interval predicting is shown to be more effective than that of other point forecasting approaches.
(2) *According to the decomposition and reconstruction techniques, this study applies the data pretreatment method to eliminate the negative influence of the initial data series.* As a superior data preprocessing strategy, FIG is used to decrease high-frequency noise and to reconstruct the novel data sequences to acquire the significant elements of the historical data and facilitate the smooth implementation of the next phase.
(3) *A combined model is developed in the predicting section to obtain the results of $PM_{2.5}$ concentrations.* It obtains more accurate prediction results when compared with the traditional $PM_{2.5}$ and PM_{10} concentrations of prediction approaches.
(4) *The developed model could be employed in air quality monitoring.* The proposed system and the predicting results are clearly improved by providing more useful information on air quality to people and analyzing and predicting $PM_{2.5}$ and PM_{10} concentrations even in more complicated conditions.

The rest of this article is organized as follows. Section "Forecasting System Development" touches upon the design of the forecasting system, including data fuzzy information granulation and the proposed combined forecasting system. Section "Framework of the Proposed Forecasting System and Parameter" describes the framework and parameters of the proposed prediction system. To further verify the accuracy and effectiveness of the proposed combined model from various aspects, detailed experimental results and analysis are presented in section "Experimental Results and Discussion". Finally, section "Conclusion" concludes this research.

FORECASTING SYSTEM DEVELOPMENT

This section develops an innovative combined predicting system to predict the $PM_{2.5}$ and PM_{10} concentrations in the air, which enhances the performance of the results by a data denoising strategy and a predicting approach.

TABLE 2 | Nomenclature.

Nomenclature			
Abbreviate			
Anns	Artificial neural networks	**WPD**	Wavelet packet decomposition
BPnn	Back propagation neural Network	**FT**	Fuzzy theory
ARIMA	Auto-regressive integrated Moving Average	**CS**	Cuckoo search
ARma	Auto-regressive moving Average	**BEGA**	Binary encoding genetic optimization Algorithm
GM	Gray method	**VMD**	Variational mode decomposition
GRnn	General regression neural Network	**MSE**	Mean squared error
LSTM	Long short-term memory Network	**MAPE**	Average absolute percent error
PSO	Particle swarm optimization	**MAE**	Mean absolute error of *N* predicting results
SVM	Support vector machine	**RMSE**	Root of mean error squares
TCN	Temporal convolution network	**SSE**	Sum of squared error
PICP	Prediction interval coverage probability	**PBW**	Prediction band width
FL	Fuzzy logic	**MOWCA**	Multi-objective Water Cycle Algorithm
SVR	Support vector regression	**FIG**	Fuzzy information granulation
IG	Information granulation	**TCNQR**	Temporal convolution network with the quantile regression
EMD	Empirical mode decomposition	**MLR**	Multiple linear regression
PM	Particulate matter	**MODOA**	Multi-objective dragonfly optimization algorithm
CNN	Convolutional neural network		

Decomposition and Denoising Strategies

Fuzzy information granules (FIG) construct information granules by building fuzzy sets for each subsequence formed by discretizing the time series (Mencar and Fanelli, 2008). The core of fuzzy information granulation is to complete the fuzzification process after the window is created, which mainly includes window division and information fuzzification.

The window division is to convert the time series $\overline{R} = \left\{\overline{R}_1, \overline{R}_2, \ldots, \overline{R}_\gamma\right\}$ into the granular time series $\bar{\bar{\Theta}} = \left\{\bar{\bar{\Theta}}_1, \bar{\bar{\Theta}}_2, \ldots, \bar{\bar{\Theta}}_\varsigma\right\}$ after information granulation by setting the time granularity $\widehat{\mathbf{E}}$ to divide $\overline{R} = \left\{\overline{R}_1, \overline{R}_2, \ldots, \overline{R}_\gamma\right\}$ into $\underline{\mathbf{H}}$ subseries $\bar{\bar{\Theta}} = \left\{\bar{\bar{\Theta}}_1, \bar{\bar{\Theta}}_2, \ldots, \bar{\bar{\Theta}}_\varsigma\right\}$, where $\underline{\mathbf{H}} = \frac{\gamma}{\widehat{\mathbf{E}}}$ and the ηth subseries is $\bar{\bar{\Theta}}_\eta = \left[\overline{\overline{R}}_1^{(\eta)}, \overline{\overline{R}}_2^{(\eta)}, \ldots, \overline{\overline{R}}_{\widehat{\mathbf{E}}}^{(\eta)}\right]$.

$$\left\{\overline{R}_1, \overline{R}_2, \ldots, \overline{R}_\gamma\right\} \Rightarrow \left\{\left[\overline{\overline{R}}_1^{(1)}, \overline{\overline{R}}_2^{(1)}, \ldots, \overline{\overline{R}}_{\widehat{\mathbf{E}}}^{(1)}\right], \ldots, \left[\overline{\overline{R}}_1^{(\underline{\mathbf{H}})}, \overline{\overline{R}}_2^{(\underline{\mathbf{H}})}, \ldots, \overline{\overline{R}}_{\widehat{\mathbf{E}}}^{(\underline{\mathbf{H}})}\right]\right\} \quad (2-1-1)$$

The information granulation of the time series $\overline{R} = \left\{\overline{R}_1, \overline{R}_2, \ldots, \overline{R}_\gamma\right\}$ is to construct the information particles $\tilde{\Gamma} = \left\{\tilde{\Gamma}_1', \tilde{\Gamma}_2', \ldots, \tilde{\Gamma}_\varsigma'\right\}$ using the fuzzy method for each of the $\underline{\mathbf{H}}$ subsequences $\bar{\bar{\Theta}} = \left\{\bar{\bar{\Theta}}_1, \bar{\bar{\Theta}}_2, \ldots, \bar{\bar{\Theta}}_\varsigma\right\}$ formed by the discretization operation.

Suppose that $\mathbf{Z}$ is a given theoretical domain, then a fuzzy subset $\Lambda = \{\chi, \Omega(\chi) \,|\, \chi \in \mathbf{Z}\}$ on $\mathbf{Z}$, $\Omega(\chi) : \chi \rightarrow [0, 1]$ represents the affiliation function of Λ. Two fuzzy subsets, Φ and Ξ, are equal, denoted by $\Phi = \Xi$, if they have the same affiliation function, i.e., $\widehat{\Omega}'_{\Phi}(\chi) = \widehat{\Omega}''_{\Xi}(\chi)$.

In this study, the triangular fuzzy particles are chosen to construct the information grain and its affiliation function is as follows:

$$\mathbf{A}_{Tf}(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}-\mathbf{I}_{Tf}}{\mathbf{K}_{Tf}-\mathbf{I}_{Tf}}, \mathbf{I}_{Tf} \leq \mathbf{x} \leq \mathbf{K}_{Tf} \\ 0, \mathbf{x} < \mathbf{I}_{Tf} \cup \mathbf{x} > \mathbf{N}_{Tf} \\ \frac{\mathbf{N}_{Tf}-\mathbf{x}}{\mathbf{N}_{Tf}-\mathbf{K}_{Tf}}, \mathbf{K}_{Tf} < \mathbf{x} \leq \mathbf{N}_{Tf} \end{cases} \quad (2-1-2),$$

where x is the variable in the theoretical domain and $\mathbf{I}_{Tf}$, K_{Tf}, and $\mathbf{N}_{Tf}$ are the three parameters of the triangular type fuzzy example affiliation function, which correspond to the lower boundary, average level, and upper boundary of the window after fuzzy particleization, respectively.

Fuzzy sets get rid of the either-or duality in classical set theory and extend the value domain of the affiliation function from the binary {0, 1} to the multi-valued interval [0, 1], which is a kind of extension of the set theory. Information fuzzification is the fuzzification of each information grain, and the fuzzification of a single sub-window, $\bar{\bar{\Theta}}_\mu$, generates multiple fuzzy sets $\tilde{\Gamma}'_\mu = \left[\tilde{\Gamma}''_{\mu;1}, \tilde{\Gamma}''_{\mu;2}, \tilde{\Gamma}''_{\mu;3}\right]$.

Considering the single-window problem, $\bar{\bar{\Theta}}_\mu = \left[\bar{\bar{\mathbf{T}}}_1^{(\mu)}, \bar{\bar{\mathbf{T}}}_2^{(\mu)}, \ldots, \bar{\bar{\mathbf{T}}}_{\widehat{\mathbf{E}}}^{(\mu)}\right]$ should first be viewed as a window for fuzzification. The task of fuzzification is to build a triangular fuzzy particle TFP on $\bar{\bar{\Theta}}_\mu = \left[\bar{\bar{\mathbf{T}}}_1^{(\mu)}, \bar{\bar{\mathbf{T}}}_2^{(\mu)}, \ldots, \bar{\bar{\mathbf{T}}}_{\widehat{\mathbf{E}}}^{(\mu)}\right]$, which can reasonably explain the fuzzy concept M of $\bar{\bar{\Theta}}_\mu$. The fuzzy particle $\tilde{\Gamma}'_\mu = \left[\tilde{\Gamma}''_{\mu;1} = \widehat{\mathbf{I}}_{Tf}^{\mu}, \tilde{\Gamma}''_{\mu;2} = \widehat{\mathrm{K}}_{Tf}^{\mu}, \tilde{\Gamma}''_{\mu;3} = \widehat{\mathbf{N}}_{Tf}^{\mu}\right]$ can be constructed by the relevant parameters in the determined affiliation function (2-1-2) of the triangular fuzzy particle.

Predicting Algorithm

In this section, the basic theory and equations of some forecasting approaches are described.

Auto-Regressive Integrated Moving Average Model

The AR($\boldsymbol{p}$) model means the auto-regressive approach of the $\boldsymbol{p}^{th}$ order, expressed as (Hamilton James, 2015):

$$y_t = \boldsymbol{c} + \boldsymbol{a}_1 y_{t-1} + \cdots + \boldsymbol{a_p} y_{t-p} + u_t \quad (2-2-1),$$

where $\boldsymbol{a_1}, \cdots, \boldsymbol{a_p}$ are indicators; $\boldsymbol{c}$ is a constant; and u_t is referred to as the random variable.

Besides, the MA($\boldsymbol{q}$) model represents the moving average model of the $\boldsymbol{p}^{th}$ order, which is defined as:

$$y_t = \boldsymbol{a} + u_t + \boldsymbol{m}_1 u_{t-1} + \cdots + \boldsymbol{m_q} u_{t-q} \quad (2-2-2),$$

where $\boldsymbol{m}_1, \cdots, \boldsymbol{m}_q$are the factors of the approach; y_t is always set as 0, and the expectation of y_t can be written as $\boldsymbol{a}$. Also, u_t, u_{t-1} and u_{t-q} could describe the white noise error terms of the initial series.

Then, the **ARMA**($\boldsymbol{p, q}$) combines the two approaches listed above, which is shown in the following formula:

$$y_t = c + \boldsymbol{a_1} y_{t-1} + \cdots + \boldsymbol{a_p} y_{t-p} + u_t + \boldsymbol{m_1} u_{t-1} + \cdots + \boldsymbol{m_q} u_{t-q} \quad (2-2-3)$$

If these three approaches are employed in dealing with samples with non-stationarity evidence, we could consider taking various steps to decrease this limitation, which is regarded as **ARIMA** ($\boldsymbol{p,d,q}$), where d is the degree of differencing.

Deep Learning Using Long Short-Term Memory Recurrent Neural Networks

Recurrent neural network (RNN) is one type of ANN, and the combinations of various samples become a directed cycle. LSTM is proposed to handle long series. Both LSTM and RNNs could employ some gates to fix the gradient problem. Some scientists have proved that RNNs are included in the hidden layer, which is one of the features of LSTM (Gers et al., 2000). The three layers of RNNs with LSTM demonstrated in **Figure 1** present the memory cell functions.

$$f_t = \boldsymbol{\sigma}(W_f \bullet [C_{t-1}, h_{t-1}, x_t] + b_f) \quad (2-2-4)$$

$$i_t = \boldsymbol{\sigma}(W_i \bullet [C_{t-1}, h_{t-1}, x_t] + b_i) \quad (2-2-5)$$

$$o_t = \boldsymbol{\sigma}(W_o \bullet [C_t, h_{t-1}, x_t] + b_o) \quad (2-2-6)$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t \quad (2-2-7)$$

$$\widetilde{C}_t = \boldsymbol{tanh}(W_c \bullet [h_{t-1}, x_t] + b_c) \quad (2-2-8)$$

$$h_t = o_t \times \boldsymbol{tanh}(C_t) \quad (2-2-9),$$

where x_t is the input value; h_t is the output vector; C_t represents the cell state variable; $\boldsymbol{W}$ and $\boldsymbol{b}$ are indicator matrices and indicator; f_t, i_t, and o_t are forget, input, and output gate variables, respectively. In the equation, σ means the sigmoid formula and ***tanh*** refers to a rescale logistic sigmoid function belonging to (−1, 1).

Temporal Convolutional Network

A TCN, an interval predicting method, is a special kind of CNN (Shelhamer et al., 2017). It includes three sections: causal convolution, dilated convolution, and residual network. The first section makes sure of the result at time β, and we assume the input value $\overleftrightarrow{\boldsymbol{\mu}} \in \underrightarrow{\widetilde{\boldsymbol{\Theta}}}^{\mathrm{B}}$and a filter $\pi : \{0, \ldots, k-1\} \to \underrightarrow{\widetilde{\boldsymbol{\Theta}}}$. The historical convolutional layer is stated by two equations: $\underleftrightarrow{\widehat{\boldsymbol{\Lambda}}}(\overleftrightarrow{\boldsymbol{\mu}}_{\beta}) = (\overleftrightarrow{\boldsymbol{\mu}} \odot \pi)(\beta) = \sum_{j=0}^{k-1} \pi_j \overleftrightarrow{\boldsymbol{\mu}}_{\beta-j}$and $\widetilde{ouse} = (\underleftrightarrow{\widehat{\boldsymbol{\Lambda}}}(\overleftrightarrow{\boldsymbol{\mu}}_1), \underleftrightarrow{\widehat{\boldsymbol{\Lambda}}}(\overleftrightarrow{\boldsymbol{\mu}}_2), \cdots, \underleftrightarrow{\widehat{\boldsymbol{\Lambda}}}(\overleftrightarrow{\boldsymbol{\mu}}_{\mathrm{B}}))$, where k is the size of convolutional kernel, $\widetilde{ouse}$ is the output series, and $\underleftrightarrow{\widehat{\boldsymbol{\Lambda}}}(\bullet)$ is the process of convolution.

The second part uses a hyperparameter to jump some input values; thus, a range longer than it used to be could be accepted by the filter. In detail, if the causal convolution mixes, the m^{th} layer dilated convolution can be described by: $\underleftrightarrow{\widehat{\boldsymbol{\Lambda}}}(\overleftrightarrow{\boldsymbol{\mu}}_{\beta}) = (\overleftrightarrow{\boldsymbol{\mu}} \odot d_m \pi)(\beta) = \sum_{j=0}^{k-1} \pi_j \overleftrightarrow{\boldsymbol{\mu}}_{\beta - d_m j}$ and $\widetilde{ouse} = (\underleftrightarrow{\widehat{\boldsymbol{\Lambda}}}(\overleftrightarrow{\boldsymbol{\mu}}_1), \underleftrightarrow{\widehat{\boldsymbol{\Lambda}}}(\overleftrightarrow{\boldsymbol{\mu}}_2), \ldots, \underleftrightarrow{\widehat{\boldsymbol{\Lambda}}}(\overleftrightarrow{\boldsymbol{\mu}}_{\mathrm{B}}))$, where d_m means the dilation indicator of the m^{th} layer and the range could be set to $2^{m-1}2^{m-1}$. Here, $\beta - d_m j$ represents the historical direction. The second formula is a temporal convolutional layer, which constructs TCN in many layers.

If the layers are deep, to deal with the issue of decreased efficiency of the CNN results, a residual block is utilized. During the training procedure, we added a residual connection into the block to ensure normal operation in the deep layers. Moreover, TCN prevents the over-fitting problem by introducing the dropout layer after each dilated convolution (Srivastava et al., 2014).

Combined Model

Combining forecasts has long been recognized as an effective and a simple way to improve forecast stability, an improvement over a single model. This study proposes a new combined forecasting model that fuses ARIMA, neural networks, and the non-positive constraint theory.

The traditional forecasting combination method attempts to find the best weight of the combined models based on minimizing SSE:

$$\min \mathbf{F}=\mathbf{D}^T\mathbf{E}\mathbf{D}=\sum_{t=1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{m} d_i d_j e_{it} e_{jt} \begin{cases} \boldsymbol{R^T D} = 1 \\ \boldsymbol{D} \geq 0 \end{cases} (2-3-1),$$

where $\mathbf{D} = \left(d_1, d_2, \ldots, d_m\right)^T$ is the weight vector; $\mathbf{R} = (1, 1, \ldots, 1)^T$ is a column vector where all elements are 1; and $\mathbf{E} = \left(\mathbf{E} =_{ij}\right)_{m \times m}$is called the error information matrix ($\mathbf{E}_{ij} = e_i^T e_j$, $e_i = (e_{i1}, e_{i2}, \ldots, e_{iN})$).

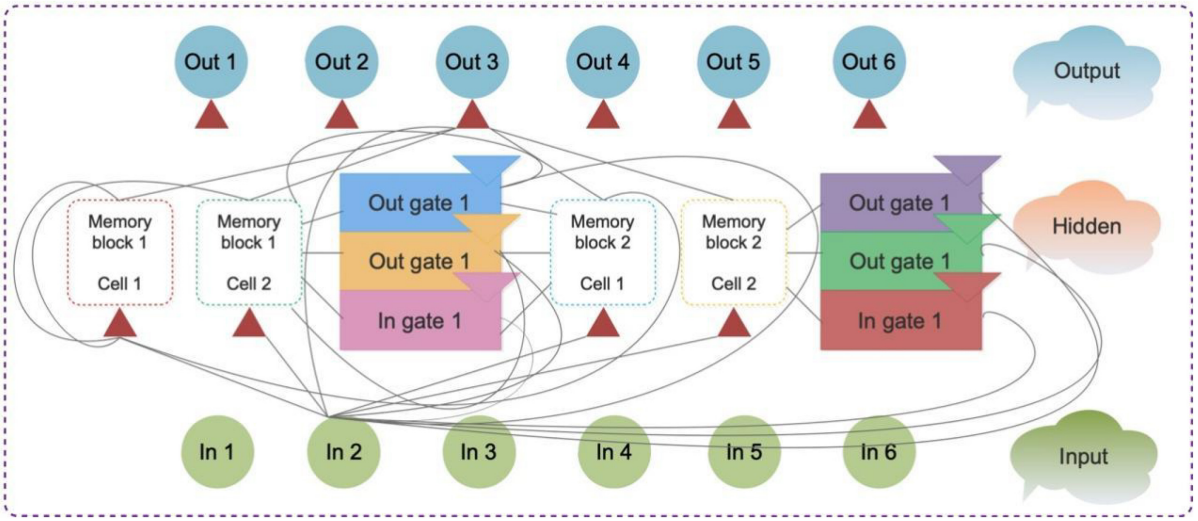


FIGURE 1 | Three-layer long short-term memory (LSTM) of two LSTM memory blocks (Gers et al., 2000).

An improvement of the traditional combination method based on the non-negative constraint theory (TCM-NNCT) and non-positive constraint theory is given as follows:

$$\min \mathbf{J}=\mathbf{D}^T\mathbf{E}\mathbf{D}=\sum_{t=1}^{T}\sum_{j=1}^{m}\sum_{i=1}^{m} d_i d_j e_{it} e_{jt} \qquad (2-3-2)$$

$$\text{st } \mathbf{R}^T\mathbf{D}=1 \qquad (2-3-3)$$

In Eq. (2-3-1), the weight vector has no limitation in the range [0,1]. The experiment results show that the combination model can obtain desirable results if the weight vector has a value in the range of [−2,2]. This section provides a weight-determined method that will be assessed by experimental simulation rather than a theoretical proof.

Interval Prediction Based on Temporal Convolutional Network

In this section, we constructed a novel multidimensional time series CNN prediction model for air contamination forecasting and uncertainty analysis.

Interval Prediction Module

There is no need to presuppose an error distribution in an interval forecasting model that, based on a linear model, is expressed as:

$$\begin{aligned}\mathbf{Y} &= \tilde{\beta}_0(\theta)+\tilde{\beta}_1(\theta)\mathbf{X_1}+\cdots+\tilde{\beta}_n(\theta)\mathbf{X_n}+u \\ &= \tilde{Q}(\theta;\mathbf{X})+\mathbf{u} \qquad (2-4-1),\end{aligned}$$

where $\boldsymbol{X}$ is n explanatory variables, $\tilde{\beta}_i$ is a vector that can be determined based on $\boldsymbol{X}$ and θ, and $\tilde{\beta}_i$ is a vector that can be identified based on $\boldsymbol{X}$ and θ. Given K samples, the vector is statistically estimated: $\tilde{\beta}_i(\theta)=\arg\min_{\beta}\sum_{i=1}^{K} f(\theta)\left[\mathbf{Y}_i-\tilde{Q}(\theta;\mathbf{X})\right](i=1,2,\ldots,K)$, where $f(\theta)$ is a piecewise linear loss function that can be defined as:

$$f(\theta)=\begin{cases}\theta u, & u\geq 0\\ (\theta-1)u, & u<0\end{cases} \qquad (2-4-2)$$

Evaluation Index of Prediction Model and Interval Forecasting

Due to the advanced non-linear characteristics and uncertainty of air contamination data, the prediction error of a single model is usually possible. For this case, the calculation of the prediction interval, i.e., the higher and lower bounds for predicting future values, is appropriate for the prediction of air contamination data. Below the given prediction interval, the predictions are often created higher and therefore the stability of prediction can be additionally improved. Different indicators can judge the prediction results and quality of the interval prediction model, such as the following common evaluation indicators.

Prediction interval coverage probability (PICP) is the most important index to measure the quality of the prediction interval, which reflects the probability result of the observed value falling into the prediction interval, namely reliability. In other words, the greater the probability value, the more the observations covered by the prediction interval and vice versa. In general, within the established prediction interval, the calculated probability p (PICP) should be higher than the rated confidence level, namely: $p=P\left(f_i\in\left[L\left(\overline{\overline{\mathbf{X}_i}}\right),U\left(\overline{\overline{\mathbf{X}_i}}\right)\right]\right)\geq\boldsymbol{\mu}$, $i=1,2,\ldots,K$, where $P(\bullet)$ is the expressed probability; $L\left(\overline{\overline{\mathbf{X}_i}}\right)$, $U\left(\overline{\overline{\mathbf{X}_i}}\right)$ represent the lower and upper bounds of the prediction interval predicted by $\overline{\overline{\mathbf{X}_i}}$, respectively; f_i is the predicted value; and $\boldsymbol{\mu}$ is a given confidence level. According to Bernoulli' law of huge numbers, $\tilde{\chi}_{cp}$ will be expressed by the frequency

that the prediction interval covers the determined value, and its likelihood converges to P, namely:

$$\tilde{\chi}_{cp} = \frac{1}{D}\sum_{i=1}^{D} c_i \times 100\% \xrightarrow{P} P\left(f_i \in \left[L\left(\overline{\overline{X_i}}\right), U\left(\overline{\overline{X_i}}\right)\right]\right) \quad (2-4-3)$$

where D is the predicted sample size and C_i is a Boolean quantity.

$$C_i = \begin{cases} 1, & \textbf{if } f_i \in \left[L\left(\overline{\overline{X_i}}\right), U\left(\overline{\overline{X_i}}\right)\right] \\ 0, & \textbf{if } f_i \notin \left[L\left(\overline{\overline{X_i}}\right), U\left(\overline{\overline{X_i}}\right)\right] \end{cases} \quad (2-4-4)$$

If $\tilde{\chi}_{cp} \geq \mu$, it indicates that the established prediction interval is valid; otherwise, it indicates that the established prediction interval is invalid, and it should be reestablished.

To decide the prediction interval more reasonably, it is necessary to depend on the prediction interval mean width percentage (**PIMWP**), which is the parameter basis for evaluating the prediction interval. If the forecast interval is wide enough, the coverage of the forecast interval will be on the brink of 100%. However, such a good interval cannot effectively provide the uncertainty data of the predicted value, rendering the results of the forecast interval meaningless.

If $\tilde{\chi}_{cp}$ is larger and $\bar{\tilde{\chi}}_{MWP}$ is smaller, the prediction interval of the model is more accurate and the performance is better.

$$\bar{\tilde{\chi}}_{MWP} = \frac{1}{D}\sum_{i=1}^{D} \frac{U\left(\overline{\overline{X_i}}\right) - L\left(\overline{\overline{X_i}}\right)}{f_i} \times 100\% \quad (2-4-5)$$

In addition, single high reliability and high clarity cannot reflect the performance of the interval prediction model, which is one of the biggest differences with the evaluation index of the deterministic prediction model. $\tilde{\chi}_{cp}$The performance evaluation indexes, $\tilde{\chi}_{cp}$ and $\bar{\tilde{\chi}}_{MWP}$, are often used to predict interval models. However, if some special situations occur, these two indicators cannot achieve a reasonable and scientific performance evaluation of the interval prediction model. For example, if the observed value is not within the prediction interval and if there is a small difference between $\tilde{\chi}_{cp}$ and $\bar{\tilde{\chi}}_{MWP}$ at the same time, it is impossible to measure the degree of deviation of the observed value from the prediction interval. The extent to which observations deviate from the predicted interval is immeasurable. To compensate for the shortcomings of $\tilde{\chi}_{cp}$ and $\bar{\tilde{\chi}}_{MWP}$, this study introduces another evaluation index of the prediction interval model, namely accumulated width deviation (AWD), which can clearly measure the deviation degree of observed values outside the prediction interval. Here, $\overline{\overline{\chi}}_{AWD} = \sum_{i=1}^{D} \zeta_i$, where ζ_i represents the degree to which the observed value deviates from the upper and lower bounds of the predicted interval.

$$\zeta_i = \begin{cases} \frac{L\left(\overline{\overline{X_i}}\right) - f_i}{f_i}, & \textbf{if } f_i < L\left(\overline{\overline{X_i}}\right) \\ 0, & \textbf{if } f_i \in \left[L\left(\overline{\overline{X_i}}\right), U\left(\overline{\overline{X_i}}\right)\right] \\ \frac{f_i - U\left(\overline{\overline{X_i}}\right)}{f_i}, & \textbf{if } f_i > U\left(\overline{\overline{X_i}}\right) \end{cases} \quad (2-4-6)$$

Under the condition of the same $\tilde{\chi}_{cp}$ and $\bar{\tilde{\chi}}_{MWP}$, the smaller the value of $\overline{\overline{\chi}}_{AWD}$, the higher the quality of the prediction interval.

The above three evaluation indicators, $\tilde{\chi}_{cp}$, $\bar{\tilde{\chi}}_{MWP}$, and $\overline{\overline{\chi}}_{AWD}$, are independent of each other, and only a certain feature of the prediction interval is considered. However, if only one evaluation index is selected, it is not enough to explain the quality and performance of the prediction interval. A high-quality prediction interval should conform to the confidence level requirements, i.e., $\tilde{\chi}_{cp}$ should be as high as possible while $\bar{\tilde{\chi}}_{MWP}$ and $\overline{\overline{\chi}}_{AWD}$ should be as low as possible. However, the definitions of $\tilde{\chi}_{cp}$, $\bar{\tilde{\chi}}_{MWP}$, and χ_{AWD} show that these three metrics are conflicting with each other: the higher the $\tilde{\chi}_{cp}$, the higher the $\bar{\tilde{\chi}}_{MWP}$; the lower $\bar{\tilde{\chi}}_{MWP}$, the lower $\tilde{\chi}_{cp}$ and the higher $\overline{\overline{\chi}}_{AWD}$; the lower the $\overline{\overline{\chi}}_{AWD}$ is, the higher the $\bar{\tilde{\chi}}_{MWP}$ is. Therefore, taking these three indicators into consideration, this study proposes a comprehensive index that can quantitatively evaluate the prediction interval, namely, prediction interval satisfaction index (PISI). It can be calculated by:

$$\chi_{PISI} = \left[1 - \left(1 + \lambda \times \overline{\overline{\chi}}_{AWD}\right) \bar{\tilde{\chi}}_{MWP} \times \left(1 + e^{-\eta\left(\chi_{cp} - \mu\right)}\right)\right] \times 100\% \quad (2-4-7),$$

where λ is the penalty factor of $\overline{\overline{\chi}}_{AWD}$, η is the penalty factor of $\tilde{\chi}_{cp}$, and μ (95%) is the given confidence level. In this study, we choose $\lambda = 0.5$ and $\eta = 50$.

If $\tilde{\chi}_{cp}$ is greater than the given confidence level μ, the curve of χ_{PISI} is flat and the value of χ_{PISI} tends to 1. At this point, χ_{PISI} is mainly determined by $\bar{\tilde{\chi}}_{MWP}$ and $\overline{\overline{\chi}}_{AWD}$. If $\tilde{\chi}_{cp}$ is less than the given confidence level μ, the value of χ_{PISI} changes according to the difference between $\tilde{\chi}_{cp}$ and μ, and χ_{PISI} is mainly determined by $\tilde{\chi}_{cp}$ at this time. Therefore, χ_{PISI} can further reflect the quality of the prediction interval by combining $\tilde{\chi}_{cp}$, $\bar{\tilde{\chi}}_{MWP}$, and $\overline{\overline{\chi}}_{AWD}$, making the evaluation of the prediction interval more effective and accurate.

FRAMEWORK OF THE PROPOSED FORECASTING SYSTEM AND PARAMETER

This section presents the description of the material analyzed (section "Dataset Description") and the entire probabilistic forecasting system applied in this study (section "Flow of the Proposed Ensemble Probabilistic Forecasting System").

Dataset Description

This study took the $PM_{2.5}$ and PM_{10} pollution data of Beijing, Shanghai, and Shenzhen as the experimental data set, which are daily data from January 2020 to December 2021. From each dataset, we extracted 4,386 point values as experimental sequences and selected 80% of the total length as training sets. The remaining 20% points were divided into test sets as shown in **Figure 2**.



FIGURE 2 | Information of the research areas.

Flow of the Proposed Ensemble Probabilistic Forecasting System

In accordance with the aforementioned data processing approaches and forecasting models, the proposed forecasting system includes Fuzzy information granulation, ARIMA, LSTM, TCN, multi-objective optimization, and interval prediction.

Step 1: The original three data sets were divided into a training set and a test set. A total of 4,386 pieces of data were collected. There were 3,500 pieces of data in training sets and 877 pieces of data in test sets.

Step 2: The pollution values of $PM_{2.5}$ and PM_{10} are reconstructed by graining Fuzzy information granulation and the data after noise reduction has been obtained.

Step 3: ARIMA, **LSTM,** and **TCN** were used for forecasting, and they were used as the comparative models of the multi-target dragonfly combination prediction results in the fourth step.

Step 4: The prediction results of **ARIMA**, **LSTM,** and **TCN** were combined with a multi-objective Dragonfly algorithm for optimization.

Step 5: Probabilistic forecasting module: The upper and lower bounds and the prediction interval were obtained by using interval prediction to forecast the progress of $PM_{2.5}$ and PM_{10} data.

By constructing the prediction interval, the probability prediction of air pollution is carried out. To determine the distribution of forecast errors resulting from point forecasts, three metrics were used: the **PICP**, the **BW**, and the **PINAW**. Furthermore, interval forecasts were created by combining upper and lower bounds with an optimal distribution with a design confidence level of 95%.

Model Selection and Parameter Setting

In general, a hybrid forecasting system adopts a decomposition strategy using a shallow neural network; all of the ARIMA and LSTM have satisfactory performance in solving regression problems. In DL, the TCN based on multidimensional time series is sensitive to the prediction of statistical data. Therefore, we selected the multi-objective dragonfly optimization algorithm based on the multidimensional time series for interval prediction. The model naming and argument details of the other models are presented in **Table 3**.

Evaluation Index

In this study, five evaluation indexes [such as the mean absolute percentage error (MAPE) and root mean square error

TABLE 3 | Compare the parameter settings of each model.

Model	Symbol	Meaning	Value	Reason
LSTM	n_i	Number of input layer nodes	5	Number of feature inputs
	n_h	Number of hidden layer nodes	[100, 100]	Trial-and-error manner
	n_0	Number of output layer nodes	1	Number of feature inputs
	e_t	Epochs of training	5000	Trial-and-error manner
ARIMA	p	Auto-Regressive term	[p-site1, p-site2, p-site3] = [5,5,3]	**AIC** and **BIC**
	d	Integrated term	1	**ADF** test
	q	Moving Average term	[q-site1, q-site2, q-site3] = [3,6,3]	**AIC** and **BIC**
TNC	$\overleftrightarrow{\mu}$	input value	DataSet	Preset
	$\widetilde{\widetilde{ouse}}$	output series	\	Preset
	$\widehat{\overleftrightarrow{\Lambda}}(\bullet)$	The process of convolution	\	Preset
	d_m	dilation indicator	\	Preset

(RMSE)] were used to assess the prediction system stability and accuracy, and other indicators were used to evaluate the interval prediction capability. **Table 4** presents the specific equations and definitions.

EXPERIMENTAL RESULTS AND DISCUSSION

This section discusses in detail the fuzzy granulation strategy based on multi-dimensional time series, the multi-objective dragonfly optimization algorithm, and the simulation results of interval prediction. To further improve the prediction results, the prediction efficiency (FE) and improvement rate (IR) of the proposed combined prediction model and interval prediction, as well as sensitivity, are analyzed in the study.

Data Pre-processing: Fuzzy Information Granulation

Through fuzzy information granulation, the pollution data of $PM_{2}.5$ and PM_{10} are processed.

Specific steps are as follows:

(1). To confirm sample extraction and fuzzification processing, sample information needs to be extracted to a certain extent. Then, the specific size of the window can be understood through the extracted data. Later, fuzzy information granulation processing is carried out according to the formula (2-1-1).

(2). The minimum, average, and maximum values are normalized after granulation treatment. The processing formula is:

$$\boldsymbol{p}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}},$$

where $\boldsymbol{p}_i$ is a variable data in the sample data; x_i is the normalized data coefficient; $x_{\min}$ is the minimum value of the extracted data; and $x_{\max}$is the maximum value in the sample.

In the subsequent combined prediction model, we use the granulated average $\boldsymbol{R}$ as the input for training and testing. The comparison result of fuzzy granulation with the original data is shown in **Figure 3**.

TABLE 4 | Evaluation metrics applied in this study.

Metric	Equation
RMSE	$\boldsymbol{RMSE} = \sqrt{\left[\sum_{i=1}^{N} \boldsymbol{PP_i} - \boldsymbol{AP_i}\right] / \boldsymbol{N}}$
MAPE	$\boldsymbol{MAPE} = \frac{1}{N}\sum_{i=1}^{N} \left\|(\boldsymbol{AP_i} - \boldsymbol{PP_i}) / \boldsymbol{AP}\right\| \times 100\%$
MSE	$\boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{y_i} - \widehat{\boldsymbol{y}}_i)^2$
Adjusted *R* square	$\boldsymbol{R} = 1 - \frac{\sum_i (\widehat{y}_i - y_i)^2}{\sum_i (\overline{y}_i - y_i)^2}$
SSE	$\boldsymbol{SSE} = \sum_{i=1}^{n} (\boldsymbol{y_i} - \widehat{\boldsymbol{y}}_i)^2$
PICP	$\tilde{\chi}_{cp} = \frac{1}{D}\sum_{i=1}^{D} \boldsymbol{c_i} \times 100\%$
BW	$\overline{\overline{X}}_{BW} = \sum_{i=1}^{D} \xi_i$
PINAW	$\overline{\overline{X}}_{PINAW} = \left[1 - (1 + \lambda \times \bar{\bar{X}}_{AWD}) X_{MWP} \times (1 + e^{-\eta(X_{cp} - \mu)})\right] \times 100\%$

Multi-Objective Optimization Combination Forecasting and Comparison Model

This section compares the proposed combined forecasting model with the commonly used single-point forecasting models. The single models of point prediction include ARIMA, LSTM, and T-convolutional neural network.

Experiment I: $PM_{2}.5$ **and** PM_{10} **forecasting.**

In this experiment, three traditional single models, **ARIMA**, **LSTM,** and **TCN**, are utilized to compare with the proposed system. The prediction results are shown in **Tables 5**, **6**.

(a) From **Tables 5**, **6**, we can see that the proposed model has achieved significant improvements compared with the three single models. In the forecast of $PM_{2}.5$ daily concentration in the three cities, the MAPE ($\times$100%) of the model in this study are **17.53130124**, **11.52643852,** and **6.00510985,** respectively, and the MAPE of PM_{10} are **20.10103656**, **19.61939713,** and **9.348984687,** respectively. In addition, there are substantial improvements in other data comparisons, which demonstrate the superior predictive power of the proposed model in simulating air contamination series.

(b) MAPE and RMSE are mainly used to measure the prediction error of each model. The smaller the value, the better the model prediction performance. In addition, the $\boldsymbol{R}$ index mainly evaluates the fit consistency between the

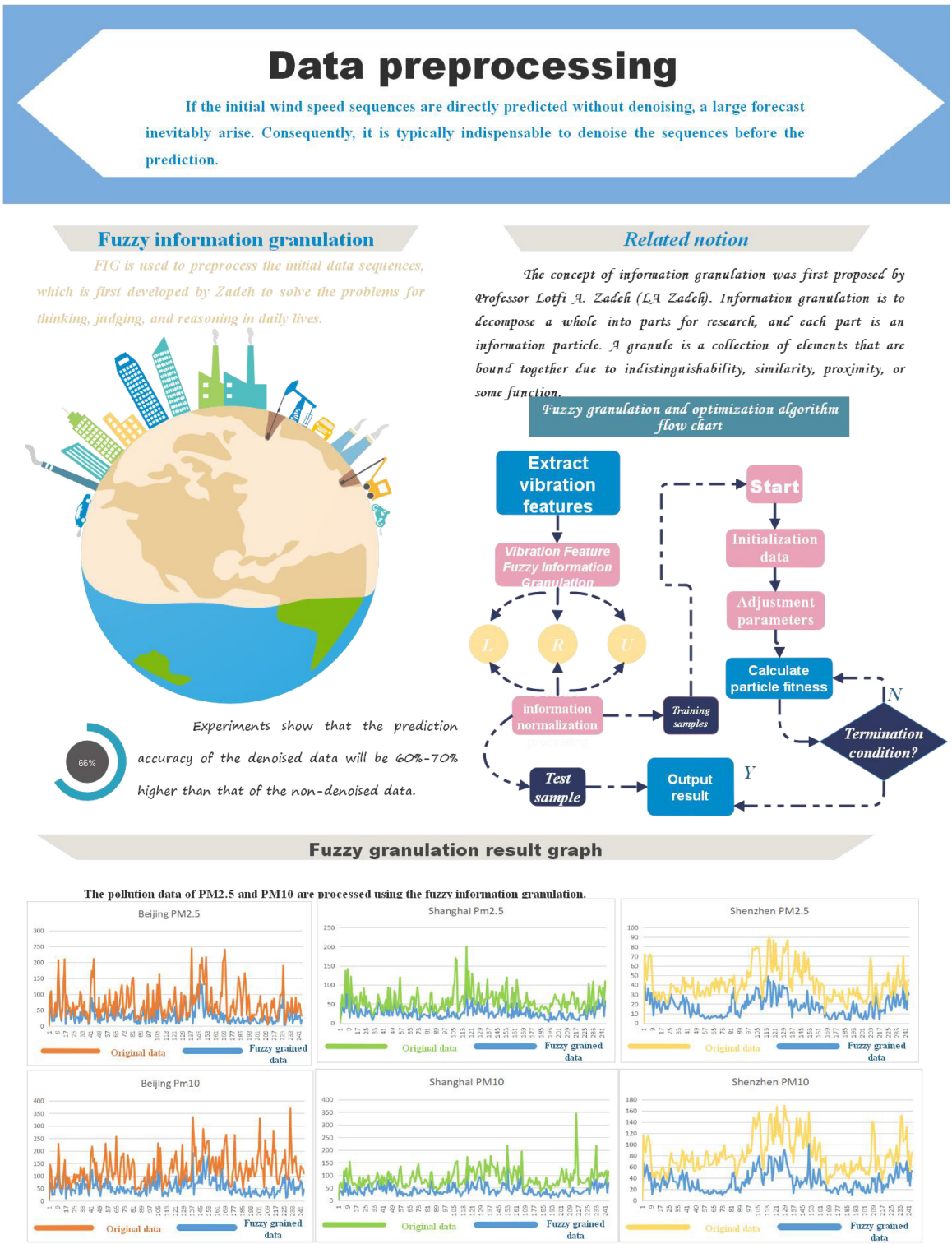


FIGURE 3 | Data preprocessing flowchart.

original value and the predicted value. The index values of the system are all larger than the reference model, indicating that the system has a better simulation effect on the air pollution sequence. The ***R*** value is negative, indicating that the model is not suitable for simulating the air pollution series.

TABLE 5 | Statistical errors of the proposed system and three traditional models for daily $PM_{2.5}$ concentrations of Beijing, Shanghai, and Shenzhen.

Dataset	Model	MAPE (×100%)	MSE	Adjusted *R* square	RMSE	SSE
Beijing	ARIMA	86.97079038	290.8284471	0.447545833148643	17.053693062077780	5.816568941111197e + 04
	LSTM	136.01785	413.20679	0.215077461604307	20.327488086028843	8.264135437752891e + 04
	TCN	142.5970955592992	483.03770834970896	0.1626560139686989	21.9781	66176.16604391013
	Proposed model	**17.53130124**	**148.870600556075**	**0.658175732872686**	**25.2317663**	**29774.1201112151**
Shanghai	ARIMA	43.06089753	139.425587	0.448273075306167	11.807861237106376	2.788511739895187e + 04
	LSTM	44.672726	173.59648	0.313053969451486	13.175601796860805	3.471929654188834e + 04
	TCN	60.826491189904864	265.8895658979947	−0.8598887306068743	16.3061	36426.870528025276
	Proposed model	**11.52643852**	**67.0458435333302**	**0.667495893927825**	**15.29028516**	**13409.1687066660**
Shenzhen	**ARIMA**	40.30779213	37.85861391	0.624570379270827	6.152935389900167	7.571722782457183e + 03
	LSTM	66.276283	89.870964	0.108783389745464	9.480030256857855	1.797419473418808e + 04
	TCN	60.826401222797664	84.42822174344232	0.0699984144048289	9.1885	11566.666378851598
	Proposed model	**6.00510985**	**17.5222223250933**	**0.985588304**	**7.806397025**	**3504.44446501865**

The indicators can be defined as $SSE = \sum_{i=1}^{N} (PP_i - AP_i)^2$, $RMSE = \sqrt{\left[\sum_{i=1}^{N} PP_i - AP_i\right] / N}$, *and* $MAPE = \frac{1}{N} \sum_{i=1}^{N} |(AP_i - PP_i) / AP| \times 100\%$. *The most satisfactory results are shown in bold.*

TABLE 6 | Statistical errors of the proposed system and three traditional models for daily PM_{10} concentrations of Beijing, Shanghai, and Shenzhen.

Dataset	Model	MAPE (×100%)	MSE	Adjusted *R* square	RMSE	SSE
Beijing	ARIMA	63.52736357	487.0786852	0.174520697390467	22.069859201835650	9.741573703776996e + 04
	LSTM	196.7689707455670	2.432077115750104e + 03	−3.121776177956965	49.316093881714760	4.864154231500208e + 05
	TCN	87.22617049732669	706.49757233221	−0.30037075679791525	26.5800 RMSE	96790.16740951277
	Proposed model	**20.10103656**	**172.374212971936**	**0.403054937**	**25.04239173**	**34474.8425943872**
Shanghai	ARIMA	29.09318151	176.4140437	0.528676587826571	13.282094853290085	3.528280873835899e + 04
	LSTM	35.3052930993369	2.069021408276871e + 02	0.447221882339130	14.384093326577350	4.138042816553741e + 04
	TCN	40.82371479881204	552.8902300983436	−0.09493369083894865	23.5136	75745.96152347307
	Proposed model	**19.61939713**	**98.0218221960804**	**0.342456465189497**	**19.8011941252118**	**19604.3644392161**
Shenzhen	ARIMA	25.09504498	88.93366541	0.659983745777606	9.430464750332934	1.778673308145440e + 04
	LSTM	28.8580663433921	1.252390110215527e + 02	0.521179080890740	11.191023680680540	2.504780220431053e + 04
	TCN	33.353551605811546	158.35503196357078	0.29102131681963195	12.5839	21694.639379009197
	Proposed model	**9.348984687**	**35.9146836500957**	**0.71286274**	**11.84927859**	**7182.93673001914**

The indicators can be defined as $SSE = \sum_{i=1}^{N} (PP_i - AP_i)^2$, $RMSE = \sqrt{\left[\sum_{i=1}^{N} PP_i - AP_i\right] / N}$, *and* $MAPE = \frac{1}{N} \sum_{i=1}^{N} |(AP_i - PP_i) / AP| \times 100\%$. *The most satisfactory results are shown in bold.*

Experiment II: SO2 and CO forecasting.

In Experiment **I**, we performed prediction experiments using PM2.5 and PM10 data, which achieved good results. To further verify the effectiveness of the prediction system, in Experiment 2, the SO_2 and CO data of three cities were used to conduct the experiment again.

Therefore, in this part, we selected the CO value and sulfur dioxide data of three cities for comparative experiments. The detailed results are shown in **Table 7**.

To further explore the application of the point prediction system, this experiment used SO_2 and CO daily datasets in Beijing, Shanghai, and Shenzhen to examine the superiority and applicability of the developed system. The results showed that the model proposed in this study not only exhibits the best prediction performance, indicating that the prediction system in this study is not only suitable for the prediction of $PM_{2.5}$ and PM_{10}, but also for the prediction of other air pollutants. Although the randomness and complexity of different datasets are different, the results show that the proposed model has strong applicability and effectiveness for the prediction of various air pollution and has potential application prospects in air pollution monitoring.

Indexes of Prediction Model and Interval Forecasting

Point forecasting only provides each forecast point for the target and does not show the probability of correct forecasting. However, in several problems, it is necessary to quantify the accuracy of estimates using countermeasures. Once the extent of uncertainty increases, the dependability of the point prediction decreases significantly. In contrast to point forecasting, prediction intervals not only provide the location in which observations are presumably made but also conjointly provide an indicator of capability known as the confidence level. Since interval forecasting is more reliable and informative than the settled point forecast, it is helpful to investigate and evaluate the data.

Experiment III: Interval forecasting and evaluation index.

To comprehensively evaluate the forecast results of the prediction model, four analysis indexes are adopted in the study, including the prediction interval coverage probability (**PICP**), coverage width criterion (**CWC**), prediction band width (**PBW**), and PI normalized averaged width (**PINAW**).

Prediction interval coverage probability is the basic evaluation index to assess the overall probability of the actual value falling into the **PBW**, and it is expressed as follows:

$$\mathbf{PICP} = \left(\frac{1}{n}\sum_{i=1}^{n}\overline{\overline{C}}_i\right) \times 100\% \qquad (4-3-1)$$

$$\overline{\overline{C}}_i = \begin{cases} 0\ \mathbf{y}_i \notin (\boldsymbol{L}_i, \boldsymbol{U}_\mathbf{I}) \\ 1\ \mathbf{y}_i \in (\boldsymbol{L}_\mathbf{I}, \boldsymbol{U}_\mathbf{I}) \end{cases} \qquad (4-3-2),$$

Where the variable y_i is the actual air contamination value. U_i and L_i represent the upper and lower bounds, respectively, and n is the number of samples.

If the PI width is sufficiently large, the PICP can easily reach 100%. Considering that the PICP meets the prediction interval nominal confidence (**PINC**) of the required prediction interval, the PBW should be as small as possible to guarantee the prediction effect.

Due to the contradiction between **PICP** and **PINRW**, **CWC** is used as a comprehensive evaluation index. In addition, because **PICP** is a basic evaluation indicator compared with **PINAW** (or **PINRW**) and is expected to achieve the desired nominal confidence level $\boldsymbol{\mu}$, an improved **CWC** design can assess the prediction effect better.

The modified CWC used in the experiment is defined as follows:

$$\mathrm{CWC}_{proposed} = \begin{cases} \boldsymbol{\eta}_1 \cdot \mathbf{PINAW}(\mathrm{PICP} \geq \boldsymbol{\mu}) \\ (0.1 + \boldsymbol{\mu}_1 \cdot \mathbf{PINAW})[1 + \exp(\boldsymbol{\eta}_2(\boldsymbol{\mu} - \mathbf{PICP}))] \\ (\mathrm{PICP} < \boldsymbol{\mu}) \qquad (4-3-3) \end{cases}$$

PBW, **PICP**, **PINAW**, and **CWC** are used to assess the **IP** performance. The detailed forecasting results of the proposed hybrid forecasting system and the comparative models are presented in **Figure 4** and **Table 8**.

(a) After optimization using the combined prediction algorithm, the interval prediction can estimate the upper and lower bounds of the probability prediction. Then, **PICP**, **PINAW**, and **CWC** indicators are selected to measure the performance of interval prediction of air pollution series. **PICP** mainly measures the probability of the original data entering the prediction interval, and **PINAW** is used to evaluate the normalized average width of the interval. This section adopts the interval prediction of $PM_{2.5}$ and PM_{10}, and the obtained results provide a practical application for analyzing the uncertainty of air pollution.

TABLE 7 | The forecasting performances of various models for SO_2 and CO in three cities.

Dataset (SO_2)	Model	MAPE (×100%)	MSE	Adjusted *R* square	RMSE	SSE
Beijing	ARIMA	16.9679992	0.594247371	0.214374179	0.77087442	1.19E + 02
	LSTM	62.5043641	3.575070593	−3.7264286	1.890785708	7.15E + 02
	TCN	21.19875252	0.358244808	−1.246512939	0.5985	48.7212939
	Proposed model	**8.801941086**	**0.017101509**	**−4.775512824**	**0.307953961**	**3.129576183**
Shanghai	ARIMA	16.804447	1.47013432	0.242891519	1.212490957	2.94E + 02
	LSTM	33.6266108	4.642820638	−1.391018856	2.154720548	9.29E + 02
	TCN	17.30624383	1.627779555	−0.25163582	1.2758	221.3780194
	Proposed model	**13.3948766**	**0.120567162**	**−3.960236439**	**0.817679472**	**22.06379071**
Shenzhen	ARIMA	13.93960103	0.183750043	0.68076782	0.428660755	36.75000861
	LSTM	7.4051157	0.003636823	0.659593382	0.060306074	0.727364522
	TCN	7.401322957	0.323880616	0.047690701	0.5691	44.0477638
	Proposed model	**4.5874185**	**0.143997793**	**−7.824128059**	**0.893606857**	**26.35159606**
Dataset (CO)						
Beijing	ARIMA	55.82851595	0.054564355	0.005877223	0.233590143	10.91287095
	LSTM	63.6736338	0.123604838	−1.251990073	0.351574798	24.72096768
	TCN	39.54360038	0.062681733	−0.71777545	0.2504	8.524715657
	Proposed model	**36.0314606**	**0.027527125**	**−3.06026958**	**0.390705028**	**5.037463826**
Shanghai	ARIMA	15.8161825	0.017936156	0.224697526	0.133925933	3.587231123
	LSTM	23.8757882	0.041810611	−0.807291964	0.204476431	8.362122138
	TCN	18.00573321	0.033547157	−0.834057427	0.1832	4.562413404
	Proposed model	**11.03447391**	**0.011493365**	**−5.983131272**	**0.07983478**	**0.210328539**
Shenzhen	ARIMA	22.46848191	0.379235401	0.341147671	0.615820916	75.84708014
	LSTM	9.8188329	0.005360446	0.498262188	0.07321507	1.072089292
	TCN	11.01559594	0.007725412	0.046786078	0.0879	1.050656042
	Proposed model	**6.6615594**	**0.003953843**	**−1.79785567**	**0.148073826**	**0.723553317**

The indicators can be defined as $\mathbf{SSE} = \sum_{i=1}^{N}(\mathbf{PP}_i - \mathbf{AP}_i)^2$, $\mathbf{RMSE} = \sqrt{\left[\sum_{i=1}^{N}\mathbf{PP}_i - \mathbf{AP}_i\right]/N}$, *and* $\mathbf{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|(\mathbf{AP}_i - \mathbf{PP}_i)/\mathbf{AP}\right| \times 100\%$. *The most satisfactory results are shown in bold.*

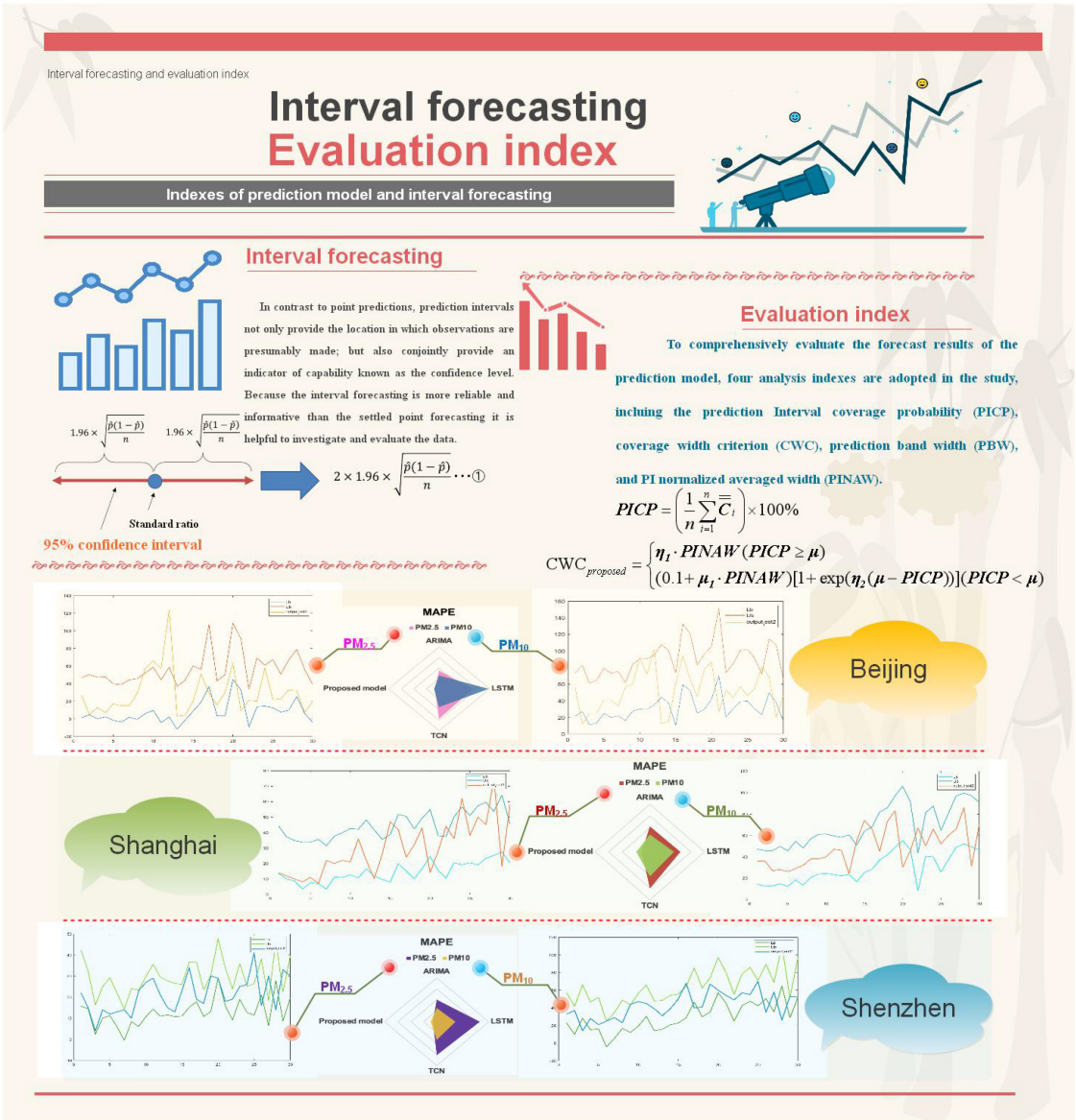


FIGURE 4 | A graph of interval prediction results.

(b) **Table 8** show the daily probability prediction evaluation results of $PM_{2}.5$ and PM_{10} in Beijing, Shanghai, and Shenzhen. It can be seen that the results of a single model interval are not good because the result of the width is too narrow and the coverage rate is low, while the interval coverage rate of the combined forecasting system proposed in this study is higher than that of a single model and the results are more accurate.

(c) It is difficult to satisfy all the optimal conditions due to the large number of indicators that measure the performance of interval prediction. However, the higher the confidence, the greater the coverage probability and the wider the interval. Therefore, the probability forecast has a certain prediction interval, which provides a reference for the actual application of air pollution monitoring.

Discussion

In this section, we used three methods to discuss the performance of the proposed combined forecasting system: forecasting effectiveness (FE), stability analysis (SA), and improvement ratio (IR).

Forecasting Effectiveness

To verify the availability of the relevant prediction system, the finite element method (Banik et al., 2020) is adopted in this study. This may be determined using the expected result of the prediction accuracy series, that is, the deviation between the expected value and normal deviation. The indicator is explained as follows.

Count the d-**th** order predicting availability element $m^d = \sum_{i=1}^{n} Q_i A_i^d$, where A_i is the prediction accuracy, Q_i is the discrete

TABLE 8 | Interval forecasting results of the proposed system for daily $PM_{2.5}$ and PM_{10}concentrations of Beijing, Shanghai, and Shenzhen.

$PM_{2.5}$	*Indicators*	Proposed model	ARIMA	LSTM	TCN
Site 1	**BW**	**27.27709623**	5.588932739	5.400260609	9.399792741
	CP	**0.8**	0.1	0.2	0.4
	CWC	**967.5573339**	4.62E + 17	3.30E + 15	4.05863E + 11
	PINAW	**0.534845024**	0.160988443	0.170554714	0.462680511
Site 2	**BW**	**25.83786828**	3.101276071	1.711007768	2.246018338
	CP	**0.7**	0.2	0.1	0.1
	CWC	**1216.366544**	1.23E + 15	9.11E + 16	3.79E + 17
	PINAW	**0.453295935**	0.063429076	0.031777656	0.132020441
Site 3	**BW**	**13.02474235**	1.822308173	2.649576346	2.393098477
	CP	**0.7**	0.1	0.1	0.2
	CWC	**1294.458165**	2.17E + 17	−2.06E + 17	1.50E + 15
	PINAW	**0.482397865**	0.075831053	−0.071902246	0.077871318
PM_{10}	***Indicators***	**Proposed model**	ARIMA	LSTM	TCN
Site 1	**BW**	**40.37286522**	2.864634888	2.867020268	7.299477064
	CP	**0.8**	0.1	0.3	0.2
	CWC	**1197.315173**	3.11E + 17	2.5209E + 13	5.96E + 15
	PINAW	**0.66185025**	0.108333022	0.193635382	0.30842717
Site 2	**BW**	**39.53693477**	3.204081672	18.80110089	4.136207351
	CP	**0.8**	0.3	0.4	0.3
	CWC	**1300.436217**	7.86722E + 12	3.06843E + 11	3.62306E + 13
	PINAW	**0.718853359**	0.060429758	0.349798846	0.278294498
Site 3	**BW**	**36.49380793**	5.868071455	4.509674695	2.006110094
	CP	**0.8**	0.3	0.1	0.1
	CWC	**1500.428327**	2.04898E + 13	1.78E + 17	3.38E + 17
	PINAW	**0.829404726**	0.157386217	0.062204966	0.11778856

The indicators can be defined as $\tilde{\lambda}_{cp} = \frac{1}{D}\sum_{i=1}^{D} c_i \times 100\%$ *,* $X_{MWP} = \frac{1}{D}\sum_{i=1}^{D} \frac{U(X_i)-L(X_i)}{f_i} \times 100\%$, $\overline{\overline{X}}_{BW} = \sum_{i=1}^{D} \xi_i$, *and* $\overline{\overline{X}}_{PINAW} = \left[1-(1+\lambda \times \bar{\bar{X}}_{AWD})X_{MWP} \times (1+e^{-\eta(X_{cp}-\mu)})\right] \times 100\%$. *The most satisfactory results are shown in bold.*

TABLE 9 | Forecasting effectiveness (FE) of different models ($PM_{2.5}$).

	Beijing		Shanghai		Shenzhen	
	1-Order	**2-Order**	**1-Order**	**2-Order**	**1-Order**	**2-Order**
Proposed model	**0.713962612**	**0.499107698**	**0.833677634**	**0.674130951**	**0.827274422**	**0.715910344**
Model ($PM_{2.5}$)						
ARIMA	0.537758525	0.365927564	0.626486448	0.434454329	0.687655842	0.501227795
LSTM	0.419353392	0.262811053	0.620248042	0.442276111	0.552618069	0.371521413
TCN	0.380004231	0.25407457	0.509236577	0.344293544	0.58990613	0.408951676
Model (PM_{10})						
Proposed model	**0.820324675**	**0.612407931**	**0.892528922**	**0.780386906**	**0.882115673**	**0.798110303**
ARIMA	0.584812738	0.406227792	0.71977323	0.539768394	0.776748883	0.625024691
LSTM	0.21938475	0.147504326	0.676791129	0.488168419	0.740723134	0.585249795
TCN	0.488147467	0.319256144	0.56964758	0.390169063	0.633711865	0.453582124

probability distribution, and $\sum_{i=1}^{n} \boldsymbol{Q_i} = \mathbf{1}$, $\boldsymbol{Q_i} > 0$. Since we could not obtain any prior information on $\boldsymbol{Q_i}$, it is determined as $\boldsymbol{Q_i} = 1/\boldsymbol{n}$, $\boldsymbol{i} = 1, 2, \ldots, \boldsymbol{n}$. The other $\boldsymbol{A_i}$ is calculated using $\boldsymbol{A_i} = 1 - \left|\overline{\overline{\xi}}_{\boldsymbol{i}}\right|$, in which $\overline{\overline{\xi}}_i$ is expressed as:

$$\overline{\overline{\xi}}_i = \begin{cases} -1, & (AP_i - PP_i)/AP_i < -1 \\ (AP_i - PP_i)/AP_i, & -1 \le (AP_i - PP_i)/AP_i < 1 \\ 1, & (AP_i - PP_i)/AP_i > 1 \end{cases}, (4-4-1)$$

where $\boldsymbol{PP_i}$ and $\boldsymbol{AP_i}$ indicate the $\boldsymbol{i}$**-th** point forecast value and observation quantitative value, respectively.

Thereafter, the continuous function $\overline{\overline{\boldsymbol{L}}}\left(\boldsymbol{g}^1, \boldsymbol{g}^2, \ldots, \boldsymbol{g}^d\right)$ of a ***d-order*** unit is introduced to assess the $\boldsymbol{d}$**-th** order predicting availability. While there is only one variable in the equation $\overline{\overline{\boldsymbol{L}}}(\boldsymbol{z}) = \boldsymbol{z}$, the *first*-order FE can be expressed as $\overline{\overline{\boldsymbol{L}}}\left(g^1\right) = g^1$.

If there are two variables in this equation, for instance, $\overline{\overline{L}}(i,j) = i\left(1-\sqrt{j-i^2}\right)$, the second-order FE can be expressed as follows:

$$\overline{\overline{L}}\left(g^1,g^2\right) = g^1\left(1-\sqrt{g^2-\left(g^1\right)^2}\right) \qquad (4-4-2)$$

According to the FE definition, the higher the value of $\overline{\overline{L}}$, the better the prediction performance of the models.

Therefore, the ***d-th*** order FE is expressed as $\boldsymbol{H}\left(\boldsymbol{m}^1,\boldsymbol{m}^2,\cdots,\boldsymbol{m}^k\right)$. Thus, the *first*-order prediction effectiveness is defined as $\boldsymbol{H}\left(\boldsymbol{m}^1\right)=\boldsymbol{m}^1$. If there are two variables in the equation, the second-order FE is given by $\boldsymbol{H}\left(\boldsymbol{m}^1,\boldsymbol{m}^2\right)=\boldsymbol{m}^1\left(1-\sqrt{\boldsymbol{m}^2-\left(\boldsymbol{m}^1\right)^2}\right)$.

By comparing the FE values with those of other related models, it can be easily concluded that the proposed system obtains the highest index value in both the first-order and second-order calculations, which shows that its performance in air pollution prediction exceeds that of other models. Specifically, we took Beijing PM_{10} data as an example in one-step, $\boldsymbol{H}_1^{\boldsymbol{Beijing}} = \mathbf{0.820324675}$ and $\boldsymbol{H}_2^{\boldsymbol{Beijing}} = \mathbf{0.612407931}$ in two orders, and the FE is much larger than that of other models. In other predictions, our proposed system exhibits the best forecast performance compared with the other models. The specific experimental results of the other models are listed in **Table 9**.

Sensitivity Analysis

In this section, the sensitivity of the proposed prediction system is analyzed experimentally. Since the weight determination method plays an important role in the final prediction, this study discusses the prediction sensitivity of the combined prediction model by adjusting the optimization parameters. In the parameter setting stage, the important parameter of the population size has a great influence on the optimization performance. Therefore, the experiment adopts the method of changing one parameter to examine its influence on the prediction result. Here, the size of the population is set to 40, 60, 80, and 100 in turn. The specific experimental results are shown in **Table 10**. The relevant conclusions are summarized as follows:

From **Table 10**, it can be seen that the performance of the proposed mode is different under various parameter settings. For example, in the Shenzhen PM2.5 forecast, MAPE values range from 5.9821 to 6.4834%.

Consequently, the fluctuation range of the forecast values in the three regions is small, indicating that the forecast system is less sensitive to the two modes and has a good stability in practical applications.

Improvement Ratio

In this section, the effectiveness of the combined forecasting model system is analyzed by the percentage improvement of MAPE and MSE. We proposed an index $\boldsymbol{IR_{MAPE}}$ to measure the improvement in the **PCFM** prediction accuracy. $\boldsymbol{IR_{MAPE}}$ can be

TABLE 10 | Experimental results of forecasting results under different population sizes.

Dataset	PM2.5				PM10			
	40	60	80	100	40	60	80	100
Beijing								
MAPE (×100%)	16.7949636457552	17.7356139438864	16.5110612195599	14.5260609900006	19.9055947149062	18.9704302780156	20.0524797526017	21.1438211511366
RMSE	24.6970552160974	25.3686235730786	24.4857948286102	23.3129364646248	24.9616400286964	24.2687122670146	25.1103732743030	26.0590974636506
SSE	30497.2268173483	32178.3530996280	29977.7074194398	27174.6503301816	31154.1736461110	29448.5197549573	31526.5422987416	33953.8280310020
Shanghai								
MAPE (×100%)	11.7258101883320	11.5430390511560	11.6253007536679	11.5970879035290	25.7764408550232	22.7232812706143	14.6954959388701	13.8943981605036
RMSE	15.6570375896180	15.3272904374202	15.5085690707090	15.4803567790693	29.5040374762424	26.4801510616650	18.4758225872704	17.5545394655169
SSE	12257.1413041355	11746.2916076516	12025.7857310475	11982.0723003638	43524.4113699757	35059.9200124298	17067.8010138146	15408.0927923195
Shenzhen								
MAPE (×100%)	5.9821028059 6730	6.1100433517 6550	6.4834299242 8906	6.2499767640 4358	9.3955940572 2180	9.6249946829 7147	10.1251000479952	10.0495620103120
RMSE	7.7788584591 5623	8.0177187303 4480	8.5561268049 5358	8.2522204788 6010	11.8654999185498	12.0187644430554	12.4795681613774	12.4454433453428
SSE	3025.5319463 7932	3214.1906819 4609	3660.3652951 2225	3404.9571415 8590	7039.5044158 5532	7222.5349368 8267	7786.9810747 2327	7744.4530031 0681

The indicators can be defined as $\boldsymbol{SSE} = \sum_{i=1}^{N}(\boldsymbol{PP}_i - \boldsymbol{AP}_i)^2$, $\boldsymbol{RMSE} = \sqrt{\left[\sum_{i=1}^{N}\boldsymbol{PP}_i - \boldsymbol{AP}_i\right]/\boldsymbol{N}}$, *and* $\boldsymbol{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|(\boldsymbol{AP}_i - \boldsymbol{PP}_i)/\boldsymbol{AP}\right| \times 100\%$. *The most satisfactory results are shown in bold.*

TABLE 11 | Improvement ratio (IR) for CO of different models (×100%).

	Model	Beijing	Shanghai	Shenzhen
$\bar{\bar{P}}_{MAPE}$	**ARIMA**	35.460472%	30.233%	70.3515%
	LSTM	43.412275%	53.7838%	32.155%
	TCN	8.8817%	38.71689%	39.526%
$\bar{\bar{P}}_{MSE}$	**ARIMA**	49.551%	35.92069%	98.957%
	LSTM	77.729%	72.51089%	26.2404%
	TCN	56.084%	65.7396%	48.8203%

The ***MAPE*** *×100% of the* ***proposed prediction system is*** $MAPE^{Beijing} = 36.0314606$, $MAPE^{Shanghai} = 11.03447391$ *and* $MAPE^{Shenzhen} = 6.6615594$.

expressed as:

$$\bar{\bar{P}}_{MAPE} = \left[\left(MAPE^{com} - MAPE^{pro}\right) / MAPE^{com}\right] \times 100\%$$

$$\bar{\bar{P}}_{MSE} = \left[\left(MSE^{com} - MSE^{pro}\right) / MSE^{com}\right] \times 100\%,$$

where $MAPE^{com}$ is the compared model **MAPE** values and $MAPE^{pro}$ indicates **MAPE** values of the prediction system. Moreover, the three models are compared with different indicators, which shows the superiority of the combined forecasting model system. The detailed calculation results are shown in **Table 11**.

(a) The model is improved by **MAPE**, which verifies the superiority of the proposed prediction system. Compared with the **ARIMA** model, this model improves by 35.460472%. For the **LSTM** model, the combined prediction system achieves a 43.412275% improvement in **MAPE**. The results show that the system has a good prediction effect on $PM_{2.5}$ and PM_{10}.

(b) For the three urban datasets considering mean square error (**MSE**) and **MAPE**, the proposed prediction system still achieves a significant improvement in prediction accuracy. Experiments show that, compared with the **TCN** model (taking Beijing as an example), the **MAPE** of the combined model is improved by 56.084%.

CONCLUSION

Predicting air quality plays a vital role in the environment and economy of energy development, which is widely discussed worldwide. In recent years, more researchers have focused on the methods to forecast $PM_{2.5}$ and PM_{10} concentrations and provide useful information for the citizens in their daily lives. However, to overcome the limitations and negative effects of an individual approach, this study develops a novel combined forecasting system that takes advantage of data preprocessing, single models, and the interval predicting approach.

The developed system includes an advanced data denoising technique, three single forecasting algorithms, and an optimization approach to predict the $PM_{2.5}$ and PM_{10} concentrations. Based on the experiments, we concluded that the combined model has the following advantages: (1) as for data denoising strategy, the combined system computes the data series without fluctuation and uncertainty by **FIG**, which yields better performance compared with single models by decomposing and reconstructing the initial data. (2) In the comparative experiments, to predict the $PM_{2.5}$ and PM_{10} concentrations of three cities, we found that the $PM_{2.5}$ **MAPE** (×100%) values of the proposed system are **17.53130124**, **11.52643852**, and **6.00510985**, which provide more satisfying results than the **ARIMA** models (**86.97079038**, **43.06089753**, and **40.30779213**). (3) Consequently, **MODOA** is utilized as an advanced optimization algorithm to determine the weight of every single model and to obtain the forecasting values of $PM_{2.5}$ and PM_{10} concentrations.

The proposed early warning system has many practical applications, such as warning and guiding the public before the occurrence of harmful air pollutants and mining the characteristics of air pollutants.

(1) The fuzzy preference rough set was applied to the early warning system to determine the main pollutants suitable for different cities. Attribute selection simplifies the process of early warning systems and makes the prediction of pollutants more effective. In addition, these results can help decision-makers in relevant sectors to monitor and analyze certain polluting pollutants, which play a crucial role in formulating effective strategies for each city.

(2) In the developed early warning system, the interval forecast based on deterministic forecast provides the forecast range and the confidence level, which can be used to analyze and monitor the uncertainty information of the future value of pollutants. Air quality warning systems trigger alerts when air pollution exceeds an upper limit. According to the forecast range, different early warning levels can also be divided as a guide for daily life.

Therefore, we concluded that the proposed combined predicting system enhances the forecasting capacity and accuracy of $PM_{2.5}$ and PM_{10} concentrations by conducting and analyzing the experiments. Accurate forecasts not only reduce the cost and risk of dealing with air pollution systems but also help policymakers come up with effective strategies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.tianqihoubao.com.

AUTHOR CONTRIBUTIONS

ZL contributed to the conception and design of the study. JW organized the model development and the experimental analysis and wrote the first draft of the manuscript. JL wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This study was supported by the Science and Technology Development Fund, FDCT, Macau SAR, under Grant 0064/2021/A2.

REFERENCES

Air Quality Expert Group (2012). *Fine particulate matter (PM2.5) in the UK. Technical report*. Available online at: http://uk-air.defra.gov.uk [accessed on March 19, 2022]

Akyüz, M., and Çabuk, H. (2009). Meteorological variations of PM2.5/PM10 concentrations and particle-associated polycyclic aromatic hydrocarbons in the atmospheric environment of Zonguldak, Turkey. *J. Hazard. Mater.* 170, 13–21. doi: 10.1016/j.jhazmat.2009.05.029

Bai, Y., Li, Y., Zeng, B., Li, C., and Zhang, J. (2019). Hourly PM2. 5 concentration forecast using stacked autoencoder model with emphasis on seasonality. *J. Clean. Prod.* 224, 739–750. doi: 10.1016/j.jclepro.2019.03.253

Banik, A., Behera, C., Sarathkumar, T. V., and Goswami, A. K. (2020). Uncertain wind power forecasting using LSTM-based prediction interval. *IET Renew. Power Gener.* 14, 2657–2667. doi: 10.1049/iet-rpg.2019.1238

Beaulant, A. L., Perron, G., Kleinpeter, J., Weber, C., Ranchin, T., and Wald, L. (2008). Adding virtual measuring stations to a network for urban air pollution mapping. *Environ. Int.* 34, 599–605. doi: 10.1016/j.envint.2007.12.004

Bergen, S., Sheppard, L., Sampson, P. D., Kim, S. Y., Richards, M., Vedal, S., et al. (2013). A national prediction model for PM2.5 component exposures and measurement error-corrected health effect inference. *Environ. Health Perspect.* 121, 1017–1025. doi: 10.1289/ehp.1206010

Bin, H., Zu, Y. X., and Zhang, C. (2014). A forecasting method of short-term electric power load based on BP neural network. *Appl. Mech. Mater.* 538, 247–250. doi: 10.4028/www.scientific.net/AMM.538.247

Borrego, C., Cascão, P., Lopes, M., Amorim, J. H., Tavares, R., Rodrigues, V., et al. (2011). Impact of urban planning alternatives on air quality: URBAIR model application. *WIT Trans. Ecol. Environ.* 147, 13–24. doi: 10.2495/AIR110021

Chen, Y., Kang, Y., Chen, Y., and Wang, Z. (2020). Probabilistic forecasting with temporal convolutional neural network. *Neurocomputing* 399, 491–501. doi: 10.1016/j.neucom.2020.03.011

Ding, S. (2019). A novel discrete grey multivariable model and its application in forecasting the output value of China's high-tech industries. *Comput. Ind. Eng.* 127, 749–760. doi: 10.1016/j.cie.2018.11.016

Djalalova, I., Delle Monache, L., and Wilczak, J. (2015). PM2. 5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model. *Atmos. Environ.* 108, 76–87. doi: 10.1016/j.atmosenv.2015.02.021

Genc, D. D., Yesilyurt, C., and Tuncel, G. (2010). Air pollution forecasting in Ankara, Turkey using air pollution index and its relation to assimilative capacity of the atmosphere. *Environ. Monit. Assess.* 166, 11–27. doi: 10.1007/s10661-009-0981-y

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: continual prediction with LSTM. *Neural Comput.* 12, 2451–2471. doi: 10.1162/089976600300015015

Guo, J., Liu, Z., Huang, W., Wei, Y., and Cao, J. (2018). Short-term traffic flow prediction using fuzzy information granulation approach under different time intervals. *IET Intell. Transp. Syst.* 12, 143–150. doi: 10.1049/iet-its.2017.0144

Hamilton James, D. (2015). *Time Series Analysis*. Beijing: China Renmin University Press.

He, Y., Yan, Y., Wang, X., and Wang, C. (2019a). Uncertainty forecasting for streamflow based on support vector regression method with fuzzy information granulation. *Energy Procedia* 158, 6189–6194. doi: 10.1016/j.egypro.2019.01.489

He, Y., Yan, Y., and Xu, Q. (2019b). Wind and solar power probability density prediction via fuzzy information granulation and support vector quantile regression. *Int. J. Electr. Power Energy Syst.* 113, 515–527. doi: 10.1016/j.ijepes.2019.05.075

Li, J., Wang, R., Wang, J., and Li, Y. (2018). Analysis and forecasting of the oil consumption in China based on combination models optimized by artificial intelligence algorithms. *Energy* 144, 243–264. doi: 10.1016/j.energy.2017.12.042

Liu, H., Jin, K., and Duan, Z. (2019). Air PM2.5 concentration multi-step forecasting using a new hybrid modeling method: comparing cases for four cities in China. *Atmos. Pollut. Res.* 10, 1588–1600. doi: 10.1016/j.apr.2019.05.007

Liu, W., Guo, G., Chen, F., and Chen, Y. (2019). Meteorological pattern analysis assisted daily PM2.5 grades prediction using SVM optimized by PSO algorithm. *Atmos. Pollut. Res.* 10, 1482–1491. doi: 10.1016/j.apr.2019.04.005

Majumder, H., and Maity, K. (2018). Application of GRNN and multivariate hybrid approach to predict and optimize WEDM responses for Ni-Ti shape memory alloy. *Appl. Soft Comput.* 70, 665–679. doi: 10.1016/j.asoc.2018.06.026

McKeen, S. A., Chung, S. H., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., et al. (2007). Evaluation of several PM2.5 forecast models using data collected during the ICARTT/NEAQS 2004 field study. *J. Geophys. Res. Atmos* 112, 1–20. doi: 10.1029/2006JD007608

Mencar, C., and Fanelli, A. M. (2008). Interpretability constraints for fuzzy information granulation. *Inf. Sci.* 178, 4585–4618. doi: 10.1016/j.ins.2008.08.015

Osowski, S., and Garanty, K. (2007). Forecasting of the daily meteorological pollution using wavelets and support vector machine. *Eng. Appl. Artif. Intell.* 20, 745–755. doi: 10.1016/j.engappai.2006.10.008

Pai, T. Y., Hanaki, K., and Chiou, R. J. (2013). Forecasting hourly roadside particulate matter in taipei county of taiwan based on first-order and one-variable grey model. *Clean Soil Air Water.* 41, 737–742. doi: 10.1002/clen.201000402

Samia, A., Kaouther, N., and Abdelwahed, T. (2012). A hybrid ARIMA and artificial neural networks model to forecast air quality in urban areas: case of Tunisia. *Adv. Mater. Res.* 518, 2969–2979. doi: 10.4028/www.scientific.net/AMR.518-523.2969

Sfetsos, A. (2000). A comparison of various forecasting techniques applied to mean hourly wind speed time series. *Renew. Energy* 21, 23–35. doi: 10.1016/S0960-1481(99)00125-1

Shelhamer, E., Long, J., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 3431–3440. doi: 10.1109/TPAMI.2016.2572683

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

State of Global Air (2020). *Health Effects Institute*. Available online at: https://www.stateofglobalair.org [accessed on October 22, 2020]

Sun, W., and Li, Z. (2020). Hourly PM2.5 concentration forecasting based on feature extraction and stacking-driven ensemble model for the winter of the Beijing-Tianjin-Hebei area. *Atmos. Pollut. Res.* 11, 110–121. doi: 10.1016/j.apr.2020.02.022

Tian, Q., and Wang, H. (2021). Predicting remaining useful life of rolling bearings based on reliable degradation indicator and temporal convolution network with the quantile regression. *Appl. Sci.* 11:4773. doi: 10.3390/app11114773

van Donkelaar, A., Martin, R. V., and Park, R. J. (2006). Estimating ground-level PM2.5 using aerosol optical depth determined from satellite remote sensing. *J. Geophys. Res. Atmos* 111:D21201. doi: 10.1029/2005JD006996

Wakamatsu, S., Morikawa, T., and Ito, A. (2013). Air pollution trends in Japan between 1970 and 2012 and impact of urban air pollution countermeasures. *Asian J. Atmos. Environ.* 7, 177–190. doi: 10.5572/ajae.2013.7.4.177

Wang, J., and Hu, J. (2015). A robust combination approach for short-term wind speed forecasting and analysis - Combination of the ARIMA (Autoregressive Integrated Moving Average), ELM (Extreme Learning Machine), SVM (Support Vector Machine) and LSSVM (Least Square SVM) forecasts using a GPR (Gaussian Process Regression) model. *Energy* 93, 41–56. doi: 10.1016/j.energy.2015.08.045

Wang, J., Zhang, L., and Li, Z. (2022b). Interval forecasting system for electricity load based on data pre-processing strategy and multi-objective optimization algorithm. *Appl. Energy* 305:117911. doi: 10.1016/j.apenergy.2021.117911

Wang, J., Wang, R., and Li, Z. (2022a). A combined forecasting system based on multi-objective optimization and feature extraction strategy for hourly PM2.5 concentration. *Appl. Soft Comput.* 114:108034. doi: 10.1016/j.asoc.2021.108034

Wang, J., Zhang, L., Wang, C., and Liu, Z. (2021). A regional pretraining-classification-selection forecasting system for wind power point forecasting and interval forecasting. *Appl. Soft Comput.* 113:107941. doi: 10.1016/j.asoc.2021.107941

Wang, S., Wang, J., Lu, H., and Zhao, W. (2021). A novel combined model for wind speed prediction – Combination of linear model, shallow neural networks, and deep learning approaches. *Energy* 234:121275. doi: 10.1016/j.energy.2021.121275

Wang, Y., Wang, J., Zhao, G., and Dong, Y. (2012). Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: a case study of China. *Energy Policy* 48, 284–294. doi: 10.1016/j.enpol.2012.05.026

Wang, Z., Wang, C., and Wu, J. (2016). Wind energy potential assessment and forecasting research based on the data pre-processing technique and swarm intelligent optimization algorithms. *Sustainability* 8:1191. doi: 10.3390/su8111191

Wei, D., Wang, J., Niu, X., and Li, Z. (2021). Wind speed forecasting system based on gated recurrent units and convolutional spiking neural networks. *Appl. Energy* 292:116842. doi: 10.1016/j.apenergy.2021.116842

Xiao, X., Xie, W., Zhou, Y., Zhao, W., Liu, X., and Zhang, C. (2019). Prediction and analysis of energy demand of high energy density AC/DC park based on spatial static load forecasting method. *J. Eng.* 2019, 3388–3391. doi: 10.1049/joe.2018.8389

Yang, H., Zhu, Z., Li, C., and Li, R. (2020). A novel combined forecasting system for air pollutants concentration based on fuzzy theory and optimization of aggregation weight. *Appl. Soft Comput. J.* 87:105972. doi: 10.1016/j.asoc.2019.105972

Yang, W., Yao, Q., Ye, K., and Xu, C. Z. (2020). Empirical mode decomposition and temporal convolutional networks for remaining useful life estimation. *Int. J. Parallel Program.* 48, 61–79. doi: 10.1007/s10766-019-00650-1

Zadeh, L. A. (1997). Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst.* 90, 111–127. doi: 10.1016/S0165-0114(97)00077-8

Zhang, X., Wang, J., and Gao, Y. (2019). A hybrid short-term electricity price forecasting framework: cuckoo search-based feature selection with singular spectrum analysis and SVM. *Energy Econ.* 81, 899–913. doi: 10.1016/j.eneco.2019.05.026

Zhang, Y., and Na, S. (2018). A novel agricultural commodity price forecasting model based on fuzzy information granulation and MEA-SVM model. *Math. Probl. Eng.* 2018, 1–10. doi: 10.1155/2018/2540681

Zhu, R., Liao, W., and Wang, Y. (2020). Short-term prediction for wind power based on temporal convolutional network. *Energy Rep.* 6, 424–429. doi: 10.1016/j.egyr.2020.11.219

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.

Copyright © 2022 Wang, Li and Li. 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 Ecology and Evolution

ORIGINAL RESEARCH
published: 02 May 2022
doi: 10.3389/fevo.2022.875000

Forecasting New COVID-19 Cases and Deaths Based on an Intelligent Point and Interval System Coupled With Environmental Variables

Zongxi Qu[1,2], Yongzhong Sha[1,2], Qian Xu[1,2] and Yutong Li[1,2]*

[1] School of Management, Lanzhou University, Lanzhou, China, [2] Research Center for Emergency Management, Lanzhou University, Lanzhou, China

The outbreak of Coronavirus disease 2019 (COVID-19) has become a global public health event. Effective forecasting of COVID-19 outbreak trends is still a complex and challenging issue due to the significant fluctuations and non-stationarity inherent in new COVID-19 cases and deaths. Most previous studies mainly focused on univariate prediction and ignored the uncertainty prediction of COVID-19 pandemic trends, which may lead to insufficient results. Therefore, this study utilized a novel intelligent point and interval multivariate forecasting system that consists of a distribution function analysis module, an intelligent point prediction module, and an interval forecasting module. Aimed at the characteristics of the COVID-19 series, eight hybrid models composed of various distribution functions (DFs) and optimization algorithms were effectively designed in the analysis module to determine the exact distribution of the COVID-19 series. Then, the point prediction module presents a hybrid multivariate model with environmental variables. Finally, interval forecasting was calculated based on DFs and point prediction results to obtain uncertainty information for decision-making. The new cases and new deaths of COVID-19 were collected from three highly-affected countries to conduct an empirical study. Empirical results demonstrated that the proposed system achieved better prediction results than other comparable models and enables the informative and practical quantification of future COVID-19 pandemic trends, which offers more constructive suggestions for governmental administrators and the general public.

Keywords: COVID-19, point forecasting, interval forecasting, artificial intelligence, environmental variables

OPEN ACCESS

Edited by:
Yan Hao,
Shandong Normal University, China

Reviewed by:
Yunxuan Dong,
University of Macau, China
Ling Xiao,
Xuzhou University of Technology, China

****Correspondence:***
Yongzhong Sha
shayzh@lzu.edu.cn

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Ecology and Evolution

Received: *13 February 2022*
Accepted: *25 March 2022*
Published: *02 May 2022*

Citation:
Qu Z, Sha Y, Xu Q and Li Y (2022) Forecasting New COVID-19 Cases and Deaths Based on an Intelligent Point and Interval System Coupled With Environmental Variables. Front. Ecol. Evol. 10:875000. doi: 10.3389/fevo.2022.875000

INTRODUCTION

Risk prevention and control of major infectious diseases are essential for human health and social stability. In recent years, with global warming, the deterioration of the ecological environment, and the acceleration of urbanization, an increasing number of pathogenic microorganisms have mutated, leading to the outbreak of major infectious diseases more frequently (Wu et al., 2017).

Abbreviations: ARIMA, auto regressive integrated moving average model; BPNN, back propagation neural network; GRNN, general regression neural network; ANFIS, Adaptive Neuro-Fuzzy Inference System; LSSVM, least square support vector machine; SCA, sine cosine algorithm; DFs, distribution functions; MLE, maximum likelihood estimation; TN-SCA-LSSVM, SCA-LSSVM with NO_2 and temperature; ECDC, European Center for Disease Prevention and Control; WAQI, World Air Quality Index project.

In December 2019, infectious pneumonia caused by a novel coronavirus disease (COVID-19) was discovered and quickly spread to more than 200 countries worldwide. With the global novel coronavirus epidemic becoming more serious, the World Health Organization raised the global risk of the COVID-19 epidemic to the highest level.

The COVID-19 epidemic was non-linear, dynamic, and fuzzy, thereby increasing the difficulty of prevention and control decision-making. Practical modeling approaches to predict the spread of a novel virus in the population play an essential role in the preparation and formulation of health and economic policies of any government or authority figure. When new cases increase at rates of thousands per day, health care systems of even the most developed countries are overwhelmed and unable to handle influxes of such large numbers of patients. In overwhelming situations, timely outbreak forecasting supports responsible agencies in being prepared and in managing the response effectively. For example, by targeting exclusion zones and scheduling economic activities, managing medical resources, and planning for emergency hospitals, effective forecasting is strategically essential for decision-makers (Swapnarekha et al., 2020).

Recently, various models have been developed to forecast the upcoming number of COVID-19 cases and its spread in the near future. Epidemiological models have been widely adopted in predicting COVID-19 cases and deaths. Many of these models were based on the traditional SEIR model and have been widely adopted (Li et al., 1999; Barmparis and Tsironis, 2020; He et al., 2020; Ndaïrou et al., 2020; Pandey et al., 2020). Additionally, statistical forecasting models, artificial intelligence (AI) models, and hybrid forecasting models have also been practical for epidemic prediction. For example, Ceylan (2020) applied auto regressive integrated moving average model (ARIMA) to forecast the epidemiological trend in Italy, Spain, and France. Ghosal et al. (2020) used linear and multiple linear regression methods to predict the number of deaths in India over a short period of 6 weeks. Moftakhar and Seif (2020) used the ARIMA model to forecast the number of patients with COVID-19 in Iran in the next 30 days. Ala'raj et al. (2021) developed a dynamic hybrid model based on SEIRD and ARIMA models to provide long- and short-term forecasts with confidence intervals. Ly (2020) employed an Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict COVID-19 cases in the United Kingdom. The results showed that data from Spain and Italy increased the ability to forecast COVID-19 cases in the United Kingdom. Borghi et al. (2021) used a machine learning model based on the multilayer Perceptron artificial neural network structure, which effectively predicted the behavior of four time series (accumulated infected cases, new cases, accumulated deaths, and new deaths). Parbat and Chakraborty (2020) used support vector regression (SVR) for a 60-day forecast of COVID-19 cases in India based on time-series data reported from March 01, 2020, to April 30, 2020. Meanwhile, the combination and mixing of different models have also regarded as effective ways to improve prediction, including applications in different fields, such as economic modeling and policy-making [18,19] (Stock and Watson, 2004; McAdam and McNelis, 2005), electricity price forecasting (Yang et al., 2022), environmental pollution (Hao et al., 2021), and COVID-19 forecasting (Castillo and Melin, 2020).

Although these methods have contributed significantly to the field of COVID-19 prediction, most of the models mainly focused on deterministic forecasts and ignored the uncertain information in the forecasts, resulting in the inability of the government disease control department to assess and manage epidemic risk. Additionally, one area of research has been on the impact of air pollution on new cases and deaths from COVID-19. It is known that air pollution can result in several diseases, including chronic respiratory diseases, stroke and cardiovascular problems. Recent studies have identified links between air pollution (mainly nitrogen oxides NO_2 and $PM_{2.5}$) and deaths and cases of COVID-19. Travaglio et al. (2021) explored potential links between air pollutants and COVID-19 mortality and infectivity. They found that air pollutant concentrations, especially nitrogen oxides and $PM_{2.5}$, were positively associated with COVID-19 mortality and infectivity. Konstantinoudis et al. (2021) used high geographical resolution to investigate the effect of long-term exposure to NO_2 and $PM_{2.5}$ on COVID-19 mortality in England. They found some evidence of an association of NO_2 with COVID-19 mortality, while the effect of long-term exposure to $PM_{2.5}$ remained uncertain. Lian et al. (2021) reported that urban lockdown was an effective method to reduce the number of new cases, and nitrogen dioxide (NO_2) concentrations can be used as an indicator of environmental lockdown to assess the effectiveness of lockdown measures. In some studies, the influence of meteorological parameters on the transmission of COVID-19 was discussed, and it was found that weather factors could affect the spread of COVID-19 (Malki et al., 2020; Shi et al., 2020). For example, Wu et al. (2020) analyzed the relationship between temperature change and n COVID-19 pneumonia and its impact on 166 countries. Wang et al. (2020) demonstrated that temperature can significantly modify the spread of COVID-19 to a certain extent and that there may be an optimal temperature for virus transmission. The above studies have pointed out the effects of environmental and meteorological factors on the survival and spread of the virus. A tremendous number of studies support that both nitrogen oxides and temperature play an important role in the spread and infection of COVID-19, motivating the current study to take environmental and meteorological factors into account in the prediction of COVID-19. We sought to determine whether the addition of these variables would improve the outbreak prediction.

Hence, by taking into consideration the results of the above works, this study utilized a novel point and interval data-driven forecasting model consisting of a distribution function analysis module, an intelligent point prediction module, and an interval forecasting module. First, several distribution functions (DFs) optimized by a metaheuristic algorithm were effectively designed to analyze the characteristics of the COVID-19 series. Furthermore, we used environmental features, such as nitrogen dioxide (NO_2) and temperature, as inputs to the multivariable hybrid prediction model, which is a combination of the sine cosine algorithm (SCA) and least square support vector machine (LSSVM). Based on the DFs and point forecasting results, interval

forecasting was designed to obtain uncertain information. The new case and new death series collected from the top three affected countries were used for the empirical study. We compared the performance of the best data-driven univariate model and the best multivariate model in an attempt to generate better predictions.

Our main contributions are as follows:

1 A practical epidemic analysis and prediction tool based on distribution function analysis, intelligent point prediction, and interval forecasting modules are proposed for the government and the public.
2 Environmental variables, such as NO_2 and temperature, were selected as inputs to construct a multivariable hybrid prediction model.
3 Interval forecasting based on DFs and point forecasting results can provide more uncertainty information for decision-making.

The rest of the paper is organized as follows. Section "Methodology" introduces the related Methodologies. Section "A Framework of the Developed Hybrid Forecasting System" describes the primary process of the proposed framework of the developed hybrid system. Section "Data Description and Evaluation Criteria" describes the research datasets and the evaluation criteria of this study. Section "Experimental Results and Analysis" discusses the forecasting results of the proposed model and the comparative results with other models. Finally, Section "Conclusion" concludes the critical conclusions of this paper.

METHODOLOGY

Some related methodologies are introduced in this section, including LSSVM, SCA, DFs, and interval prediction theory.

Least Squares Support Vector Machine

The support vector machine (SVM) proposed by Vapnik is an essential method in machine learning that effectively resolved pattern identification and classification tasks. The support vector machine is aimed at a small sample problem, is based on structural risk minimization, better solves the previous machine learning model overlearning, non-linear, dimensional disaster and local minimum problems, and has a good generalization ability. However, this method has some defects, such as slow training speed and poor stability when training samples on a large scale, limiting its application scope (quadratic programming problem needs to be solved in the learning process). Therefore, Suykens and Vandewalle (1999) proposed the least squares support vector machine (LSSVM) based on SVM, which significantly reduced the algorithm's computational complexity and improved the training speed. The LSSVM is an extension of the standard SVM. The algorithm transforms the solution of the support vector machine from a quadratic programming problem to linear equations. More details on the LSSVM can be found in Suykens and Vandewalle (1999).

It is worth noting that different types of kernel functions can be used in the LSSVM model, such as sigmoid, polynomial, and radial basis function (RBF), which are commonly used in the LSSVM model. RBF is a general choice of the kernel function proposed in Keerthi and Lin (2003), requiring fewer parameters and superior performance in applications. Accordingly, this study identifies RBF as the appropriate kernel function:

$$K\left(x_i, x_j\right) = \exp\left\{-\left|\left|x_j - x_i\right|\right|^2 / 2\,\sigma^2\right\} \tag{1}$$

Sine Cosine Algorithm

Mirjalili (2016) proposed the SCA, which is based on sine and cosine functions to explore different regions of the search space. It can effectively avoid local optimization, converge to global optimization, and effectively use the promising area of the search space during optimization. In SCA, the search space dimension is determined by the number of parameters required for optimization. The SCA creates different initial random agent solutions and requires them to use mathematical models based on sine and cosine functions to swing outward or toward the best solution.

$$\overrightarrow{x_i^{t+1}} = \overrightarrow{x_i^t} + rand_1 \times \sin\left(rand_2\right) \times \left|rand_3 \times l_i^t - \overrightarrow{x_i^t}\right| \tag{2}$$

$$\overrightarrow{x_i^{t+1}} = \overrightarrow{x_i^t} + rand_1 \times \cos\left(rand_2\right) \times \left|rand_3 \times l_i^t - \overrightarrow{x_i^t}\right| \tag{3}$$

where $\overrightarrow{x_i^t}$ is the current position at the tth iteration in the ith dimension, l_i^t is the targeted optimal global solution and $rand_1, rand_2, rand_3 \in [0, 1]$ are random numbers. Eqs. (2) and (3) use $0.5 \leq rand_4 < 0.5$ conditions for exploitation and exploration.

$$\overrightarrow{x_i^{t+1}} = \begin{cases} \overrightarrow{x_i^t} + rand_1 \times \sin\left(rand_2\right) \times \left|rand_3 \times l_i^t - \overrightarrow{x_i^t}\right|, rand_4 < 0.5 \\ \overrightarrow{x_i^t} + rand_1 \times \cos\left(rand_2\right) \times \left|rand_3 \times l_i^t - \overrightarrow{x_i^t}\right|, rand_4 \geq 0.5 \end{cases} \tag{4}$$

Distribution Functions

The probability distribution function has played an essential role in time series analysis, resource evaluation, and interval prediction in recent years. Researchers have tried to fit the basic characteristics of historical data by various DFs, hoping to mine the relevant characteristics, thereby deeply understanding data uncertainty. This study used the weibull distribution, gamma distribution, lognormal distribution, and Rayleigh DFs to study the statistical characteristics of new Covid-19 cases and deaths in three countries. The above DFs are shown in **Table 1**.

Interval Prediction Theory

Based on deterministic prediction, many studies (Song et al., 2015; Xu et al., 2017; Tian and Hao, 2020) have proposed interval prediction technology that can reflect the uncertain trend of future values to provide uncertain information about time series,

TABLE 1 | Four distribution functions.

Distribution functions	Equations	Parameters
Lognormal	$f(x,\mu,\sigma)=\frac{1}{x\sqrt{2\pi}\sigma}\exp\left(-\frac{(ln(x)-\mu)^2}{}\right)$	μ,σ
Gamma	$f(x,\xi,\theta)=\frac{x^{\xi-1}}{\theta^{\xi}\Gamma(\xi)}\exp\left(-\frac{x}{\theta}\right)$	σ
Weibull	$f(x,k,c)=\frac{k}{c}\left(\frac{x}{c}\right)^{k-1}\exp\left(-\left(\frac{x}{c}\right)^{k}\right)$	ξ,θ
Rayleigh	$f(x,\sigma)=\frac{1}{\sigma^2}\exp\left(-\frac{x^2}{2\pi\sigma^2}\right)$	k,c

such as air pollutants, wind energy, macroeconomic economy, and carbon trading prices. This type of interval prediction is a dynamic interval prediction method that calculates the uncertain information of future values based on point prediction and DFs. Therefore, the performance of the interval prediction model depends on the accuracy of the point prediction and the estimation of the distribution function. To be specific, assuming that the observation is Y_t, at the significance level α, the probability formula for the lower limit: L and upper limit: U can be expressed:

$$P\left(L_t \le Y_t \le U_t\right) = 1 - 2\alpha \tag{5}$$

The above formula can also be described by the following equation.

$$P\left\{L_t \le Y_t \le U_t \left| E\left(Y_t\right) = \widehat{y}\right.\right\} \cdot P\left\{E\left(Y_t\right) = \widehat{y}\right\} = 1 - 2\alpha \tag{6}$$

Additionally, we suppose that the forecasting values possess similar DFs with the historical datasets. Therefore, once the DFs of the original time series are determined, the estimated variance can be obtained. As a result, the values of the upper and lower bounds can be calculated with a certain confidence level$_\alpha$.

$$\left\{ \left(L_t, U_t\right) \middle| L_t \le Y_t \le U_t, \int_{L_t}^{U_t} f\left(t \left| \widehat{\Theta}\right.\right) dt = 1 - 2\alpha \right\} \tag{7}$$

The above equation can also be expressed as:

$$\begin{cases} [\widehat{L},\widehat{U}]=[\widehat{L},\widehat{y}][\widehat{y},\widehat{U}] \\ \int_{\widehat{L}}^{\widehat{y}} f\left(t|\widehat{\Theta}\right)dt+\int_{\widehat{y}}^{\widehat{U}} f\left(t|\widehat{\Theta}\right)dt=1-2\alpha \end{cases} \tag{8}$$

A FRAMEWORK OF THE DEVELOPED HYBRID FORECASTING SYSTEM

This section describes the details of the developed hybrid architecture framework, as shown in **Figure 1**. The framework consists of three modules: distribution function analysis, intelligent point prediction with environmental features, and interval forecasting.

Distribution Function Analysis Module

This module mainly implements characteristic data analysis of raw epidemic data. First, the Weibull distribution, Rayleigh distribution, Lognormal distribution, and Gamma distribution are introduced to fit the epidemic time series. To obtain the optimal estimation of model parameters, two different estimation methods, namely, maximum likelihood estimation (MLE) and a robust heuristic algorithm (SCA), are applied to evaluate the parameters of different DFs. Finally, the most suitable epidemic sequence distribution function is obtained by comparing the fitting ability of 8 hybrid probability DFs.

Intelligent Point Prediction Module With Environmental Features

The volatility and non-linearity of new cases and new deaths of COVID-19 make modeling very difficult. A successful predictive model requires optimization as well as sufficient data to drive it. Previous studies have shown that some environmental variables are highly correlated with epidemic changes, especially nitrogen dioxide and temperature, which have a significant impact on the epidemic trend of COVID-19 (Bauwens et al., 2020; Shi et al., 2020; Wang et al., 2020; Travaglio et al., 2021). Thus, we took environmental features, such as nitrogen dioxide (NO_2) and temperature, as inputs to construct a multivariable hybrid prediction model. To develop an intelligent point prediction model, we designed a LSSVM prediction model based on SCA optimization, namely, the hybrid SCA-LSSVM. Specifically, the SCA was introduced when training the LSSVM model, and the parameters (i.e., σ, γ) of the LSSVM model were optimized by the SCA algorithm to achieve high-performance forecasting.

Interval Forecasting Module

According to interval forecasting theory, interval prediction of the COVID-19 epidemic can be achieved based on the appropriate distribution function and point prediction values of COVID-19.

DATA DESCRIPTION AND EVALUATION CRITERIA

Data Description

The accuracy of the prediction mainly depends on the quality of the data and requires sufficient historical data. This study collected the data from the open dataset Our World in Data [Coronavirus (COVID-19) Cases – Our World in Data], which contains global daily data from the European Center for Disease Prevention and Control (ECDC). Due to the significant fluctuations and non-stationarity inherent in COVID-19, new case and death series bring great challenges to predictions. To verify the performance of the model, we used new cases per 100 thousand of the population per day as one of the predictive variables:

$$New\ cases\ per\ 100\ thousand = \frac{new\ cases\ per\ day}{Total\ Population} * 100,000 \tag{9}$$

The new deaths per thousand of the population calculated according to Equation (10) were also predicted based on available data.

$$New\ deaths\ per\ 100\ thousand = \frac{new\ deaths\ per\ day}{Total\ Population} * 100,000 \tag{10}$$

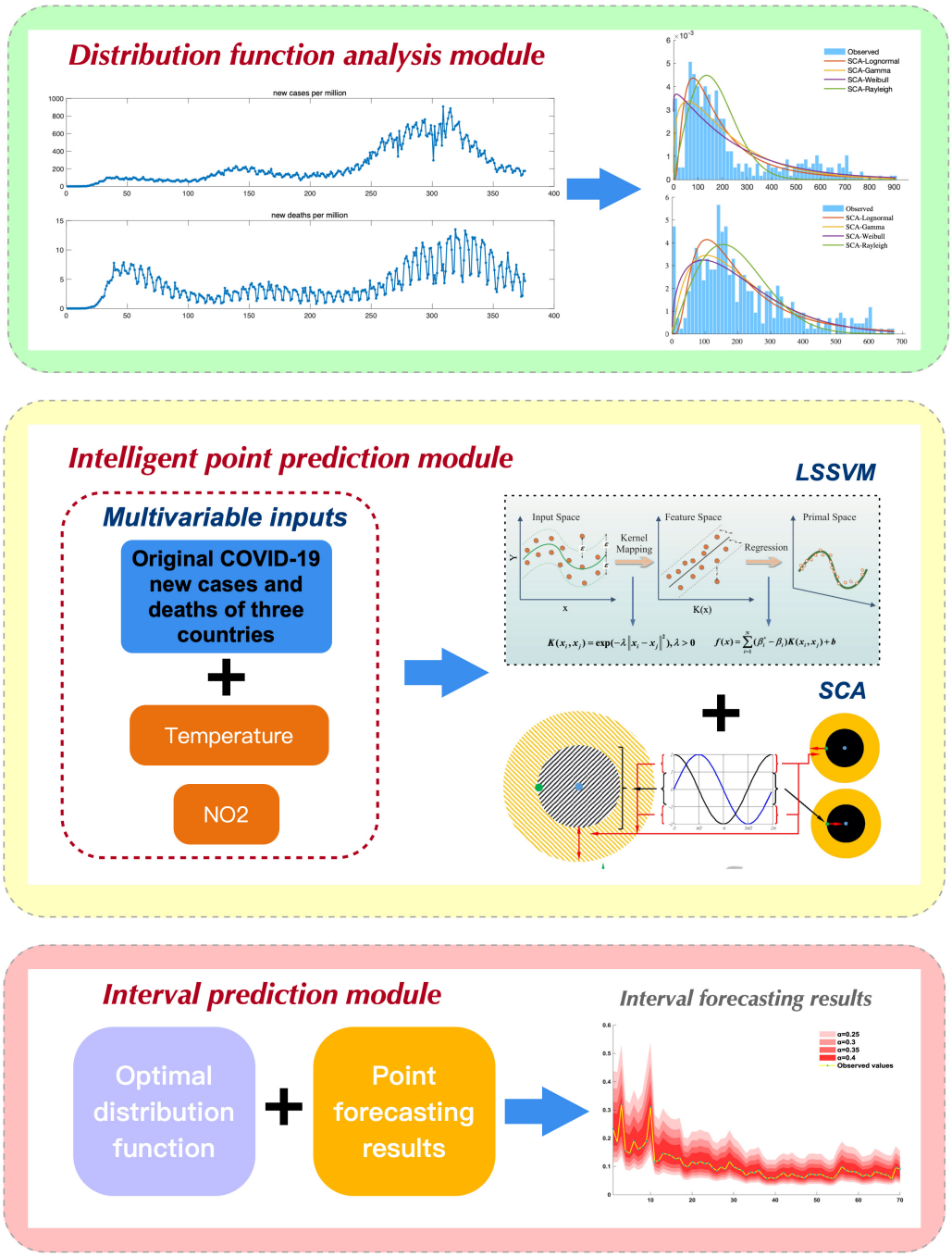


FIGURE 1 | The main procedure of the proposed system.

The World Air Quality Index project (WAQI) (Covid-19 Worldwide Air Quality data) provides a dataset covering air quality for more than 130 countries, updated daily starting in the first quarter of 2020. The dataset contains the data of each air pollutant, i.e., CO, NO_2, O_3, SO_2, PM_{10}, and $PM_{2.5}$, as well as meteorological data including humidity and temperature.

We focused on the three major countries that have been most strongly affected by COVID-19: the United States, India, and Brazil. The data of new cases and new deaths per 100 thousand of the population for the three countries, as well as the data of NO_2 and temperature for the same period, were selected as input variables for the outbreak modeling. Notably, the first observation time (or start time) and the length of the time series are different for each country. Sample data from the United States were collected from February 29, 2020, to March 10, 2021. Sample data from India were collected from March 18, 2020 to March

TABLE 2 | Eight evaluation rules.

Metric	Equation	Definition
MAE	$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}\left\lvert y_n - \hat{y}_n\right\rvert$	The average absolute forecast error of *n* times forecast results
RMSE	$\mathrm{RMSE} = \left(\frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{y}_n\right)^2\right)^{1/2}$	The root-mean-square forecast error
MAPE	$\mathrm{MAPE} = \frac{1}{N}\sum_{n=1}^{N}\left\lvert\frac{y_n - \hat{y}_n}{y_n}\right\rvert \times 100\%$	The average of absolute error
TIC	$\mathrm{TIC} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}y_n^2} + \sqrt{\frac{1}{N}\sum_{n=1}^{N}\hat{y}_n^2}}$	Theil's inequality coefficient
IA	$IA = 1 - \frac{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}\left(\lvert\hat{y}_n - \bar{y}\rvert + \lvert y_n - \bar{y}\rvert\right)^2}$	The index of agreement of forecasting results
R^2	$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)}{\sum_{n=1}^{N}(y_n - \bar{y})}$	Coefficient of determination
IFAW	$IFAW = \frac{1}{N}\sum_{n=1}^{N}(U_n - L_n)$	Interval forecasting average width
IFCP	$IFCP = \frac{1}{N}\sum_{n=1}^{N} b_n,\ b_n = \begin{cases}1, & \text{if } y_n \in [U_n, L_n] \\ 0, & \text{otherwise}\end{cases}$	Interval forecasting coverage probability

10, 2021. Sample data from Brazil were collected from March 17, 2020, to March 10, 2021. Sample data were divided into two parts: a training subset and a testing subset. We used 80% of the total data as the training subset and the remaining 20% as the test subset.

Evaluation Criteria

This study considered eight evaluation criteria to effectively evaluate the model's performance, as shown in **Table 2**. Specifically, the MAE, RMSE, and R2 were chosen as error criteria to determine the fitting level of these DFs. The MAE, RMSE, MAPE, IA, DA, and R2 were used to reflect the prediction performance of the point forecasting models. The PIAW and PICP were used to measure the validity of the interval prediction.

Here y_n and $\hat{y}_n$ represent the actual and predicted values at time n, respectively. N denotes the sample size. L_n and U_n are the lower and upper values of the interval forecasting, and b_n means a Boolean value.

EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we establish three experiments (Experiment I: DFs of COVID-19 cases; Experiment 2: point prediction of COVID-19 cases; Experiment 3: interval prediction of COVID-19 cases) to illustrate that the proposed hybrid system can effectively analyze the deterministic and uncertain information of COVID-19. Specifically, Experiment I used four probability DFs (Weibull, Rayleigh, Lognormal, and Gamma) to fit the distribution of epidemic cases. The parameters of the four probability DFs were optimized using the SCA algorithm. In experiment II, a hybrid model with environmental features, TN-SCA-LSSVM, was proposed for the point prediction of new cases and deaths from COVID-19. Three countries were selected as experimental cases and compared with the benchmark model to verify the prediction accuracy of the proposed model. To show the superior forecast performance of the hybrid model, five benchmark models, namely, ARIMA, back propagation neural network (BPNN), general regression neural network (GRNN), LSSVM, and SCA-LSSVM, were introduced. Experiment III calculated the interval prediction of new cases and new deaths in three countries based on the best distribution function determined in Experiment I and the point prediction results with the highest accuracy in Experiment II. Details are shown in the following sections.

Experiment I: Distribution Functions of COVID-19 Cases

To obtain the characteristics of the COVID-19 series and determine the optimal distribution function, four DFs (Weibull, Rayleigh, Lognormal, and Gamma), were used to calculate the distribution function of new COVID-19 cases and deaths. In addition, the parameter assessment of DFs was an essential step. Traditionally, the MLE method is used for parameter estimation of DFS. However, this study employed a robust optimization algorithm SCA to optimize the relevant parameters, and MLE was used as a comparison method to illustrate the optimization performance of SCA. **Table 3** shows the estimated parameters of the different DFs determined by the MLE and SCA methods. To further select the optimal DFs, the MAE, RMSE, and R2 were chosen as error criteria to determine the fitting level of these DFs. **Table 4** shows the values of the error results for the different distributions of new cases and new deaths of the epidemic in the three countries, and the bold values are the optimal results. Among the four DFs of all datasets, the R^2 determined by the SCA algorithm was significantly larger than that of the MLE method. At the same time, the SCA algorithm determined that the values of MAE and RMSE were also smaller than those of the MLE method. Thus, the SCA algorithm used in this paper had better optimization performance and simulated the distribution of the epidemic data exactly.

Furthermore, among the four DFs optimized by SCA, SCA-Lognormal only achieved optimal simulation capability for the new cases in the United States. SCA-Gamma achieved optimal simulation performance for both the new deaths in the United States and the new cases in India. SCA-Weibull obtained optimal simulation ability for new cases and new deaths in Brazil and India.

Experiment II: Intelligent Point Prediction for COVID-19 Cases

In this experiment, an intelligent hybrid prediction model coupled with environmental variables (TN-SCA-LSSVM) was used to perform a point prediction analysis of new cases and new deaths in three countries. The new cases and new deaths of COVID-19 and the environmental variables (temperature and NO_2) were taken as inputs of the multivariable point prediction. Thus, the number of input neurons of LSSVM was set to 4. To evaluate the predictive advantages of the proposed hybrid model, five univariate approaches, namely, ARIMA, BPNN, GRNN, LSSVM, and SCA-LSSVM, were selected as benchmark models for comparison. In addition, six evaluation criteria (MAE, RMSE, MAPE, IA, DA, and R2) were used to reflect the prediction

TABLE 3 | The parameters values of the different distribution functions are determined by MLE and SCA.

Countries	Types of cases	Methods	Lognormal		Gamma		Weibull Rayleigh		
			μ	σ	θ	k	λ	k	σ
United States	New cases	MLE	1.5859	4.8799	235.4008	0.9975	232.5667	1.1314	221.8384
		SCA	5.0481	0.8191	162.3509	1.3498	238.4217	1.0433	135.1476
	New deaths	MLE	1.4159	0.9973	2.0827	2.1522	4.5237	1.2630	3.6913
		SCA	1.2491	0.7112	2.1612	1.9676	4.5382	1.4115	3.0889
Brazil	New cases	MLE	1.2901	4.5723	117.6314	1.2773	160.8250	1.3241	126.2440
		SCA	4.8261	0.6909	70.6131	2.2244	176.4688	1.4984	122.9704
	New deaths	MLE	1.1251	0.9589	2.1763	1.6661	3.9725	1.5907	2.9499
		SCA	1.2175	0.5933	1.5846	2.4602	4.2921	1.7905	2.9930
India	New cases	MLE	1.6692	2.4204	27.6680	0.8291	22.3740	0.9368	21.4061
		SCA	2.7967	1.0588	27.4867	0.9043	24.9569	0.9138	15.6776
	New deaths	MLE	1.3955	−1.7107	0.3206	1.0024	0.3273	1.0529	0.2944
		SCA	−1.4439	1.1266	0.2868	1.0798	0.3507	0.9885	0.2440

TABLE 4 | The criteria values of different distribution functions of six datasets.

Countries	Types of cases	Criteria	Lognormal		Gamma		Weibull Rayleigh			
			MLE	SCA	MLE	SCA	MLE	SCA	MLE	SCA
United States	New cases	*MAE*	0.0839	**0.0235**	0.0441	0.0375	0.1634	0.0618	0.0420	0.0403
		RMSE	0.1023	**0.0351**	0.0540	0.0447	0.1930	0.0827	0.0505	0.0476
		R^2	0.8750	**0.9853**	0.9652	0.9761	0.5553	0.9184	0.9696	0.9730
	New deaths	*MAE*	0.0973	0.0165	0.0287	**0.0148**	0.0280	0.0214	0.0816	0.0472
		RMSE	0.1140	0.0217	0.0349	**0.0181**	0.0375	0.0245	0.0949	0.0533
		R^2	0.8455	0.9944	0.9855	**0.9961**	0.9833	0.9929	0.8930	0.9662
Brazil	New cases	*MAE*	0.0930	0.0526	0.0611	0.0327	0.0465	**0.0223**	0.0503	0.0491
		RMSE	0.1084	0.0587	0.0695	0.0392	0.0513	**0.0298**	0.0585	0.0572
		R^2	0.8591	0.9587	0.9421	0.9816	0.9684	**0.9893**	0.9590	0.9608
	New deaths	*MAE*	0.0917	0.0386	0.0611	0.0284	0.0424	**0.0237**	0.0317	0.0323
		RMSE	0.1069	0.0486	0.0686	0.0354	0.0464	**0.0296**	0.0375	0.0368
		R^2	0.8662	0.9724	0.9449	0.9853	0.9747	**0.9897**	0.9836	0.9842
India	New cases	*MAE*	0.0734	0.0385	0.0353	**0.0226**	0.0315	0.0232	0.1281	0.1213
		RMSE	0.0853	0.0474	0.0408	**0.0269**	0.0368	0.0279	0.1595	0.1322
		R^2	0.9131	0.9732	0.9801	**0.9913**	0.9838	0.9907	0.6962	0.7912
	New deaths	*MAE*	0.0565	0.0396	0.0380	0.0325	0.0322	**0.0231**	0.1140	0.1010
		RMSE	0.0673	0.0482	0.0441	0.0380	0.0367	**0.0291**	0.1366	0.1189
		R^2	0.9452	0.9719	0.9765	0.9825	0.9838	**0.9897**	0.7743	0.8290

The bold values present the optimal results.

performance of the models; the results are shown in **Tables 5**, **6**. The boldly marked values indicate the best values of the model in different evaluation metrics, and the optimal point prediction model is selected accordingly. **Figure 2** shows the predicted and observed values between the proposed model and other models. Further discussion of the experimental results follows.

From **Table 5**, we can draw the following conclusions:

For the single model comparisons, including ARIMA, BPNN, GRNN, LSSVM, it can be seen from **Table 5** and **Figure 3** that LSSVM had more accurate prediction accuracy than other single models and had the best performance among a variety of error indicators of MAE, RMSE, MAPE, IA, DA, and R2. For instance, the MAPE of new cases predicted by ARIMA, BPNN, GRNN, and LSSVM in the United States were 39.1424, 17.6103, 15.6918, and 14.1000%, respectively. In Brazil, the MAPE values of ARIMA, BPNN, GRNN, and LSSVM were 65.3496, 51.0333, 53.7500, and 51.2592%, respectively. In India, the MAPE values of ARIMA, BPNN, GRNN, and LSSVM were 36.4135, 18.8504, 17.5906, and 15.2222%, respectively.

The proposed hybrid model with environmental features showed stronger predictive performance compared with other models. For example, in the United States, compared with the LSSVM and SCA-LSSVM, TN-SCA-LSSVM led to 7.6160 and 4.3233% reductions in MAE, 3.5175 and 3.7255% reductions in RMSE, and 7.9957 and 6.2626% reductions in MAPE, respectively. In Brazil, compared with LSSVM and SCA-LSSVM,

TABLE 5 | The comparative forecasting error of different models for COVID-19 new cases.

Countries	Criteria	ARIMA	BPNN	GRNN	LSSVM	SCA-LSSVM	TN-SCA-LSSVM
United States	MAE	103.3365	72.3343	65.6524	60.5696	58.4851	55.9566
	RMSE	125.7786	103.3695	102.0531	88.3619	88.5528	85.2538
	MAPE (%)	39.1424	17.6103	15.6919	14.1000	13.8393	12.9726
	TIC	0.1337	0.1136	0.1131	0.0927	0.0927	0.0951
	IA	0.8571	0.9261	0.9250	0.9524	0.9524	0.9533
	R^2	0.6493	0.7631	0.7710	0.8269	0.8402	0.8262
Brazil	MAE	147.5017	80.0527	78.6863	84.3194	59.0297	58.6014
	RMSE	191.1310	101.7826	103.0591	105.4444	75.1114	73.9705
	MAPE (%)	65.3496	51.0333	53.7500	51.2592	28.9399	28.0350
	TIC	0.3616	0.2230	0.2266	0.2369	0.1521	0.1478
	IA	0.3509	0.6131	0.6225	0.5957	0.7050	0.7084
	R^2	−4.2445	−0.4873	−0.5248	−0.5962	0.1901	0.2145
India	MAE	3.6180	2.8170	2.3159	1.9432	1.8260	1.7828
	RMSE	4.3580	5.4566	3.3962	3.1049	3.2072	3.1651
	MAPE (%)	36.4135	18.8504	17.5906	15.2222	14.5030	14.3134
	TIC	0.1697	0.2183	0.1406	0.1320	0.1370	0.1358
	IA	0.5170	0.5111	0.6722	0.7120	0.7283	0.7308
	R^2	−0.4228	−1.2524	0.1275	0.2778	0.2219	0.2422

TABLE 6 | The comparative forecasting error of different models for COVID-19 new death cases.

Countries	Criteria	ARIMA	BPNN	GRNN	LSSVM	SCA-LSSVM	TN -SCA-LSSVM
United States	MAE	3.9138	1.7581	1.8688	1.7476	1.7006	1.6252
	RMSE	4.8346	2.1134	2.2823	2.2780	2.1507	2.0040
	MAPE (%)	55.9991	25.2980	27.9532	26.1569	25.3390	24.3988
	TIC	0.2922	0.1289	0.1434	0.1387	0.1321	0.1262
	IA	0.5075	0.8354	0.7983	0.8166	0.8376	0.8470
	R^2	−1.4148	0.5386	0.4618	0.4639	0.5221	0.5821
Brazil	MAE	2.0207	1.5995	1.5551	1.4997	1.2921	1.1995
	RMSE	2.4866	2.3221	2.3352	2.2038	1.7577	1.4880
	MAPE (%)	47.3042	48.2826	46.8319	44.1043	33.4255	26.4318
	TIC	0.2277	0.2232	0.2301	0.2185	0.1694	0.1381
	IA	0.3561	0.5578	0.5851	0.6269	0.7256	0.7876
	R^2	−0.5565	−0.3574	−0.3727	−0.2225	0.2223	0.4427
India	MAE	0.1176	0.0591	0.0330	0.0261	0.0227	0.0251
	RMSE	0.1235	0.1428	0.0471	0.0396	0.0341	0.0400
	MAPE (%)	144.3315	43.5405	23.6569	18.5391	17.6502	17.7402
	TIC	0.3678	0.4568	0.1844	0.1582	0.1395	0.1583
	IA	0.3677	0.4284	0.8035	0.8538	0.8727	0.8557
	R^2	−4.3463	−6.5096	0.1844	0.4236	0.5931	0.4409

TN -SCA-LSSVM led to 30.5007 and 0.7256% reductions in MAE, 29.8488 and 1.5190% reductions in RMSE, and 45.3074 and 3.1267% reductions in MAPE, respectively. In India, compared with LSSVM and SCA-LSSVM, TN-SCA-LSSVM led to 21.1537 and 6.0300% reductions in MAE, 5.5636 and −3.2965% reductions in RMSE, and 17.5524 and 4.7246% reductions in MAPE, respectively. According to the six evaluation criteria, it can be concluded that the proposed hybrid multivariable model was significantly better than other benchmark models for forecasting new cases.

From **Table 6**, we can draw the following conclusions:

It can be seen from **Table 6** and **Figure 3** that the proposed TN-SCA-LSSVM showed stronger predictive performance than ARIMA, BPNN, GRNN, LSSVM, and SCA-LSSVM. LSSVM had more accurate prediction accuracy than other single models and had the best performance among various error indicators of MAE, RMSE, MAPE, IA, DA, and R2. The proposed TN-SCA-LSSVM showed stronger predictive performance than other single or hybrid univariate models. According to the six evaluation criteria, it can be concluded that the proposed hybrid multivariable model was significantly better than other benchmark models for forecasting new death cases.

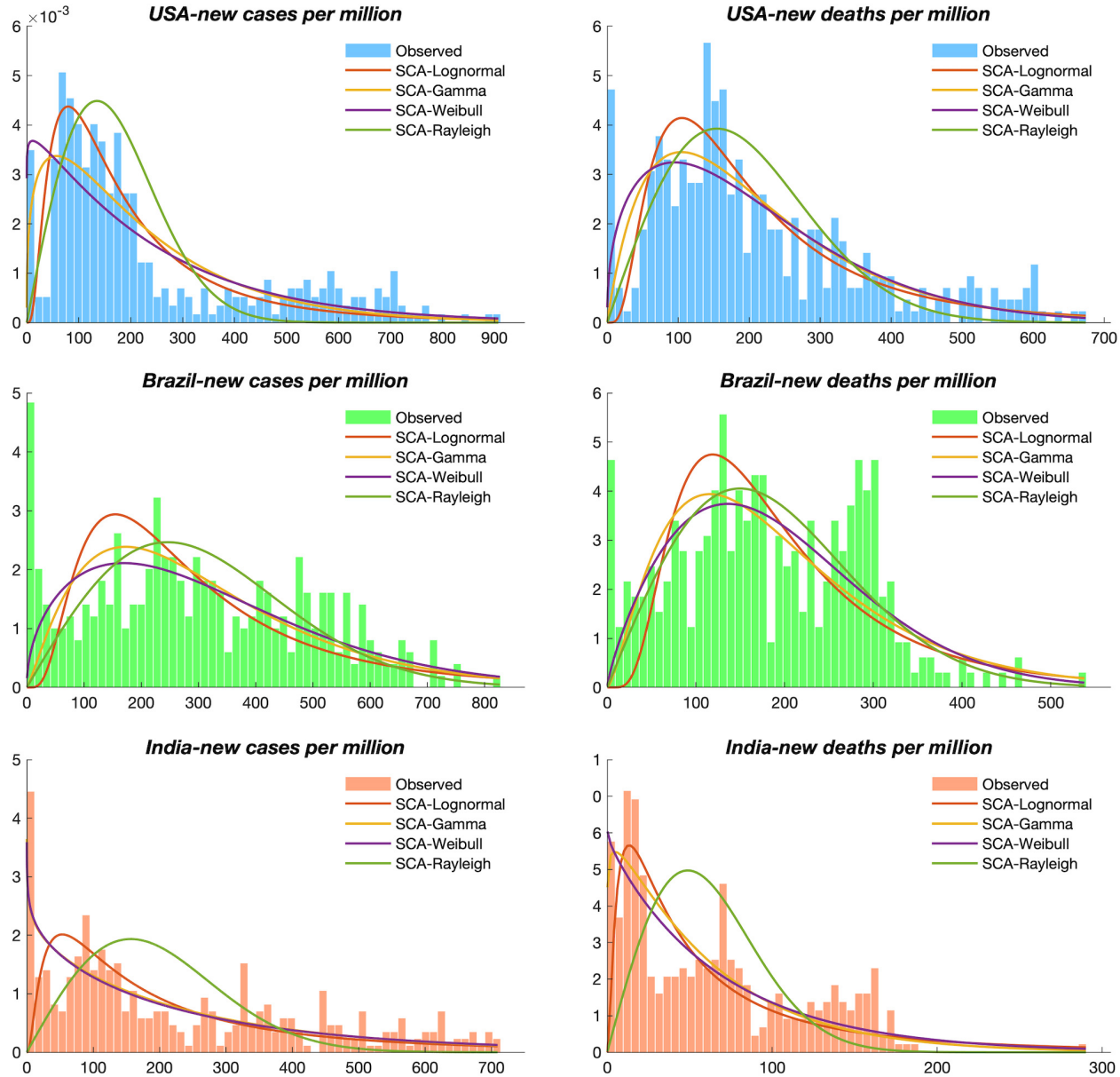


FIGURE 2 | The observed sequences and probability density functions of four distributions in the United States, Brazil, and India.

Remark

The proposed hybrid multivariable model with environmental features had strong prediction ability and effectively addressed the complexity and non-linearity of new cases and new deaths. The optimization method played an essential role in improving the prediction accuracy of the hybrid model. Results indicated that the SCA significantly improved the prediction performance of the LSSVM. In addition, the forecasting model with environmental variables further improved the prediction ability of the hybrid model.

Experiment III: Interval Forecasting of COVID-19 Cases

In Experiment III, based on the interval forecasting theory discussed in Section "Interval Forecasting Module," the interval prediction of new cases and new deaths in three countries was calculated by incorporating the optimal distribution function determined in Section "Experiment I: Distribution Functions of COVID-19 Cases" and the point prediction results with the highest accuracy in Section "Experiment II: Intelligent Point Prediction for COVID-19 Cases." In addition, two metrics, PIAW and PICP listed in **Table 1**, were used to measure the validity

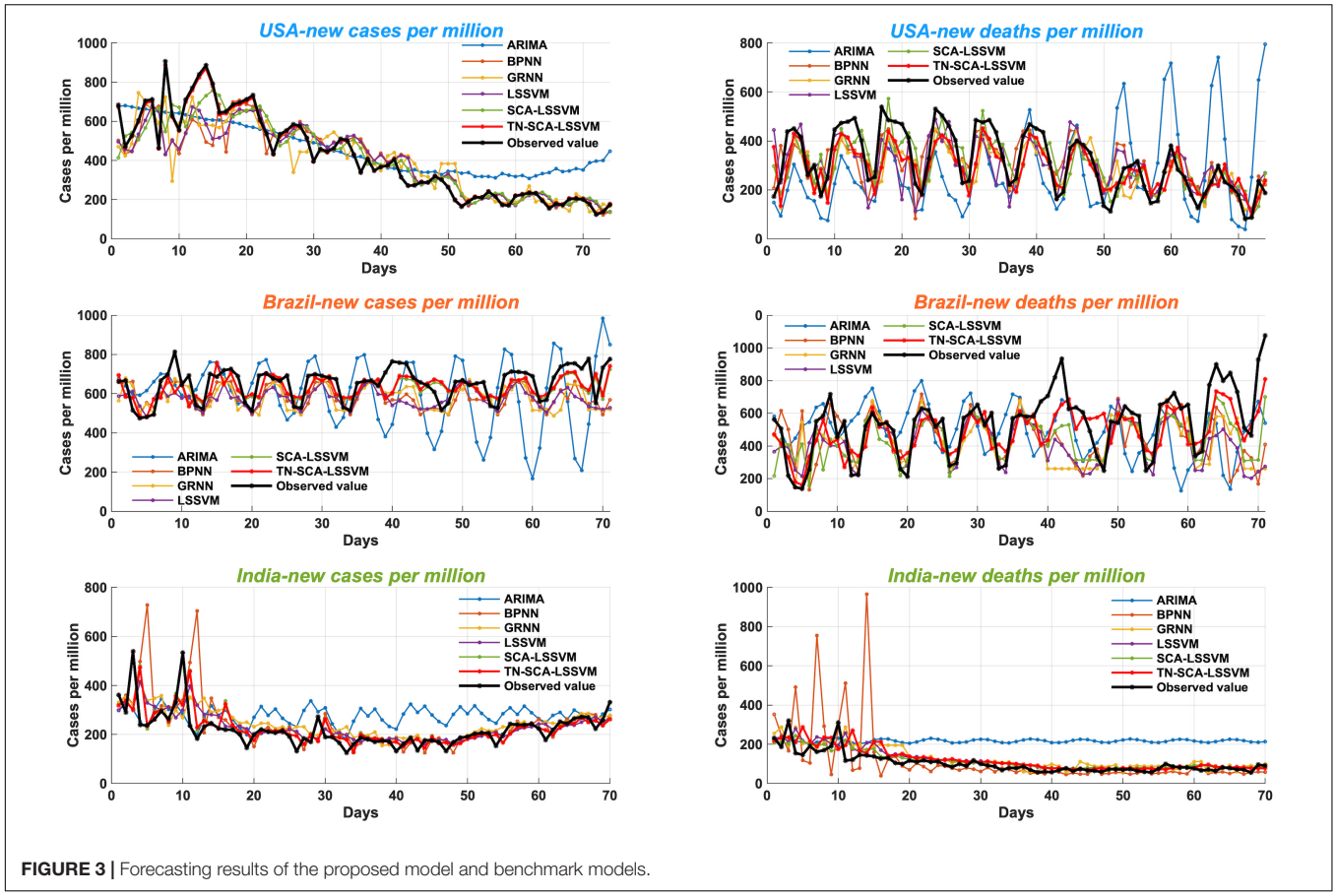


FIGURE 3 | Forecasting results of the proposed model and benchmark models.

TABLE 7 | The interval prediction results under five different significance levels of COVID-19 cases.

Countries	Types of cases	Criteria	α					
			0.2	0.25	0.3	0.35	0.4	0.45
United States	New cases	IFCP	100.00%	100.00%	100.00%	100.00%	98.65%	87.84%
		IFAW	627.5977	489.2653	372.9357	270.2132	176.0058	86.8294
	New death cases	IFCP	100.00%	100.00%	98.65%	93.24%	90.54%	66.22%
		IFAW	9.9036	7.7723	5.9529	4.3280	2.8256	1.3958
Brazil	New cases	IFCP	100.00%	98.59%	98.59%	92.96%	81.69%	59.15%
		IFAW	300.3756	236.0082	180.9128	131.6092	85.9563	42.4711
	New death cases	IFCP	100.00%	100.00%	98.59%	97.18%	87.32%	59.15%
		IFAW	5.5115	4.3527	3.3489	2.4426	1.5981	0.7904
India	New cases	IFCP	100.00%	100.00%	100.00%	100.00%	100.00%	85.71%
		IFAW	19.1788	14.8640	11.2819	8.1496	5.2975	2.6103
	New death cases	IFCP	100.00%	100.00%	100.00%	100.00%	100.00%	84.29%
		IFAW	0.2010	0.1548	0.1170	0.0842	0.0546	0.0269

of the interval prediction. It should be noted that the optimal interval prediction should satisfy the following conditions: The larger the IFCP value ($0 \leq \text{IFCP} \leq 100\%$) and the smaller the IFAW value at a given significance level α are, the better the predictive performance of the interval prediction. **Table 7** shows the United States, India, and Brazil interval prediction results under five different significance levels (0.20, 0.25, 0.30, 0.35, and 0.40). From **Table 7**, it can be observed that the values of IFCP and IFAW were different at five significance levels. For example, when α was 0.3, the IFCP and IFAW of COVID-19 new cases in the United States were 100.00% and 372.9357; when α was 0.35, the IFCP and IFAW of COVID-19 new cases in the United States were 100.00% and 270.2132, respectively.

To present the interval prediction results more visually, the interval prediction results of COVID-19 cases at four significance levels of 0.25, 0.3, 0.35, and 0.4 were selected to make a visual

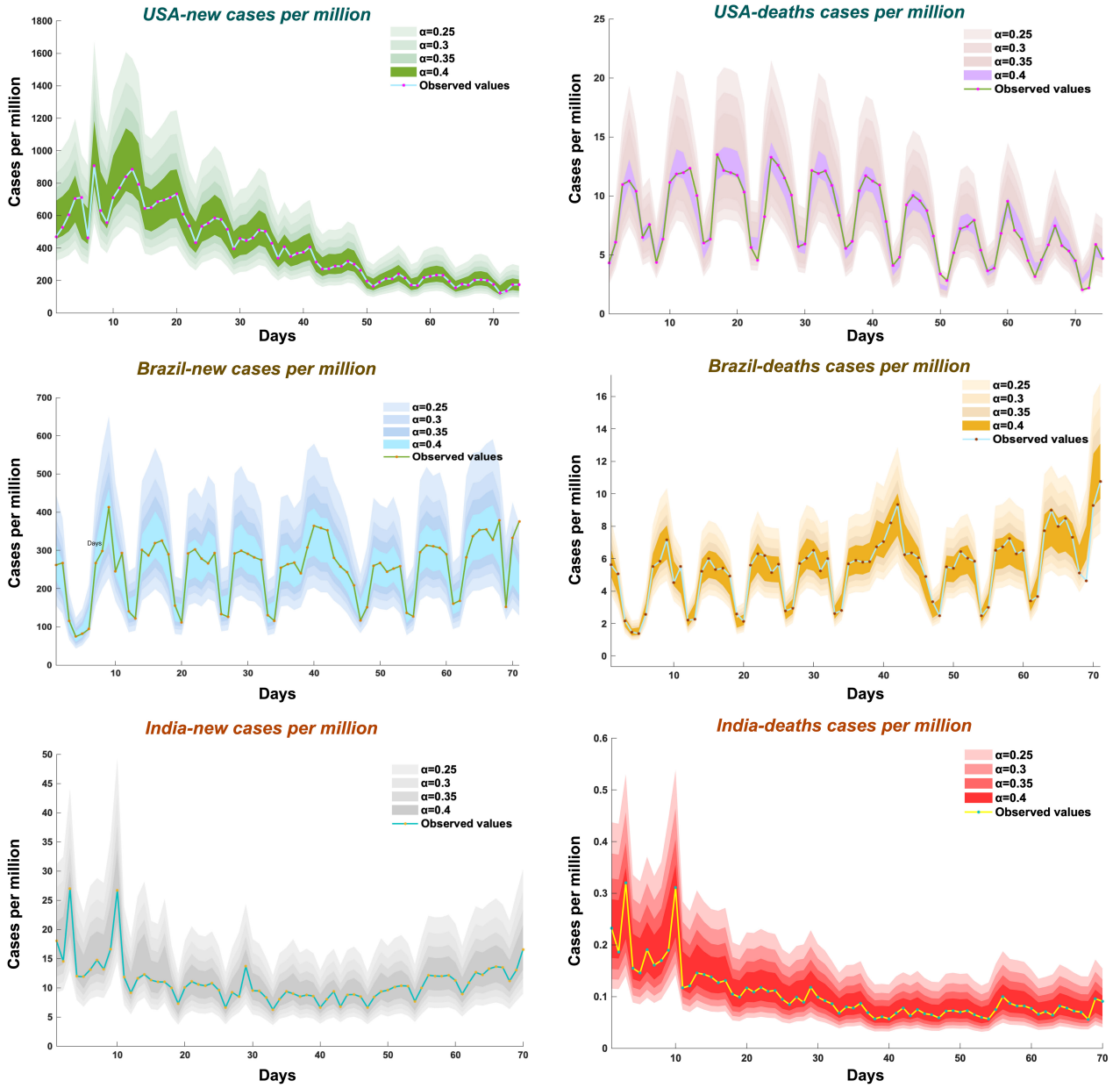


FIGURE 4 | Interval prediction results of the proposed model with different significance levels.

effect, as shown in **Figure 4**. **Figure 4** contains six subplots showing the interval prediction results of new cases and new deaths for each of the three countries. The dots represent the actual value, and the color depth of the shaded area indicates the range of interval forecasting at different significance levels. When a smaller significance level is selected, there are individual actual values that exceed the corresponding shaded areas. When a smaller significance level is chosen, there are individual actual values that exceed the corresponding shaded areas. When the significance level is large, although the shaded area can cover all the actual values well, it will lead to a large range of prediction intervals and lose practical significance.

Discussion

The proposed point and interval forecasting approach with environmental variables obtained better prediction results than other comparable models. The specific reasons were determined to be as follows: First, the optimal DFs and their parameters that best fit the epidemic data of different countries were obtained by SCA. Second, the proposed hybrid multivariable model SCA-LSSCM had a strong prediction ability and effectively addressed the complexity and non-linearity of new cases and new deaths. Third, the addition of environmental variables further improved the prediction ability of the hybrid model. Finally, interval forecasting was calculated based on the optimal DFs and

point prediction results to capture uncertainty information for decision-making.

Notably, because the interval prediction results were calculated based on the point prediction results, the interval prediction performance depends mainly on the point prediction results. In addition, a suitable significance level needs to be selected according to the actual situation in the practical application. In conclusion, the interval forecasting model proposed in this study could provide uncertain information about future epidemic development and could be combined with the accurate deterministic information provided by the point prediction hybrid model in Experiment 2. It could provide public health decision-makers with rich information for epidemic prevention and control decisions.

In practice, the proposed model could be driven by real-time data to dynamically and continuously optimize the model parameters by updating the data daily, making the model adaptable to complex epidemic scenarios that are non-linear, dynamic, and ambiguous. At the same time, this data-driven prediction would also help to establish a predictable safeguard mechanism, leaving a window of time for relevant decision-making departments to take measures and adjust strategies in advance to avoid the continuous spread of the epidemic.

CONCLUSION

This study presented a novel point and interval forecasting approach with environmental variables, which was composed of a distribution function analysis module, an intelligent point prediction module, and an interval forecasting module. In the distribution function analysis module, according to the results of the MAE, RMSE, and R2, SCA-Lognormal achieved optimal simulation capability for the new cases in the United States, while SCA-Gamma achieved optimal simulation performance in both the new deaths in the United States and the new cases in India. SCA-Weibull obtained optimal simulation ability for new cases and new deaths in Brazil and new deaths in India. In the intelligent point prediction module, according to the MAE, RMSE, MAPE, IA, DA, and R2, the hybrid multivariate model TN-SCA-LSSVM achieved more robust predictive performance than other univariate approaches, such as ARIMA, BPNN, GRNN, LSSVM, and SCA-LSSVM, which indicated that SCA significantly improved the prediction performance of LSSVM and that the addition of environmental features (temperature and NO_2) further improved the prediction ability of the hybrid model. For instance, the average MAPE values of the proposed TN-SCA-LSSVM model were 62.1521, 33.9225, 27.5146, 18.3956, and 5.8034% lower than those of ARIMA, BPNN, GRNN, LSSVM, and SCA-LSSVM, respectively. In the interval forecasting module, for interval prediction of Covid-19 data in three countries, interval prediction results for new cases and new deaths were obtained based on the point prediction values and optimal DFs of the proposed hybrid TN-SCA-LSSVM model. The results showed that the performance of interval prediction was excellent because most of the observed values were located in the shaded area, with higher values of IFCP and smaller values of IFAW at different significance levels. Overall, the proposed system achieved better prediction results than other comparable models and enabled the informative and practical quantification of future COVID-19 pandemic trends, which offers more constructive suggestions for governmental administrators and the general public.

In this study, epidemiological data and two environmental variables were considered inputs for point and interval prediction models. However, predicting COVID-19 is a complex problem related to multiple factors, such as meteorological, environmental, socioeconomic or policy factors. Thus, the forecasting model can be improved by incorporating more influencing factors from different data sources, which may be an interesting research pursuit.

DATA AVAILABILITY STATEMENT

The original datasets used in the study are included in the article, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZQ: writing, conceptualization, and methodology. YS: writing-reviewing and editing. QX: formal analysis. YL: data curation and visualization. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 72004086).

REFERENCES

Ala'raj, M., Majdalawieh, M., and Nizamuddin, N. (2021). Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections. *Infect. Dis. Model.* 6, 98–111. doi: 10.1016/J.IDM.2020.11.007

Barmparis, G. D., and Tsironis, G. P. (2020). Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach. *Chaos Solitons Fract.* 135:109842. doi: 10.1016/j.chaos.2020.109842

Bauwens, M., Compernolle, S., Stavrakou, T., Müller, J. F., van Gent, J., Eskes, H., et al. (2020). Impact of Coronavirus Outbreak on NO2 Pollution Assessed Using TROPOMI and OMI Observations. *Geophys. Res. Lett.* [Epub Online ahead of print] doi: 10.1029/2020GL087978

Borghi, P. H., Zakordonets, O., and Teixeira, J. P. (2021). A COVID-19 time series forecasting model based on MLP ANN. *Procedia Comput. Sci.* 181, 940–947. doi: 10.1016/j.procs.2021.01.250

Castillo, O., and Melin, P. (2020). Forecasting of COVID-19 time series for countries in the world based on a hybrid approach combining the fractal dimension and fuzzy logic. *Chaos Solitons Fract.* 140:110242. doi: 10.1016/j.chaos.2020.110242

Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. *Sci. Total Environ.* 729:138817. doi: 10.1016/j.scitotenv.2020.138817

Covid-19 Worldwide Air Quality data. *COVID-19 Worldwide Air Quality data*. Available Online at: https://aqicn.org/data-platform/covid19/ (accessed Feb. 13, 2022).

Ghosal, S., Sengupta, S., Majumder, M., and Sinha, B. (2020). Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks. *Diabetes Metab. Syndr.* 14, 311–315. doi: 10.1016/j.dsx.2020.03.017

Hao, Y., Niu, X., and Wang, J. (2021). Impacts of haze pollution on China's tourism industry: a system of economic loss analysis. *J. Environ. Manag.* 295:113051. doi: 10.1016/J.JENVMAN.2021.113051

He, S., Peng, Y., and Sun, K. (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dyn.* 101, 1667–1680. doi: 10.1007/s11071-020-05743-y

Keerthi, S. S., and Lin, C. J. (2003). Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Comput.* 15, 1667–1689. doi: 10.1162/089976603321891855

Konstantinoudis, G., Padellini, T., Bennett, J., Davies, B., Ezzati, M., and Blangiardo, M. (2021). Long-term exposure to air-pollution and COVID-19 mortality in England: a hierarchical spatial analysis. *Environ. Int.* 146:106316. doi: 10.1016/J.ENVINT.2020.106316

Li, M. Y., Graef, J. R., Wang, L., and Karsai, J. (1999). Global dynamics of a SEIR model with varying total population size. *Math. Biosci.* 160, 191–213. doi: 10.1016/S0025-5564(99)00030-9

Lian, X., Huang, J., Zhang, L., Liu, C., Liu, X., and Wang, L. (2021). Environmental Indicator for COVID-19 Non-Pharmaceutical Interventions. *Geophys. Res. Lett.* 48:e2020GL090344. doi: 10.1029/2020GL090344

Ly, K. T. (2020). A COVID-19 forecasting system using adaptive neuro-fuzzy inference. *Finance Res. Lett.* 41:101844. doi: 10.1016/j.frl.2020.101844

Malki, Z., Atlam, E. S., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., and Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: machine learning approaches. *Chaos Solitons Fract.* 138:110137. doi: 10.1016/J.CHAOS.2020.110137

McAdam, P., and McNelis, P. (2005). Forecasting inflation with thick models and neural networks. *Econ. Model.* 22, 848–867. doi: 10.1016/J.ECONMOD.2005.06.002

Mirjalili, S. (2016). SCA: a Sine Cosine Algorithm for solving optimization problems. *Knowl.-Based Syst.* 96, 120–133. doi: 10.1016/j.knosys.2015.12.022

Moftakhar, L., and Seif, M. (2020). The exponentially increasing rate of patients infected with COVID-19 in Iran. *Arch. Iran. Med.* 23, 235–238. doi: 10.34172/aim.2020.03

Ndaïrou, F., Area, I., Nieto, J. J., and Torres, D. F. M. (2020). Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. *Chaos Solitons Fract.* 135:109846. doi: 10.1016/j.chaos.2020.109846

Pandey, G., Chaudhary, P., Gupta, R., and Pal, S. (2020). SEIR and Regression Model based COVID-19 outbreak predictions in India. *medRxiv* [Preprint]. doi: 10.1101/2020.04.01.20049825

Parbat, D., and Chakraborty, M. (2020). A python based support vector regression model for prediction of COVID19 cases in India. *Chaos Solitons Fract.* 138:109942. doi: 10.1016/j.chaos.2020.109942

Shi, P., Dong, Y., Yan, H., Zhao, C., Li, X., Liu, W., et al. (2020). Impact of temperature on the dynamics of the COVID-19 outbreak in China. *Sci. Total Environ.* 728:138890. doi: 10.1016/J.SCITOTENV.2020.138890

Song, Y., Qin, S., Qu, J., and Liu, F. (2015). The forecasting research of early warning systems for atmospheric pollutants: a case in Yangtze River Delta region. *Atmos. Environ.* 118, 58–69. doi: 10.1016/j.atmosenv.2015.06.032

Stock, J. H., and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *J. Forecast.* 23, 405–430. doi: 10.1002/for.928

Suykens, J. A. K., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Process. Lett.* 9, 293–300. doi: 10.1023/A:1018628609742

Swapnarekha, H., Behera, H. S., Nayak, J., and Naik, B. (2020). Role of intelligent computing in COVID-19 prognosis: a state-of-the-art review. *Chaos Solitons Fract.* 138:109947. doi: 10.1016/j.chaos.2020.109947

Tian, C., and Hao, Y. (2020). Point and interval forecasting for carbon price based on an improved analysis-forecast system. *Appl. Math. Model.* 79, 126–144. doi: 10.1016/j.apm.2019.10.022

Travaglio, M., Yu, Y., Popovic, R., Selley, L., Leal, N. S., and Martins, L. M. (2021). Links between air pollution and COVID-19 in England. *Environ. Pollut.* 268:115859. doi: 10.1016/J.ENVPOL.2020.115859

Wang, M., Jiang, A., Gong, L., Luo, L., Guo, W., Li, C., et al. (2020). Temperature Significantly Change COVID-19 Transmission in 429 cities. *medRxiv* [Preprint]. doi: 10.1101/2020.02.22.20025791

Wu, T., Perrings, C., Kinzig, A., Collins, J. P., Minteer, B. A., and Daszak, P. (2017). Economic growth, urbanization, globalization, and the risks of emerging infectious diseases in China: a review. *Ambio* 46, 18–29. doi: 10.1007/s13280-016-0809-2

Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., et al. (2020). Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. *Sci. Total Environ.* 729:139051. doi: 10.1016/J.SCITOTENV.2020.139051

Xu, Y., Du, P., and Wang, J. (2017). Research and application of a hybrid model based on dynamic fuzzy synthetic evaluation for establishing air quality forecasting and early warning system: a case study in China. *Environ. Pollut.* 223, 435–448. doi: 10.1016/j.envpol.2017.01.043

Yang, W., Sun, S., Hao, Y., and Wang, S. (2022). A novel machine learning-based electricity price forecasting model based on optimal model selection strategy. *Energy* 238:121989. doi: 10.1016/J.ENERGY.2021.121989

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.

Copyright © 2022 Qu, Sha, Xu and Li. 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 Ecology and Evolution

ORIGINAL RESEARCH
published: 10 May 2022
doi: 10.3389/fevo.2022.885171

Tourism Demand Forecasting Considering Environmental Factors: A Case Study for Chengdu Research Base of Giant Panda Breeding

Jianhong He[1], *Dong Liu*[1]*, *Yulei Guo*[2]* *and Daohua Zhou*[3]

[1] School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing, China, [2] Chengdu Research Base of Giant Panda Breeding, Chengdu, China, [3] Chengdu Zhongke Daqi Software Co., Ltd, Chengdu, China

Effectively prediction of the tourism demand is of great significance to rationally allocate resources, improve service quality, and maintain the sustainable development of scenic spots. Since tourism demand is affected by the factors of climate, holidays, and weekdays, it is a challenge to design an accurate forecasting model obtaining complex features in tourism demand data. To overcome these problems, we specially consider the influence of environmental factors and devise a forecasting model based on ensemble learning. The model first generates several sub-models, and each sub-model learns the features of time series by selecting informative sequences for reconstructing the forecasting input. A novel technique is devised to aggregate the outputs of these sub-models to make the forecasting more robust to the non-linear and seasonal features. Tourism demand data of Chengdu Research Base of Giant Panda Breeding in recent 5 years is used as a case to validate the effectiveness of our scheme. Experimental results show that the proposed scheme can accurately forecasting tourism demand, which can help Chengdu Research Base of Giant Panda Breeding to improve the quality of tourism management and achieve sustainable development. Therefore, the proposed scheme has good potential to be applied to accurately forecast time series with non-linear and seasonal features.

Keywords: tourism demand forecasting, ensemble learning, RNN, time series, machine learning

OPEN ACCESS

Edited by:
Wendong Yang,
Shandong University of Finance and Economics, China

Reviewed by:
Leo Yu Zhang,
Deakin University, Australia
Shenghui Zhang,
University of Macau, China

***Correspondence:**
Dong Liu
dongliu525@gmail.com
Yulei Guo
yulei.guo@panda.org.cn

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Ecology and Evolution

Received: *27 February 2022*
Accepted: *11 April 2022*
Published: *10 May 2022*

Citation:
He J, Liu D, Guo Y and Zhou D (2022) Tourism Demand Forecasting Considering Environmental Factors: A Case Study for Chengdu Research Base of Giant Panda Breeding. Front. Ecol. Evol. 10:885171. doi: 10.3389/fevo.2022.885171

1. INTRODUCTION

Tourism, as a multibillion-dollar business, has a certain impact on the ecological environment,and also servers as an engine of economic growth. For the economic growth in the past decade, the share of tourism industry has increased steadily (Claveria et al., 2015; Yu, 2021), which has caused rising concerns about the efficiency and effectiveness of the allocation of tourism resources. Tourism forecasting plays a vital role in the allocation of tourism resources, and the accurate forecasting results can not only help decision makers make the reasonable allocation, but also support tourists to plan their schedules.

Many researchers devote to the tourism demand forecasting, and the existing studies have proved that tourism demand forecasting is crucial to allocate the tourism resources (Li and Cao, 2018; Khademi et al., 2022). For natural resource scenic spots, accurate forecasting of tourism

demand can also help scenic spots control the impact of visitors on the environment, achieve a good balance between economic income and environmental protection, and thus promote the sustainable development of scenic spots. Tourism activity is affected by many factors (Hao et al., 2021), such as weather, environment, emergencies such as pandemic and government policy, which makes the accurate tourism forecasting as a challenge thing. Therefore, it is necessary to consider all of these factors in order to improve the forecasting accuracy.

The above challenges have promoted the development of tourism demand forecasting algorithm. Most of these schemes can be classified into two categories: statistical based models and machine learning based models. Statistical based models aim to make the unbiased prediction of the future demand, and they hold strong assumption on the time series data. One of the most representatives of these models are based on the Autoregressive Integrated Moving Average (ARIMA) technique (Faruk, 2010; Wang, 2022). ARIMA based models predicts a future value with several past observations and random errors with a linear function. They commonly used stationary time series fitting prediction model and can only capture linear, but not non-linear relations. Liu et al. (2014) proposes gray forecasting model. This method accumulating the original series and create new series to weaken the inherent randomness of the original data. Thus, it can reveal the regularity of the orderly data sequence and in turn make predictions for future values. However, gray forecasting model also impose heavy restrictions on time series. Due to the strict assumption on data, both ARIMA or gray forecasting based models need to pre-process the forecasting input. With constant assumptions, the models can solve the closed form of the solution. Nevertheless, these models are still far from practical due to the inherent complexity of the real tourism data.

Machine learning based models have been widely applied in tourism demand forecasting. In contrast to statistical learning based models, machine learning based models have the potential to recognize the nonlinear, seasonal and other complex features in the tourism time series by imposing no restriction on raw data. Singh et al. (2021) apply Support Vector Machine (SVM) to forecast the forest fire. SVM finds a hyperplane in the n-dimensional space that can classify the data points with non-linear relations.

Due to the effectiveness of capturing non-linear relations, quite a few works introduce neural networks to address the time series forecasting problems (Qian et al., 2022; Wang et al., 2022). Among them, some works based on the Convolutional Neural Network (CNN) to build the forecasting models (LeCun and Bengio, 1995; Shin et al., 2016). As an effective deep learning models, CNN can effectively extract robust features from the time series data. To learn the long-term and short term dependencies in time series data, many works propose to apply LSTM method for forecasting task (Ji et al., 2019; Khademi et al., 2022; Ozkok and Celik, 2022). The performance of these models highly depends on the feature engineering. How to reconstruct the forecasting input to identify and combine the critical information in data features is utmost significant. And, several complex models (e.g., deep learning based models) with large number of parameters not only require laborious computation but hinder the efficiency in model training and predicting. Moreover, these models suffer from the overfitting problem due to the complicated characteristics of the tourism data.

Resolving the aforementioned problems paves the way to practical tourism forecasting models. For this aim, this article presents a novel model based on ensemble learning, which considers the environmental factors. Our method is built upon the following key observations. First, the forecasting model is comprised of two components to learn the sequential relation and complex interactions of features in tourism data. Second, time series from different category are combined to expand the feature space, which is beneficial to augment and smooth the series data. Third, different feature are sampled to build several forecasting sub-models, which ensure the diversity of the sub-models and the ability to capture informative sequences in tourism time series. More robust prediction is produced by aggregating the outputs from the sub-models. This alleviate the overfitting problem of the single model based scheme, thus improve the accuracy of the forecasting model. Overall, the contributions of this article are three-folded.

1. We propose a forecasting framework that can both learn dependencies in tourism time series and extract high and low-order correlation in features of the target time. This design effectively addresses the problem of the non-linear, seasonal features in tourism time series data.
2. Considering the impact of environmental factors and the speciality of the tourism data, we propose a more robust model that is a marriage between combination technique and ensemble learning. To the best of our knowledge, we are the first to incorporate the combination technique and the idea of ensemble learning in tourism demand forecasting problem. These two techniques can alleviate the overfitting problem of previous work on tourism time series data.
3. Chengdu Research Base of Giant Panda Breeding, a famous education tourist attraction in China, is used as a case to validate the effectiveness of our scheme. The experimental results show that the proposed scheme provides accurate estimates on the daily tourism demands. Thus, the proposed scheme is conducive to improving the sustainable development of the scenic spot.

The remainder of this article is organized as follows. In Section 2, we describe the design of our model. The experimental results are reported and analyzed in Section 3. In Section 4, the conclusion of this article is drawn.

2. THE PROPOSED METHODOLOGY

For effectively improving the generalization ability and preventing the overfitting problem. The proposed model is comprised of two components: the sequential learning component is responsible for exploring relations in time series; the feature extracting component is designed to explore the high-order and low-order information in features of the target time t. The outputs of the two components are aggregated to

form the final predictive value. The overview of the forecasting model is shown in **Figure 1**.

2.1. Sequential Learning Component

The recurrent neural network (RNN) based models are more suitable for time series fitting tasks with time dependencies than the normal artificial neural network (ANN). Due to the weather and seasonal features of the tourism data, (a) we design an RNN model that has the ability for sequential learning to explore the short-term dependencies in tourism time series. The main function of RNN based models is to interact the current information with the historical state. In our network, the state of the hidden layers is updated as follows:

$$\begin{aligned} \boldsymbol{S}_{t-1} &= u(\boldsymbol{H}_{t-1}) \\ \boldsymbol{M}_t &= v(\boldsymbol{X}_t) \\ \boldsymbol{H}_t &= p(\boldsymbol{M}_t + \boldsymbol{S}_{t-1}) \end{aligned} \tag{1}$$

where $\boldsymbol{X}_t$ denotes the features at time t, $\boldsymbol{H}_t$ denotes the state of the hidden layer at time t, $u(\cdot)$ transform the state of hidden layer at $t-1$, $v(\cdot)$ extracts the information within $\boldsymbol{X}_t$ and $p(\cdot)$ joints the relevant contextual at time $t-1$ and t. Thus, the current state of the hidden layer is updated and transferred. In our network, the functions $u(\cdot)$, $v(\cdot)$, and $p(\cdot)$ are defined as follows:

$$\begin{aligned} u\left(\boldsymbol{H}_{t-1}\right) &= relu\left(\boldsymbol{V}_2^T relu\left(\boldsymbol{V}_1^T \boldsymbol{H}_{t-1} + \boldsymbol{b}_1\right) + \boldsymbol{b}_2\right) \\ v\left(\boldsymbol{X}_t\right) &= \sigma\left(\boldsymbol{U}_2^T\left(relu\left(\boldsymbol{U}_1^T \boldsymbol{X}_t + \boldsymbol{b}_3\right) + \boldsymbol{b}_4\right)\right) \\ p\left(\boldsymbol{M}_t + \boldsymbol{S}_{t-1}\right) &= \sigma\left(\boldsymbol{M}_t + \boldsymbol{S}_{t-1}\right) \end{aligned} \tag{2}$$

where $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{b}$ are parameters that are trained at each step t. The symbol $relu(\cdot), \sigma(\cdot)$ are denoted as relu and sigmoid function, respectively.

2.1.1. Feature Extracting Component

Different correlations of features in the tourism data contains different information. In order to better extract the informative correlations in features, we design two modules that can be executed in parallel in feature extracting component, which are utilized to extract high-order and low-order information in tourism data.

For extracting the low-order information, we design a single-layer fully connected neural network to explore the linear relation in tourism data. It transforms features by combining features in different dimensions linearly. The module is defined as:

$$r_l = \boldsymbol{w}_l^{\mathrm{T}} \boldsymbol{X}_t + b_l, \tag{3}$$

where $\boldsymbol{X}_t$ denotes the features of the target time t, r_l is the linear combination of features. $\boldsymbol{w}_l \in \mathbb{R}^n$ is the parameters serve as the coefficients for linear weighting.

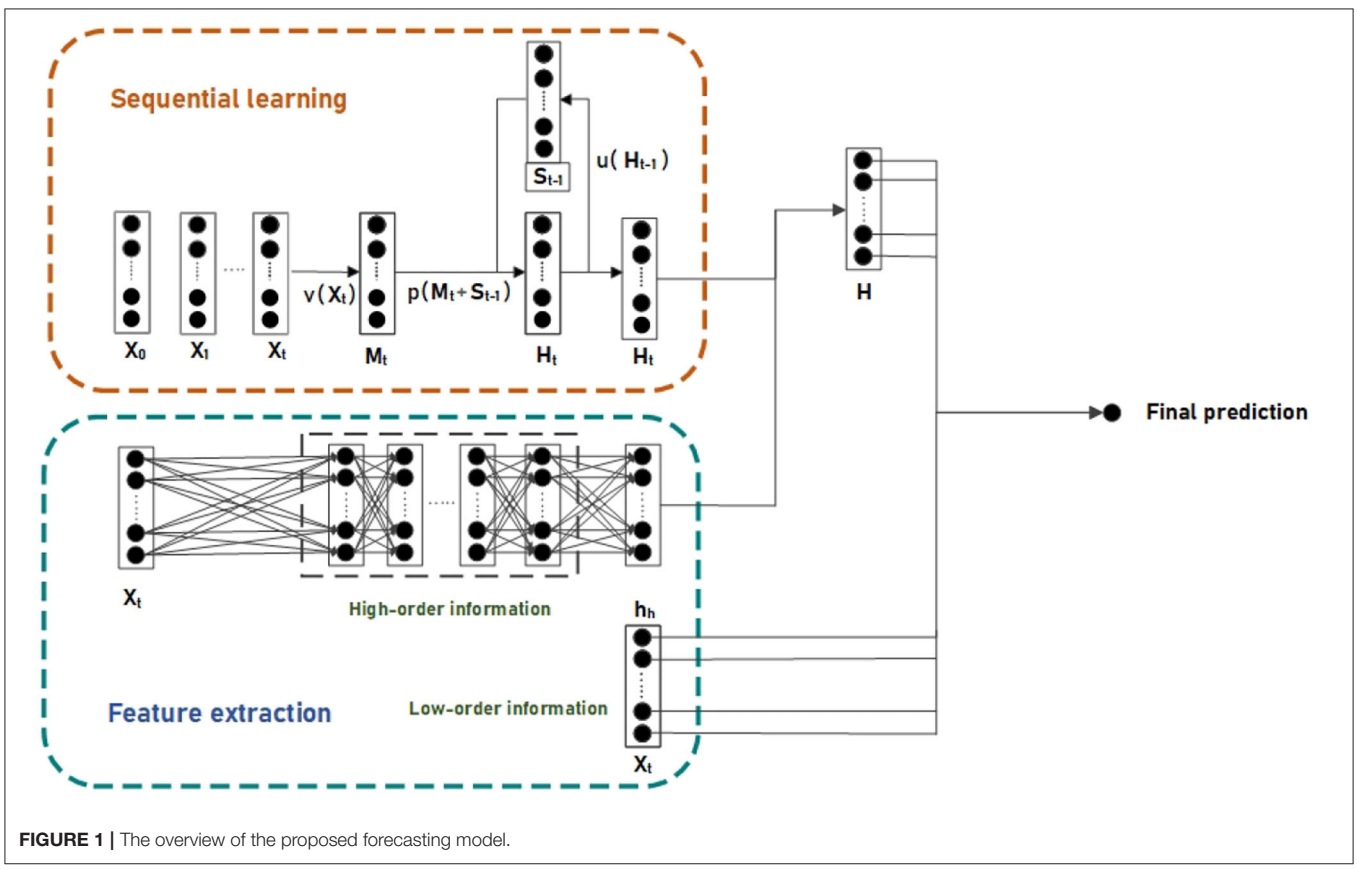


FIGURE 1 | The overview of the proposed forecasting model.

We devise another model for extracting high-order correlations in features. This module leverages the powerful non-linear expressing capability of multi-layer perceptrons to extract complex but valuable correlations in features. This module can be formally defined as follows:

$$\begin{aligned} \boldsymbol{h}_1 &= \gamma_0\left(\boldsymbol{W}_0^{\mathrm{T}}\boldsymbol{X}_t + \boldsymbol{b}_0\right) \\ \boldsymbol{h}_2 &= \gamma_1\left(\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{h}_1 + \boldsymbol{b}_1\right) \\ &\cdots\cdots \\ \boldsymbol{h}_{n+1} &= \gamma_n\left(\boldsymbol{W}_n^{\mathrm{T}}\boldsymbol{h}_n + \boldsymbol{b}_n\right) \\ \boldsymbol{h}_h &= \sigma\left(\boldsymbol{W}_{n+1}^{\mathrm{T}}\boldsymbol{h}_{n+1} + \boldsymbol{b}_{n+1}\right) \end{aligned} \tag{4}$$

where n denotes n-th hidden layer, $\boldsymbol{W}_n$, $\boldsymbol{b}_n$, $\gamma_n(\cdot)$, and $\boldsymbol{h}_{n+1}$ are the weight matrix, bias, activation function and output of $(n+1)$-th hidden layer, respectively. $\boldsymbol{W}_{n+1}$, $\boldsymbol{b}_{n+1}$, $\sigma(\cdot)$, and $\boldsymbol{h}_h$ are the weight matrix, bias, activation function, and output of output layer, respectively. This module takes the features $\boldsymbol{X}_t$ as inputs, and output the $\boldsymbol{h}_h$ as the high-order combination of features. Finally, the output of the forecasting model is combined as follows:

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{H}_t + \boldsymbol{h}_h \\ r_f &= \boldsymbol{W}_u^T f\left(\boldsymbol{W}_v^T\boldsymbol{H} + \boldsymbol{b}_v\right) + \boldsymbol{b}_u \\ \hat{Y}_t &= r_l + r_f \end{aligned} \tag{5}$$

Here, the high-order information is combined with the output of sequential learning part. Then the result is in turn combined with the output of low-order information to form the prediction of the future value at time t.

2.2. Ensemble Method

The above design rely on single model for forecasting. By combining multiple sub-models, ensemble learning often results in significantly superior generalization performance over a single model. In this section we propose an ensemble learning scheme that can boost the accuracy of the forecasting model. Our design based on the observation that tourism data happen in same time can be classified into different category (e.g., year, region, etc.), which we defined as band. Sequence data happens in different band but same time can be combined to improve the robustness of the model prediction. The proposed approach can be divided into the following steps:

(1) *Features division*
The method extract features $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{x}_T\}$ from times series $\{x_1, x_2, \cdots, x_T\}$, and then the features are divided into several bands as

$$\mathbf{X}_t = \{X_t^1, X_t^2, \cdots, X_t^m\}(t = 1, 2, \cdots, T) \tag{6}$$

where each features at time point t are divided into m sub-features.

(2) *Features reconstruction.*
To obtain more representative features for the models, the original features have been reconstructed in this step as following,

$$D = \{X_t^{1\prime}, X_t^{2\prime}, \cdots, X_t^{N\prime}\} \tag{7}$$

where D denotes the new features set, and each features at time point t are reconstruct into N sub-features.

(3) *Sub-models results on the new feature data*
Each model has been trained using the reconstructed features to obtaining the following results,

$$\phi(D) = \{\phi_1(X_t^{1\prime}), \phi_2(X_t^{2\prime}), \cdots, \phi_N(X_t^{N\prime})\} \tag{8}$$

where $\phi_i, i = 1, 2, \cdots, N$ are the sub-models adopted in the proposed approach.

(4) *Models ensemble*
After the above three steps, we can obtain the sub-models for each sub-feature, after then, the operator of *Com* has been adopted to obtain the ensemble model. The overview of the ensemble method is presented as **Figure 2**.

$$\Phi_{c,t} = Com\{\phi_1(X_t^{1\prime}), \phi_2(X_t^{2\prime}), \cdots, \phi_N(X_t^{N\prime})\} \tag{9}$$

where $\phi_{c,t}$ denotes the final results of at time point t. ϕ presents a particular transformation method, such as stacking, hadmard product or linear combination. Based on the combination, we can form multiple sub-models to explore the relations in different bands. The process of the ensemble method is shown in **Algorithm 1**:

Algorithm 1 *The proposed ensemble algorithm.*

Require:
$\{x_1, x_2, \cdots, x_T\} \leftarrow$ Input original time series;
$N \leftarrow$ Number of sub-models;
$\phi \leftarrow$ sub-models function
Ensure: $\Phi_{c,t}$;
1: Extract features of time series as $\{X_1, X_2, \cdots, x_T\}$
2: **while** (i<N) **do**
3: Divide the features into several bands as $X_t = \{X_t^1, X_t^2, \cdots, X_t^m\}$;
4: Select the appropriate sub-features from the divided bands as $D = \{X_t^{i,1}, X_t^{i,2}, \cdots, X_t^{i,N}\}$;
5: Put the sub-models ϕ on the reconstructed features;
6: Calculate the sub-models results as $\{\phi_{i,1}(X_t^{i,1}), \phi_{i,2}(X_t^{i,2}), \cdots, \phi_{i,N}(X_t^{i,N})\}$;
7: Obtain the final combination model as $\Phi_{c,t} = Com\{\phi_{i,1}(X_t^{i,1}), \phi_{i,2}(X_t^{i,2}), \cdots, \phi_{i,N}(X_t^{i,N})\}$;
8: $i = i + 1$;
9: **end while**

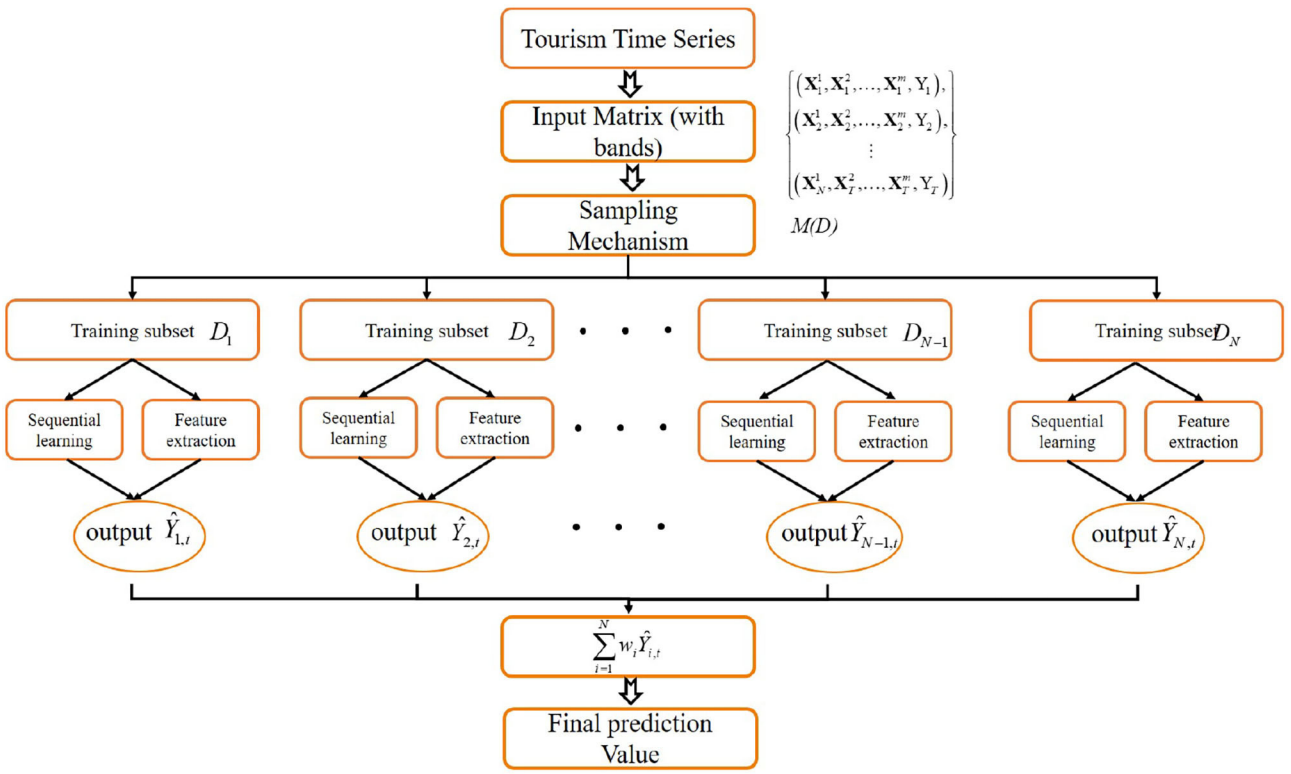


FIGURE 2 | The overview of the proposed ensemble scheme.

3. EXPERIMENTAL EVALUATION

3.1. Dataset Description

In this article, the tourism demand time series from Cheng du Research Base of Giant Panda Breeding[1] (TDPB) is used in our experiment, as shown in **Figure 3**. Cheng du Research Base of Giant Panda Breeding, located at 1375# Panda Road, Northern Suburb, Chenghua District, Chengdu City, Sichuan Province, P.R.China, is 10 km away from the city center and over 30 km from Chengdu Shuangliu International Airport. Proclaimed the "ecological demonstration project for the *ex-situ* conservation of giant pandas." The Base, covering an area of 1,000 mu (66.67 hectares), serves as the world's torchbearer for the ex-situ conservation of giant pandas, scientific research and breeding, public education, and educational tourism. The Base wears its title very well as the sanctuary for giant pandas, red pandas, and other endangered wild animals exclusive to China. TDPB covers the number of visitors to the panda base from 2017 to 2021, with a total of 1,826 pieces of data. The largest number of visitors was on October 4,2018, with 102,305 visitors. The corresponding statistics are shown in the following **Table 1**.

3.2. Evaluation Indicator

To evaluate the effectiveness of our scheme, we use the Root Mean Square Error (RMSE), mean absolute error (MAE), and Mean Absolute Percentage Error (MAPE) metrics to measure the accuracy of the estimates. These metrics are defined as:

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100\%, \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(Y_t - \hat{Y}_t\right)^2}, \tag{11}$$

$$\text{MAE} = \frac{\sum_{t=1}^{N}|Y_t - \hat{Y}_t|}{N}, \tag{12}$$

where Y_t and $\hat{Y}_t$ are the actual demand and predicted demand of future value at time t, respectively, and N denotes the maximum time being predicted in the test set. Clearly, smaller value of the three metrics indicates better forecasting results.

3.3. Other SCHEMES and Parameter Selection

To evaluate the performance of our scheme, we compare it with some state-of-the-art tourism demand forecasting techniques, listed as follows.

- **Decision Tree (DT)** (Song and Ying, 2015): DT falls under the category of supervised learning. It uses the tree representation

[1]http://www.panda.org.cn/english/

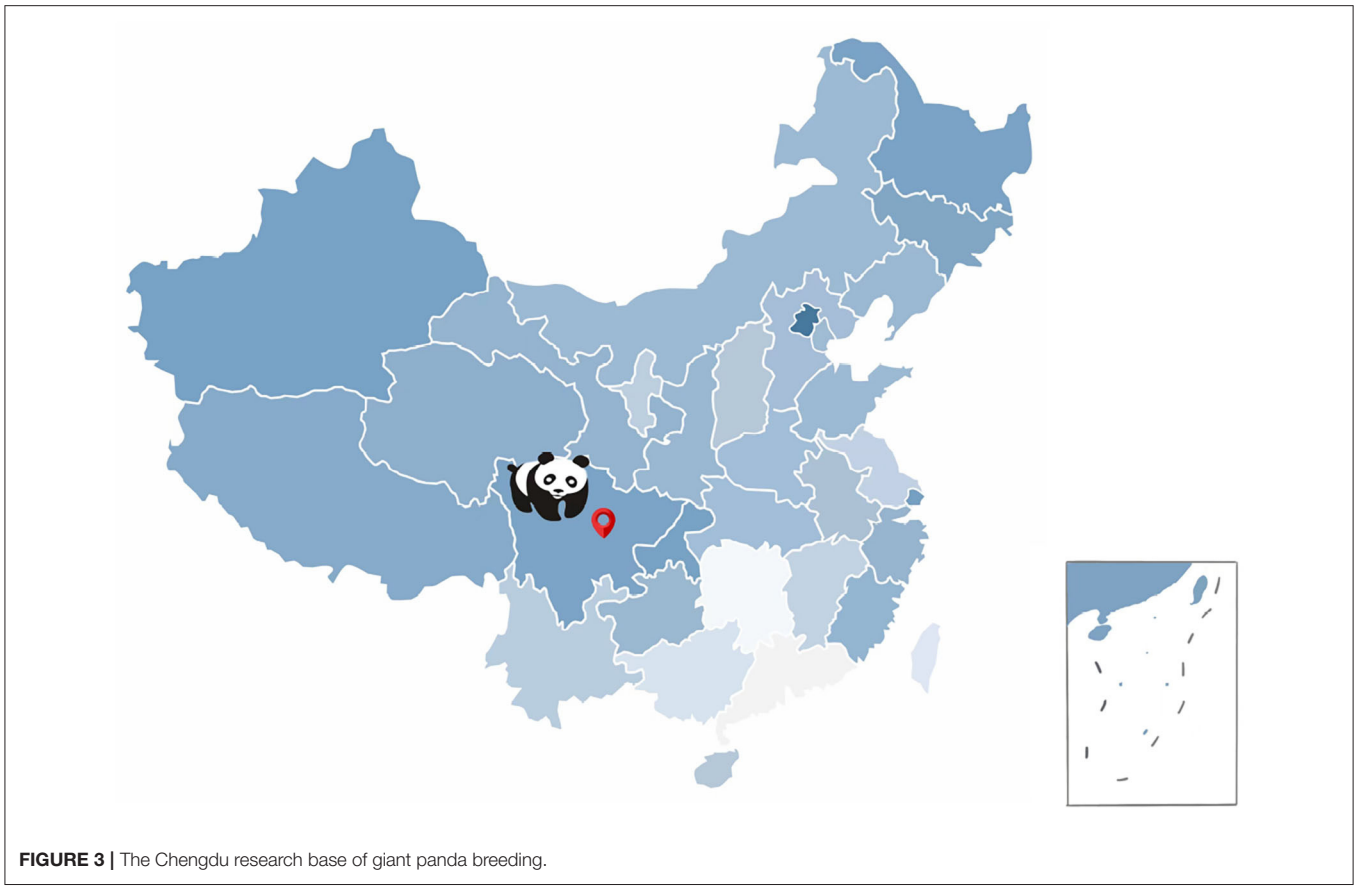


FIGURE 3 | The Chengdu research base of giant panda breeding.

to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

- **Random Forest (RF)** (Kane et al., 2014): RF is a supervised machine learning algorithm that is constructed from multiple decision tree. And RF is a classic ensemble learning algorithm for regression and classification.
- **Extra Trees (ET)** (Hammed et al., 2021): ET also known as Extremely Randomized Trees. Similar to RF, ET builds multiple trees and splits nodes using random subsets of features. Different from RF, ET uses the whole learning sample and splits nodes by choosing cut-points fully at random.
- **Gradient Boosting (GB)** (Gong et al., 2020): GB is an supervised machine learning algorithm used for classification and regression problems. It is an ensemble technique which uses multiple weak learners to produce a strong model for regression and classification.GB relies on the intuition that the best possible next model, when combined with the previous models, minimizes the overall prediction errors. The key idea is to set the target outcomes from the previous models to the next model in order to minimize the errors.
- **Light Gradient Boosting Machine (LGB)** (Fan et al., 2019): LGB is a gradient boosting framework based on decision trees to increase the efficiency of the model. LGB has been widely used in time series problems. And LGB is one of the most popular time series models.

TABLE 1 | The datasets used in our experiments.

Property	2017	2018	2019	2020	2021
Mean	13,252.54	19,870.63	24,347.87	98,65.582	12,055.78
Std	9,611.484	13,261.87	15,725.53	91,31.715	12,643.28
Min	3,419	6,545	7,979	0	86
Max	79,548	102,305	92,608	58,215	79,572

- **Extreme Gradient Boosting (XGBoost)** (Chen et al., 2015): XGBoost is a decision trees based ensemble method which makes use of gradient boosting. It is one of the most powerful algorithms for regression and classification with high speed and performance.

In this dataset, the main factors we consider include holidays, weather, and the number of visitors a few days before the forecasted date. In our experiment, each model uses the same feature input to ensure that the model does not differ in performance due to inconsistent input information. In the proposed model, we separately verified the performance of the RNN module and DNN module with different numbers of neurons in the hidden layer (32, 64, 128, 256, 512). The parameters are initialized using the popular Xavier's approach (Glorot and Bengio, 2010), and the optimizer is stochastic gradient descent algorithm (Bottou, 2012).

3.4. Experimental Results and Analysis

The forecasting results on test set are reported in **Table 2** and **Figure 4**. **Table 2** reports the test results of our experiments on RMSE, MAE, MAPE. It can be seen from **Table 2** that our scheme performs better than other algorithms in RMSE, MAE, and MAPE. Compared with the Decision Tree, ensemble models achieve better performance since they can effectively mine more valuable information. In RMSE, our scheme achieves the best performance with 5673.635, and Random Forest achieves the second best performance with 7790.284, and other models are over 7,800. Similar to RMSE, the MAE value of the proposed scheme outperforms other models, and our scheme achieves at least 1,119 improvement. In terms of MAPE metric, our scheme achieves the lowest value of 26.05%, while the corresponding values of other ensemble models are about 44%. The main reason for this accuracy improvement is that our scheme incorporates historical information and some environmental factors, and extracts high-order features and low-order features from the original data through the feature extracting component and sequential learning component.

TABLE 2 | The values of RMSE, MAE, and MAPE results for different methods.

Model	RMSE	MAE	MAPE (%)
Random forest	7,790.284	4,122.607	41.8044
Extra tree	8,990.606	4,543.783	43.8647
Extreme gradient boosting	8,136.099	4,204.991	45.8960
Light gradient boosting	8,187.854	4,249.289	44.0973
Gradient boosting	7,898.574	4,006.404	44.5745
Decision tree	1,1249.43	5,566.168	51.2416
Our	**5,673.635**	**2,887.705**	**26.0540**

The bold style represents the best test result compared to other model.

Figure 4 draws the observed tourism demand and the forecasting results of different algorithms. Compared with other algorithms, the results of our schemes fits the actual results with the smallest gap. Compared with **Figure 4G**, the forecasting results of these algorithms have low accuracy at specific times. Around the 50-th forecasting date, the **Figure 4G** shows there

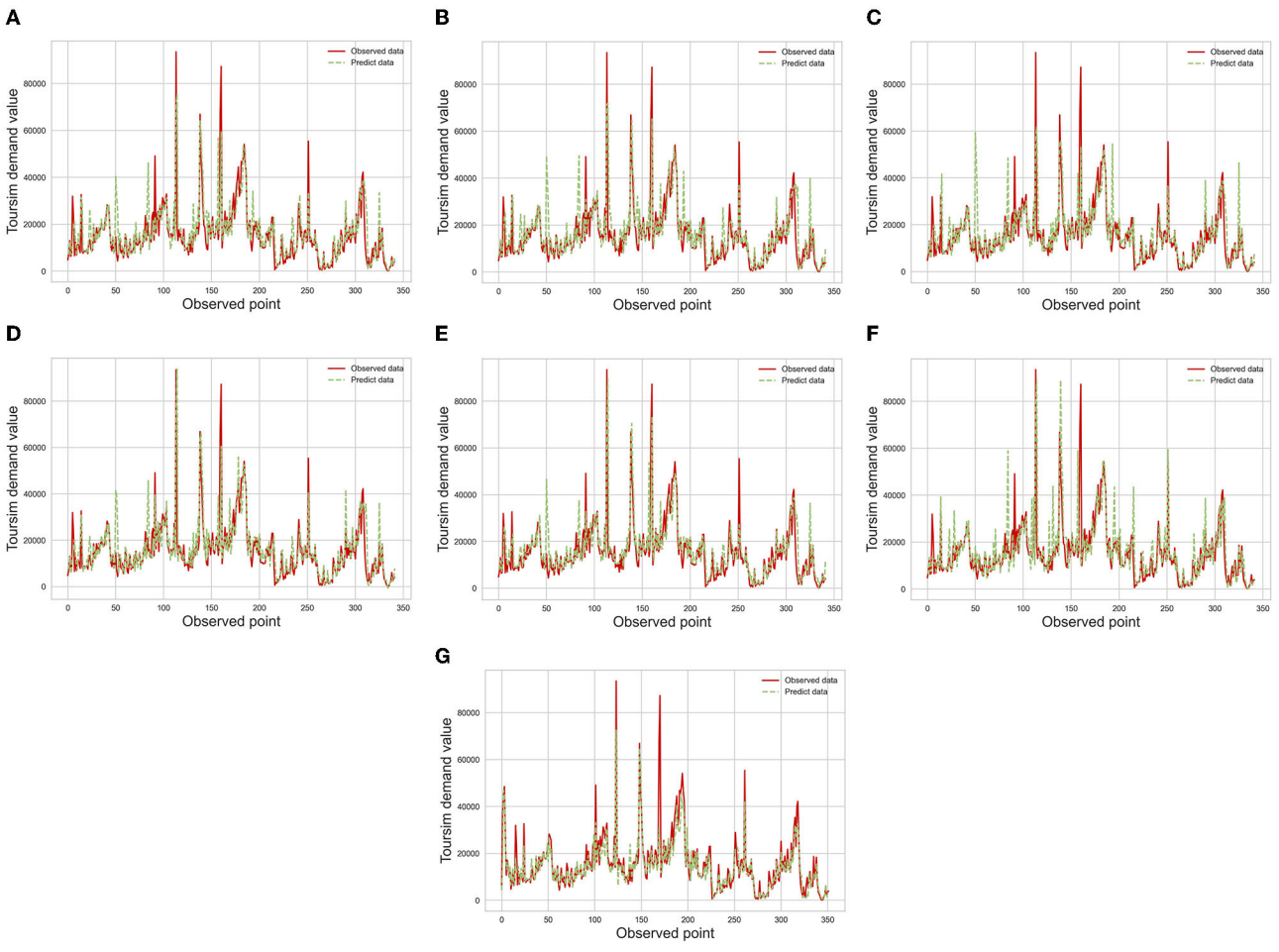


FIGURE 4 | The overview of the forecasting results in different model. **(A)** Forecasting results of RF, **(B)** forecasting results of ET, **(C)** forecasting results of LGB, **(D)** forecasting results of XGBoost, **(E)** forecasting results of GB, **(F)** forecasting results of DT, **(G)** forecasting results of Ours.

TABLE 3 | Forecasting results in 2019.

Metrics	Month	Ours	DT	ET	GB	LGB	RF	XGBoost
MAE	Jan.	**1,715.81**	6,133.29	2,827.98	2,184.47	4,051.26	4,195.45	3,817.47
	Feb.	**3,557.77**	13,209.00	9,335.63	7,497.24	7,545.80	8,121.82	8,535.36
	Mar.	2,091.67	3,007.80	1,776.53	**1,485.57**	1,535.52	1,663.79	1,716.90
	Apr.	**1,535.54**	17,150.67	6,272.12	15,062.00	10,964.92	15,623.47	10,002.17
	May	12,628.74	16,617.00	8,857.97	**8,394.18**	1,1702.60	9,917.97	10,336.93
	Jun.	**3,558.17**	3,829.43	5,162.59	5,208.52	4,842.70	3,981.84	5482.73
	Jul.	8,083.73	7,566.25	**5,498.45**	7,738.89	6,529.65	72,05.41	7,480.24
	Aug.	6,145.03	2,951.20	3,351.42	3,705.38	2,410.52	4,212.86	**2,094.07**
	Sep.	5,083.70	3,169.60	2,755.62	2,766.85	2,583.64	**1,818.64**	2,404.59
	Oct.	3,482.43	9,081.25	8,465.67	3,194.80	11,608.65	7,047.15	**3,132.15**
	Nov.	797.49	3,742.20	**788.34**	879.51	1,602.35	1,188.19	1,219.21
	Dec.	**1,423.71**	1,487.56	2,461.08	1,793.96	2,055.82	1,504.11	2,105.56
	Mean	**4,175.32**	7,328.77	4,796.12	4,992.61	5,619.45	5,540.06	4,860.62
MAPE	Jan.	**11.39**	42.24	17.84	14.31	24.83	27.94	23.31
	Feb.	**15.71**	53.13	52.82	31.57	33.99	46.70	42.46
	Mar.	13.27	17.09	10.70	9.98	**9.20**	10.29	11.62
	Apr.	**8.50**	122.06	44.79	107.47	77.94	112.95	71.20
	May	23.19	33.25	23.78	21.62	32.22	23.67	**21.14**
	Jun.	15.36	**15.36**	20.70	20.15	19.35	16.14	22.69
	Jul.	20.05	19.34	**14.00**	19.74	16.06	18.13	18.62
	Aug.	13.30	6.95	9.27	10.78	9.35	11.52	**6.65**
	Sep.	36.07	27.51	22.80	25.29	22.20	**18.43**	20.78
	Oct.	21.15	51.58	61.94	22.10	84.78	49.27	**19.90**
	Nov.	**4.58**	32.17	4.91	6.19	11.17	9.07	8.89
	Dec.	**10.43**	11.83	21.86	14.96	18.06	12.96	18.98
	Mean	**16.08**	36.04	25.45	25.35	29.93	29.76	23.85
RMSE	Jan.	**2187.95**	1,1047.12	3,406.68	2,594.98	4,799.35	5,720.96	4,701.10
	Feb.	**5,002.60**	169,,67.29	11841.20	9,,420.13	8,520.43	9,541.07	10,509.79
	Mar.	2,414.08	4,787.21	2,321.24	**1,773.48**	1,850.04	2,162.45	2,213.01
	Apr.	**2,008.43**	26,370.80	9,576.12	23364.51	16,473.05	25,022.96	14,946.13
	May	24,616.58	30,788.08	**15,198.75**	16,677.08	19,696.03	17,961.96	19,730.23
	Jun.	**3,918.73**	5,099.65	6,465.84	6,751.16	5,989.56	4,650.59	6,795.89
	Jul.	9,500.12	8,174.99	**6,628.18**	8,956.95	7,832.94	7,889.64	8,826.97
	Aug.	7,743.35	3,589.14	3,770.32	4,260.17	3,648.02	4,316.13	**2,526.00**
	Sep.	6,167.16	4,224.72	3,313.19	3,416.73	3,485.41	**2,903.30**	3,048.56
	Oct.	3,703.89	13,309.60	14,913.55	4,597.15	20,704.75	10,660.30	**3,553.41**
	Nov.	**995.92**	6,746.17	1,248.22	1,220.55	1,693.63	1,552.80	1,377.37
	Dec.	2,131.27	**1,928.57**	3,603.91	2,328.28	2,500.87	2,126.25	2,576.12
	Mean	**5,865.84**	1,1086.11	6,857.27	7,113.43	8,099.51	7,875.70	6,733.71

The bold style represents the best test result compared to other model.

is a significant gab between predicted demands and the actual demands from **Figures 4A–F**, and our scheme avoids this phenomenon as shown in **Figure 4G**. The accuracy improvement can be attributed to the effectiveness of our models for extracting useful information through the recurrent neural network, and the tree-based ensemble models only consider the information of the forecasting date.

To make the conclusions are more believable and robust. We have compared the forecasting results before the outbreak of COVID-19 and after the outbreak of COVID-19, as shown in **Tables 3**, **4**. **Table 3** reports the test results for each month of 2019. It can be seen that in most of cases, the proposed approach can obtain the desire forecasting performance. The results indicate that the proposed model achieves the best performance in January, February, April, June, and December in terms of MAE values. Furthermore, for MAPE values, the proposed model performs better than other algorithms in January, February, April, November, and December, as well as outperforms other algorithms on RMSE in January, February, April, June, and November. Our method holds a good

TABLE 4 | Forecasting results in 2021.

Metrics	Month	Ours	DT	ET	GB	LGB	RF	XGBoost
MAE	Jan.	1,026.58	762.86	840.44	472.51	**437.24**	648.03	892.38
	Feb.	1,704.86	5,668.67	2,049.68	**1,169.37**	1,436.63	3,047.89	1,456.21
	Mar.	1,352.96	2,985.89	1,396.46	2,070.19	**953.09**	1,196.66	1,233.52
	Apr.	2,107.10	2,706.33	1,972.09	**1,540.73**	4,589.47	1,639.77	5,264.18
	May	1,931.20	1,474.20	1,356.71	1,361.01	1,344.46	**694.01**	1,491.45
	Jun.	**1,961.63**	2,264.67	2,161.33	2,753.19	4,391.46	2464.27	3012.06
	Jul.	**4,645.64**	6,414.20	11,967.62	9,351.52	8801.87	9696.32	10,009.47
	Aug.	3,311.54	2,305.43	2,102.71	**1,382.32**	2,070.20	2,201.82	2,006.48
	Sep.	2,046.92	3,696.60	2,798.92	3,121.44	**1,169.52**	3005.99	1889.78
	Oct.	2,742.71	**2,119.40**	6,383.75	4,172.75	5943.67	4,117.29	3,987.71
	Nov.	**836.31**	2,069.71	1,061.51	1,907.52	1,232.21	1,181.79	882.40
	Dec.	**698.14**	1,135.40	1,616.47	2,192.80	1,317.63	937.44	1,500.83
	Mean	**2,030.47**	2,800.28	2,975.64	2,624.61	2,807.29	2,569.27	2,802.21
MAPE	Jan.	38.01	65.80	47.73	35.25	**24.00**	46.90	34.70
	Feb.	31.81	91.25	58.88	31.29	36.67	65.25	**26.93**
	Mar.	14.71	34.22	14.46	21.61	**8.23**	12.66	11.98
	Apr.	11.76	12.02	9.96	10.23	26.14	**9.09**	29.00
	May	11.14	9.41	9.37	9.64	8.07	**4.97**	8.42
	Jun.	**9.43**	18.55	14.50	17.33	28.76	16.78	20.20
	Jul.	**33.92**	51.95	148.52	111.18	98.25	110.27	113.54
	Aug.	64.97	47.03	49.15	**35.57**	50.68	53.06	45.19
	Sep.	27.20	51.72	40.49	48.54	**17.22**	44.60	28.37
	Oct.	19.04	**18.91**	47.81	25.69	33.60	26.54	23.84
	Nov.	151.73	**112.48**	155.61	348.19	181.14	130.48	193.54
	Dec.	**22.87**	26.60	45.22	61.86	37.61	25.92	38.93
	Mean	**36.38**	44.99	53.48	63.03	45.86	45.54	47.89
RMSE	Jan.	1,726.24	1,086.76	1,039.70	**573.62**	586.01	820.50	1,498.36
	Feb.	2,258.79	7,381.18	2,464.34	**1,528.46**	1,823.97	3,761.52	2,504.96
	Mar.	1,579.75	5,983.88	1,904.97	2,541.96	**1386.78**	1,696.02	1,734.95
	Apr.	2,479.79	5,525.57	2,905.72	**1,704.78**	6,484.21	2,192.40	7,484.01
	May	2,484.52	2,124.25	1696.11	1737.53	1,592.35	**885.71**	1,953.02
	Jun.	**2,346.38**	3,752.66	2,392.81	2,898.90	4,638.89	2,742.14	3,410.14
	Jul.	**5,304.81**	9,207.52	15,931.66	12,263.30	11,314.82	12,624.66	13,127.96
	Aug.	4,026.71	2,802.32	2,589.09	**1,798.77**	2,571.36	2,823.40	2,583.59
	Sep.	2,606.43	4,136.84	4,149.44	4,980.07	**1,750.01**	4,178.09	3,032.40
	Oct.	3,721.60	**2,575.87**	10,143.01	7,939.14	12,435.73	6,730.35	7,928.54
	Nov.	**905.88**	3,656.69	1,326.72	2,216.78	1,456.82	1,561.39	1,180.68
	Dec.	**734.18**	1,455.59	2,730.80	3,540.15	1,975.89	1,092.71	2,062.12
	Mean	**2,514.59**	4,140.76	4,106.20	3,643.62	4,001.40	3,425.74	4,041.73

The bold style represents the best test result compared to other model.

performance in January in terms of all metrics, the number of tourists in January is unstable, because the Spring Festival of 2019 is not in January but the winter holiday is beginning in January. What's more, the proposed model also performs well in November and December, there is no official holiday in these 2 months, therefore, the proposed model can both fit the data tendency of peaks and valleys. In general, the mean values of the proposed model in this study performs better than other algorithms. The mean MAE value of our proposed model is 4175.32, which is lower than other compared models and can improve about 700. Similarly, the proposed model can also achieve the desire results in terms of the mean value of MAPE, it owns MAPE value of 16.08% and improves the worst model (DT) about 20%. Otherwise, the same tendency also existed in the metric of RMSE. The averaged MAPE values of our method is 19.96, 9.37, 9.26, 13.85, 13.67, and 7.77%, which are lower than DT, ET, GB, LGB, RF, and XGBoost, respectively. Therefore, the proposed approach in this study achieves the highest accuracy in most cases, which validates the superiority in extracting nonlinear features and useful

information in TDPB through sequential learning and feature extracting components.

Table 4 lists the forecasting results for each month of 2021, the results are worse than the results of 2019, which is mainly caused by the outbreak of COVID-19. Where the COVID-19 occurs in the end 2019, the local government introduce policies to restrict tourism, therefore, the data set of 2020 shows a large fluctuation comparing to the previous years, in this way, the use of data set of 2020 has a negative effect on the model building. The results indicate that our proposed model has a good performance in June, July, November, and December both in MAE, MAPE, and RMSE values. However, comparing with other models, the mean results of our proposed model are superior. After 2020, at the beginning of each year, the repeat outbreak of COVID-19 makes the number of tourists is uncertainty. Therefore, the proposed model is failed to forecast the results of January, February and so on, but it is adapted by itself after the May. It can be seen from **Table 4** that the proposed scheme obtains the best MAPE in 2021 with 36.38%, and achieves 8.61, 17.09, 26.65, 9.48, 9.16, and 11.50% improvement over DT, ET, GB, LGB, RF, and XGBoost, respectively. Other two metrics of MAE and RMSE also support the same conclusions. Similar to 2019, the proposed scheme outperforms other models.

Comparing **Tables 3**, **4**, it can be found that the MAPE value of each model in 2021 is obviously higher than in 2019. Different from 2019, 2021 is the time after the COVID-19 outbreak. As we all known, China government imposes some travel restrictions during COVID-19. Although some contingencies are not considered in our scheme, our scheme still achieves the best accuracy performance, which indicates that our scheme has stronger robustness. Overall, the above experimental results demonstrate that the proposed scheme universally and consistently provide the best accuracy in all test set. This demonstrates the robustness of our method against the unexpected factors including travel restrictions. This also validates the effectiveness of sequential learning component for capturing the corresponding the difference in time.

4. CONCLUSION

Chengdu Research Base of Giant Panda Breeding is the top attractions both at home and abroad. Effective tourism demand forecasting of Chengdu Research Base of Giant Panda Breeding can help managements balance the hotel, traffic, and other public resources. This article proposes an ensemble learning based model for tourism demand forecasting considering environmental factors. The proposed scheme can both explore the sequential relation in tourism time series and extract valuable correlation in features of the estimate time through sequential learning component and feature extracting component. The ensemble method is proposed to fuse multiple forecasting results from sub-models with reconstructed forecasting input. Experimental results on the tourism demand time series from Chengdu Research Base of Giant Panda Breeding demonstrate that the proposed ensemble learning based model not only can achieve higher forecasting accuracy, but also has stronger robustness. Thus, the proposed model holds potential to be widely applied in tourism industry.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

JH: writing, conceptualization, and methodology. DL: modeling and writing and editing. YG: data collection and processing. DZ: data processing and model programming. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Central Guidance on Local Science and Technology Development Fund of Sichuan Province (2021ZYD0156).

REFERENCES

Bottou, L. (2012). "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*, Vol. 7700, eds G. Montavon, G. B. Orr, and K. R. Müller (Berlin; Heidelberg: Springer). doi: 10.1007/978-3-642-35289-8_25

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). *Xgboost: Extreme Gradient Boosting. R Package Version 0.4-2*, 1–4. Available online at: https://cran.r-project.org/web/packages/xgboost/ (accessed April 26, 2022).

Claveria, O., Monte, E., and Torra, S. (2015). Common trends in international tourism demand: are they useful to improve tourism predictions? *Tour. Manage. Perspect*. 16, 116–122. doi: 10.1016/j.tmp.2015.07.013

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., and Zeng, W. (2019). Light gradient boosting machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agric. Water Manage*. 225, 105758. doi: 10.1016/j.agwat.2019.105758

Faruk, D. Ö. (2010). A hybrid neural network and ARIMA model for water quality time series prediction. *Eng. Appl. Artif. Intell.* 23, 586–594. doi: 10.1016/j.engappai.2009.09.015

Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks," in *Appearing in Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010*, (Sardinia), 249–256.

Gong, M., Bai, Y., Qin, J., Wang, J., Yang, P., and Wang, S. (2020). Gradient boosting machine for predicting return temperature of district heating system: a case study for residential buildings in Tianjin. *J. Build. Eng.* 27, 100950. doi: 10.1016/j.jobe.2019.100950

Hammed, M. M., AlOmar, M. K., Khaleel, F., and Al-Ansari, N. (2021). An extra tree regression model for discharge coefficient prediction: novel, practical applications in the hydraulic sector and future research directions. *Math. Probl. Eng.* 2021, 7001710. doi: 10.1155/2021/7001710

Hao, Y., Niu, X., and Wang, J. (2021). Impacts of haze pollution on china's tourism industry: a system of economic loss analysis. *J. Environ. Manage.* 295, 113051. doi: 10.1016/j.jenvman.2021.113051

Ji, L., Zou, Y., He, K., and Zhu, B. (2019). Carbon futures price forecasting based with ARIMA-CNN-LSTM model. *Proc. Comput. Sci*. 162, 33–38. doi: 10.1016/j.procs.2019.11.254

Kane, M. J., Price, N., Scotch, M., and Rabinowitz, P. (2014). Comparison of ARIMA and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinformatics* 15, 276. doi: 10.1186/1471-2105-15-276

Khademi, Z., Ebrahimi, F., and Kordy, H. M. (2022). A transfer learning-based CNN and LSTM hybrid deep learning model to classify motor imagery EEG signals. *Comput. Biol. Med*. 2022, 105288. doi: 10.1016/j.compbiomed.2022.105288

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed M. Arbib (Cambirdge, MA: MIT Press), 255–258.

Li, Y., and Cao, H. (2018). Prediction for tourism flow based on lstm neural network. *Proc. Comput. Sci*. 129, 277–283. doi: 10.1016/j.procs.2018.03.076

Liu, X., Peng, H., Bai, Y., Zhu, Y., and Liao, L. (2014). Tourism flows prediction based on an improved grey GM (1, 1) model. *Proc. Soc. Behav. Sci*. 138, 767–775. doi: 10.1016/j.sbspro.2014.07.256

Ozkok, F. O., and Celik, M. (2022). A hybrid CNN-LSTM model for high resolution melting curve classification. *Biomed. Signal Process. Control* 71, 103168. doi: 10.1016/j.bspc.2021.103168

Qian, L., Zhao, J., and Ma, Y. (2022). Option pricing based on GA-BP neural network. *Proc. Comput. Sci*. 199, 1340–1354. doi: 10.1016/j.procs.2022.01.170

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* 35, 1285–1298. doi: 10.1109/TMI.2016.2528162

Singh, K. R., Neethu, K., Madhurekaa, K., Harita, A., and Mohan, P. (2021). Parallel SVM model for forest fire prediction. *Soft Comput. Lett*. 31, 00014. doi: 10.1016/j.socl.2021.100014

Song, Y.-Y., and Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* 27,130. doi: 10.11919/j.issn.1002-0829.215044

Wang, H.-J., Jin, T., Wang, H., and Su, D. (2022). Application of ieho-bp neural network in forecasting building cooling and heating load. *Energy Rep*. 8, 455–465. doi: 10.1016/j.egyr.2022.01.216

Wang, X. (2022). Research on the prediction of per capita coal consumption based on the ARIMA-BP combined model. *Energy Rep*. 8, 285–294. doi: 10.1016/j.egyr.2022.01.131

Yu, H. (2021). Development of tourism resources based on fpga microprocessor and convolutional neural network. *Microprocess. Microsyst*. 82, 103795. doi: 10.1016/j.micpro.2020.103795

Conflict of Interest: DZ was employed by Chengdu Zhongke Daqi Software Co., Ltd.

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.

Copyright © 2022 He, Liu, Guo and Zhou. 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 Ecology and Evolution

ORIGINAL RESEARCH
published: 26 May 2022
doi: 10.3389/fevo.2022.885955

COVID-19 Pandemic, Air Quality, and $PM_{2.5}$ Reduction-Induced Health Benefits: A Comparative Study for Three Significant Periods in Beijing

Fangfang Cai[1,2], *Kedong Yin*[1,2] *and Mengying Hao*[3]*

[1] School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan, China, [2] Institute of Marine Economy and Management, Shandong University of Finance and Economics, Jinan, China, [3] School of Mathematics and Computer Application Technology, Jining University, Jining, China

Previous studies have estimated the influence of control measures on air quality in the ecological environment during the COVID-19 pandemic. However, few have attached importance to the comparative study of several different periods and evaluated the health benefits of $PM_{2.5}$ decrease caused by COVID-19. Therefore, we aimed to estimate the control measures' impact on air pollutants in 16 urban areas in Beijing and conducted a comparative study across three different periods by establishing the least squares dummy variable model and difference-in-differences model. We discovered that restriction measures did have an apparent impact on most air pollutants, but there were discrepancies in the three periods. The Air Quality Index (AQI) decreased by 7.8%, and SO_2, NO_2, PM_{10}, $PM_{2.5}$, and CO concentrations were lowered by 37.32, 46.76, 53.22, 34.07, and 19.97%, respectively, in the first period, while O_3 increased by 36.27%. In addition, the air pollutant concentrations in the ecological environment, including O_3, reduced significantly, of which O_3 decreased by 7.26% in the second period. Furthermore, AQI and O_3 concentrations slightly increased compared to the same period in 2019, while other pollutants dropped, with NO_2 being the most apparent decrease in the third period. Lastly, we employed health effects and environmental value assessment methods to evaluate the additional public health benefits of $PM_{2.5}$ reduction owing to the restriction measures in three periods. This research not only provides a natural experimental basis for governance actions of air pollution in the ecological environment, but also points out a significant direction for future control strategies.

Keywords: COVID-19, control measures, air quality, $PM_{2.5}$ reduction, health benefit

OPEN ACCESS

Edited by:
Pei Du,
Xi'an Jiaotong University, China

Reviewed by:
Xiaojia Huang,
Lanzhou University, China
Lu Bai,
University of Macau, China

***Correspondence:**
Mengying Hao
mengying.hao@hotmail.com

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Ecology and Evolution

Received: *28 February 2022*
Accepted: *04 April 2022*
Published: *26 May 2022*

Citation:
Cai F, Yin K and Hao M (2022) COVID-19 Pandemic, Air Quality, and $PM_{2.5}$ Reduction-Induced Health Benefits: A Comparative Study for Three Significant Periods in Beijing. Front. Ecol. Evol. 10:885955. doi: 10.3389/fevo.2022.885955

INTRODUCTION

The international spread of Coronavirus Disease 2019 (COVID-19) is affecting public health worldwide (Wang et al., 2020a). The first case was detected in Wuhan before it rapidly disseminated throughout China. Afterward, it spread to more than 210 countries and regions (Ali and Alharbi, 2020), developing as an international health threat (Huang et al., 2020). According to the latest real-time statistics of the World Health Organization, as of 16 March 2022, the total number of confirmed cases of COVID-19 reached 4.5847 trillion, and the cumulative death toll reached

more than 6 million[1]. Since the outbreak of COVID-19, China has gone through three critical periods for initial epidemic prevention: the initial outbreak on 23 January 2020, the cluster of epidemic outbreaks in Beijing's Xinfadi market on 11 June 2020, and the policy that people stayed at their current residing localities during the Spring Festival in 2021. As the control measure of the COVID-19 outbreak was an external interference, its impact can be quantified by adopting quasi-experimental approaches. As a result, the three periods provide a unique opportunity to investigate the air quality response to such anthropogenic disruptions. The impact of the blockade triggered by the initial outbreak of COVID-19 in the world has attracted a great deal of attention (Briz-Redón et al., 2021; Gupta et al., 2021; Ju et al., 2021; Querol et al., 2021). People were frightened, and the blockade rules were strictly observed since there was no known drug or vaccine against the disease when COVID-19 first broke out during the Spring Festival in 2020 and the epidemic recurrence in Beijing's Xinfadi market. This is due to the efforts of the government and the fear of the people.

During the policy period of staying in place for Chinese New Year 2021, people were relieved from initial fear following a year of cohabitation with COVID-19, partial vaccination, and partial relaxation in industrial sectors to avoid the economic hardships experienced during the first lockdown cycle (Mahato and Pal, 2022). It is necessary to analyze the impact and differences of control measures in different stages on air pollution along with the results of the differences for a more detailed and scientific air pollution prevention and control strategy. It is universally known that China implemented an array of dramatic control measures during COVID-19 to minimize human interaction and prevent the virus from spreading further (Kraemer et al., 2020), including staying put during the spring festival, closing schools and workplaces and implementing remote office and teaching, canceling public events, restricting gatherings and traffic, carrying out strict home isolation, and even blocking out the entire city[2].

Undoubtedly, enacting these restrictions was accompanied by a significant economic loss in addition to affecting the daily lives of people around the world (Anderson et al., 2020; Meo et al., 2020; Brodeur et al., 2021). Firstly, the capital market has been severely impacted due to the disease quickly expanding across the country. Additionally, countries have imposed travel restrictions and restricted production activities to prevent the spread from escalating again, which has triggered concern of an impending economic crisis and downturn (Nicola et al., 2020). In particular, the pandemic has also affected the supply of the energy sector, such as the oil and power sector (Chiaramonti and Maniatis, 2020). The blockade has achieved great success in curbing COVID-19. For it to work, the world has to deal with severe economic crises, chronic hunger, mass unemployment, and a range of other problems (Berkowitz and Basu, 2021; Rasul et al., 2021), all of which the world is still coping with (Kassa and Grace, 2020). In addition, the fiscal policy effectively contributed to the economic development in the time duration of the COVID-19 pandemic thus far (Ren et al., 2022). Generally, the global pandemic of COVID-19 has had an extensive and far-reaching impact on the development of the world economy. It will not only reconstruct the core concepts and fundamental connotations of economic globalization but also reshape the ecosystem of the world economy and promote fundamental changes in the global governance system (Jones, 2020).

Nevertheless, there have been unintentional ecological and environmental benefits during the pandemic. Much empirical research has focused on the impact of prevention measures in the COVID-19 pandemic on air pollution. Zhang et al. (2021) employed a two-way fixed effects model and an interrupted time-series analysis to explore the impacts of the control measures on air pollution during the COVID-19 outbreak. They detected that the related decrease in air pollutant concentrations was more evident over time since the lockdown began. Wang et al. (2021) utilized a difference-in-differences (DID) model to assess the implication of intra-city mobility declines on air pollution in 325 Chinese cities, finding that cities with restriction measures have a 12.2% greater decrease in AQI. Furthermore, this reducing impact varies with distinct types of air contaminants. Lu et al. (2021) constructed a machine learning prediction model to quantify changes in NO_2, SO_2, $PM_{2.5}$, and PM_{10} levels induced by the first-level public health emergency response of 174 cities in China to COVID-19. They found the short-term emission control effect ranges from 53.0 to 98.3% for all cities, and southern cities show a significantly stronger effect than northern cities ($p < 0.01$). Compared with megacities, small-medium cities show a similar control effect on NO_2 and SO_2, but a larger effect on $PM_{2.5}$ and PM_{10}.

Additionally, the change in air pollutants indicated high spatial heterogeneity. The provinces with a reduction in $PM_{2.5}$ and PM_{10} >20 and >40% reduction in NO_2 during the impact period were mainly concentrated southeast of the "Hu Line." In addition, different types of cities show different response and resilience patterns to the pandemic (Zeng and Wang, 2022). Although COVID-19's blockade has led to a temporary improvement in air quality (He et al., 2020), it comes at the cost of curbing economic development. Furthermore, most of the reduction in pollutant concentrations in 2020–2021 appears to result from a long-term declining trend rather than COVID-19 (Hwang and Lee, 2022). Moreover, during the resumption of work, the economy recovered, and there was an increase in energy consumption. CO_2 and NO_2 emissions increased significantly, reaching the level before the blockade (Zhou et al., 2022).

Air pollution has caused great damage to the ecological environment and human health, causing serious economic losses (Hao et al., 2021). As a result, numerous scholars have carried out studies on the health benefits of improving air quality in recent years (Zhang et al., 2007; Chen et al., 2010; Xie, 2011). Several studies have been conducted on the health effects of COVID-19 lockdown through their impact on air quality. Wang et al. (2021) assessed the excess risk (ER) of six pollutants and the AQI based on health risk (HAQI) to determine the health impacts of

[1]Reference to the China national emergency broadcasting: http://www.cneb.gov.cn/2022/03/16/ARTI1647384446343556.shtml.

[2]Reference to the report on China's practices in combating COVID-19: http://cn.chinadaily.com.cn/a/202004/21/WS5e9e45afa310c00b73c786ed.html.

various air pollutants. They found that $PM_{2.5}$ was the most health risk factor and HAQI values were all lower during COVID-19. Liu et al. (2021) used a novel COVID-19 government response tracker dataset to quantify the causal impacts of lockdown measures on air pollution using a DID approach. They discovered that across the 76 nations and areas involved in the restriction measures of the COVID-19 outbreak, the estimated avoided premature deaths owing to air pollution decreases range from 99,270 to 146,649 fatalities. Additionally, Shi et al. (2021) studied the long-run health implications of reducing $PM_{2.5}$ during the Thirteenth Five-Year Plan. They determined the total premature deaths acquiring the relative risk of $PM_{2.5}$ exposure from former research and found that the yearly $PM_{2.5}$ level decreased from 49.7 $\mu g/m^3$ in 2015 to 33.2 $\mu g/m^3$ in 2020. Premature deaths declined from 1,186,201 and 446,415 in 2015 to 997,955 and 368,786 in 2020, respectively. In Seoul and Daegu, improved air quality has lowered premature mortality and saved health costs (Seo et al., 2020).

Moreover, Lam et al. (2022) selected fifteen cities worldwide to investigate the public health co-benefits of $PM_{2.5}$ reduction during a period when various non-pharmaceutical interventions (NPIs) were adopted in the COVID-19 pandemic. Due to the high $PM_{2.5}$ background with a large population, there were tremendous health co-benefits for cities in India and China. New Delhi has received the largest co-benefits, saving over 14,700 premature deaths. Bai et al. (2022) examined the $PM_{2.5}$ variations between the COVID-19 lockdown and found that the national average of $PM_{2.5}$ decreased by 18 $\mu g/m^3$, and the mean $PM_{2.5}$ for most sites decreased by 30–60%. The total avoided premature death due to $PM_{2.5}$ reduction is 9,952 in China, with a dominant contribution (94%) from anthropogenic emission changes.

Beijing is one of the capital economic centers. It is regarded as a major metropolis globally, playing a vital role in political, economic, cultural, scientific, and technological innovation. However, air pollution has long been a major concern in this area, particularly during the winter (Vu et al., 2019). Due to the peculiar geography, air pollution in this location is frequently more severe than in other places when weather conditions are adverse (Wang et al., 2017). In addition, Beijing was impacted by contaminants carried from other locations in addition to local air pollution, particularly during periods of serious air pollution (Zhang et al., 2015). Furthermore, since the implementation of the restriction policy after the initial COVID-19 epidemic, except for power plants and large enterprises, almost all factories have been closed and traffic was also restricted. Moreover, Beijing is the city with the highest proportion of people, as high as 70.9%, staying put during the Spring Festival in 2021[3]. Such restrictions should have greatly improved the state of the ecological environment. However, severe regional air pollution persisted despite the adoption of stringent controls, and greater efforts should be made to avert heavy air pollution (Wang et al., 2020b).

There have been substantial studies conducted on the influence of preventive and control measures on air pollutants all over the world since the outbreak of COVID-19 which have drawn some meaningful conclusions. However, there are still at least three deficiencies in most existing studies, which are of great research value or significance. Firstly, most studies merely analyze the situation in a specific period without comparative studies on diverse periods. Suppose the distinctions in different periods and the causes of the differences are ignored. It is not adequate to dig out more factors that need to be considered to prevent and control air pollution. Secondly, the data scale used in most studies is provincial data. However, there are gaps in economic development and the natural environment among various regions in the same province. Therefore, a unified air pollution control strategy cannot be adapted to local conditions. Thirdly, most studies solely analyze its impact on air quality without additional research value. However, our ultimate goal in improving air quality is to minimize the damage to human health and economic losses. As a result, the assessment of related health benefits can make us feel the benefits of air pollution prevention, encouraging human beings to take the initiative to reduce emissions and air pollution. The research focus of this study attempts to make up for the deficiency of the above research. This study aims to (i) determine whether the COVID-19 control measures have a causal impact on the air quality; (ii) evaluate the health benefits and avoidable economic costs due to changes in $PM_{2.5}$ concentration during COVID-19; and (iii) carry on the comparative analysis of three critical periods in China regarding the impact of restriction measures on air pollution, health benefits, and avoidable economic losses of $PM_{2.5}$ changes due to COVID-19.

Our study adopted data including daily air pollutant concentrations, meteorological information of 16 urban areas in Beijing, and designed the least squares dummy variable model and DID method to evaluate the implication of the initial COVID-19 breakout, the reemergence of the epidemic in the Xinfadi market, and the policy period that people stayed in place for Lunar New Year of 2021, respectively. We discovered that restriction measures did have an apparent impact on most air pollutants, but there were discrepancies in the three periods. The AQI fell by 7.8%, and SO_2, NO_2, PM_{10}, $PM_{2.5}$, and CO concentrations were lowered by 37.32, 46.76, 53.22, 34.07, and 19.97%, respectively, during the initial outbreak of COVID-19 in 2020, while O_3 increased by 36.27%. In addition, the air pollutants concentrations in the ecological environment, including O3, reduced significantly, of which O3 decreased by 7.26%, and AQI and $PM_{2.5}$ fell by 22.61% and 45.12%, respectively, when the epidemic outbreak occurred in Beijing's Xinfadi market. Moreover, AQI and O_3 concentration increased slightly in comparison to the same period in 2019, while other pollutants dropped, with NO_2 being the most apparent decrease during the policy of staying in place for the Lunar New Year of 2021. And Spring Festival had a great impact on the concentration of NO_2 and CO.

In addition, based on health-related data, such as exposed population and outpatient morbidity and mortality, we innovatively estimated the health benefits and avoided economic costs brought by the changes in $PM_{2.5}$ pollution as a result of the pandemic. Specifically, we applied Poisson regression relative risk models and environmental value evaluation approaches to

[3]The data comes from Zhaopin recruitment: https://www.zhaopin.com/.

analyze the avoided health risks and economic losses of $PM_{2.5}$ reduction in the 16 municipal districts of Beijing in these three periods, adopting the secondary standard limit of 35 $\mu g \cdot m^{-3}$. In other words, we used the impact of control measures on changes in air pollutants to calculate the indirect health impacts of the pandemic. We found differences in health effects and avoided economic loss among the three periods in each urban area due to the gaps in $PM_{2.5}$ changes, exposure population, outpatient incidence, and mortality in different urban areas. In addition, the avoided total health and economic loss owing to the $PM_{2.5}$ reduction affected by the restriction actions in three pandemic periods were 82,747.65 million yuan [95% CI (3,406.4, 10,879.1)], 11,143.71 million yuan [95% CI (3,826.43, 16,949.1)], and 871.65 million yuan [95% CI (350.54, 1,165.95)], respectively.

The main innovations and contributions of this research are as follows:

(a) Previous research merely studied and compared the air pollutant concentrations over a year or months before and after the outbreak. The unique characteristics of different critical periods have received little attention. In this study, we carried out targeted research for various crucial periods in China. We made comparisons both in the changes in air pollutant concentrations and indirect health benefits of the COVID-19 epidemic by affecting $PM_{2.5}$ pollutants. Targeted and more detailed control measures for serious air pollution can be formulated by exploring the particularity of different vital epidemic periods.
(b) Most former studies have relied on relatively macro and large-scale data from various countries, provinces, or cities to carry out their research. However, there are distinctions in the ecological environment, resource elements, air quality, and the intensity of control measures in specific regions, such as towns or urban areas, which cannot be generalized. Consequently, district-level and low-scale data are adopted to conduct research for elaborate prevention, control, and governance of regional air pollution.
(c) We still have no idea how the specific value of such changes is reflected in our daily lives if the study solely reports how and to what extent COVID-19 affects variations in air pollutant concentrations. Further research value is assessed in this study, which includes the health effects and economic benefits of $PM_{2.5}$ reductions induced by COVID-19. This work will present more valuable information for air pollution prevention and control decision-makers. Still, it will also play a vital role in implementing early warning and prevention measures related to air pollution and human health.

The following is the organization of the rest of the study. Section Methodology explains the empirical strategy and the value assessment methods. Section Data describes the data sources and processing. Section Empirical Study on the Effect of Control Measures on Air Quality discloses the effects of restriction efforts on air quality over three periods, respectively. Section $PM_{2.5}$ Reduction-Induced Health Benefit Evaluation assesses the indirect health effects and avoided economic loss of the prevention actions in the COVID-19 pandemic by affecting air pollution. Section Discussions on Recommendations for Meticulous Control of Air Pollution discusses the recommendations of control measures to improve air quality, and Part 7 contains the conclusions. The flowchart of this study is as follows (**Figure 1**).

METHODOLOGY

This section describes the methods adopted in this article, and it consists of two parts. Part one is the empirical strategy and part two is the value assessment methods.

Impact of Restriction Measures on Air Pollution

The empirical model was mainly utilized to evaluate the impact of control measures on air pollutants in three periods. The least-squares dummy variable model and DID model were used in this study.

Least Squares Dummy Variable Model

First, we adopted the least squares dummy variable (LSDV) method to explore the impact of the Chinese New Year in 2020 and COVID-19 restrictions on air pollution. Hence, this study mainly used an urban panel data model with a fixed effect, as follows:

$$ln\ p_{it} = \alpha_0 + \alpha_1 Covid + \alpha_2 Holidays + \alpha_3 Other\ \ holidays + \theta_1 W_{it} + \vartheta_1 X_{it} + \mu_i + \pi_t + \varepsilon_{it} \quad (1)$$

where $\ln P_{it}$ represents the explained variable, which is obtained by logarithmizing the daily average concentration of air pollutants in region i on day t. The value of "Covid" was 1 if the day falls within the initial COVID-19 epidemic period (24 January 2020–29 February 2020). "Covid" denoted the core explanatory variable for this model. If COVID-19's prevention and control measures can improve air quality, its regression coefficient should be significantly negative. "Holiday" was 1 when it falls within Chinese New Year (4 February 2019–10 February 2019 or 24 January 2020–2 February 2020). If not, it has a value of 0. It belongs to a dummy variable. "Other holidays" connotes the dummy variable for vacations except for Chinese New Year. W_{it} represents the weather variables in region i on day t. They are variables added to control the influence of meteorological conditions on air pollutants. X_{it} denotes other control factors, incorporating the impact of Month. μ_i connotes urban fixed effects, which is used to control regional heterogeneity. π_t represents the date fixed effects and ε_{it} denotes the error term.

Second, we still employed the LSDV model to analyze the influence of the outbreak of COVID-19 in Beijing's Xinfadi market on air contaminants. Our main method is following:

$$ln\ \tilde{p}_{it} = \beta_0 + \beta_1 Market + \beta_3 Other\ \ holidays + \theta_2 W_{it} + \vartheta_2 X_{it} + \mu_i + \pi_t + \varepsilon_{it} \quad (2)$$

The COVID-19 epidemic broke out in Beijing's Xinfadi market on 11 June 2020, and the number of cases was not cleared until 6 August 2020. As a result, the value of "Market" was 1 if the day falls within the epidemics outbreak period in Beijing's Xinfadi

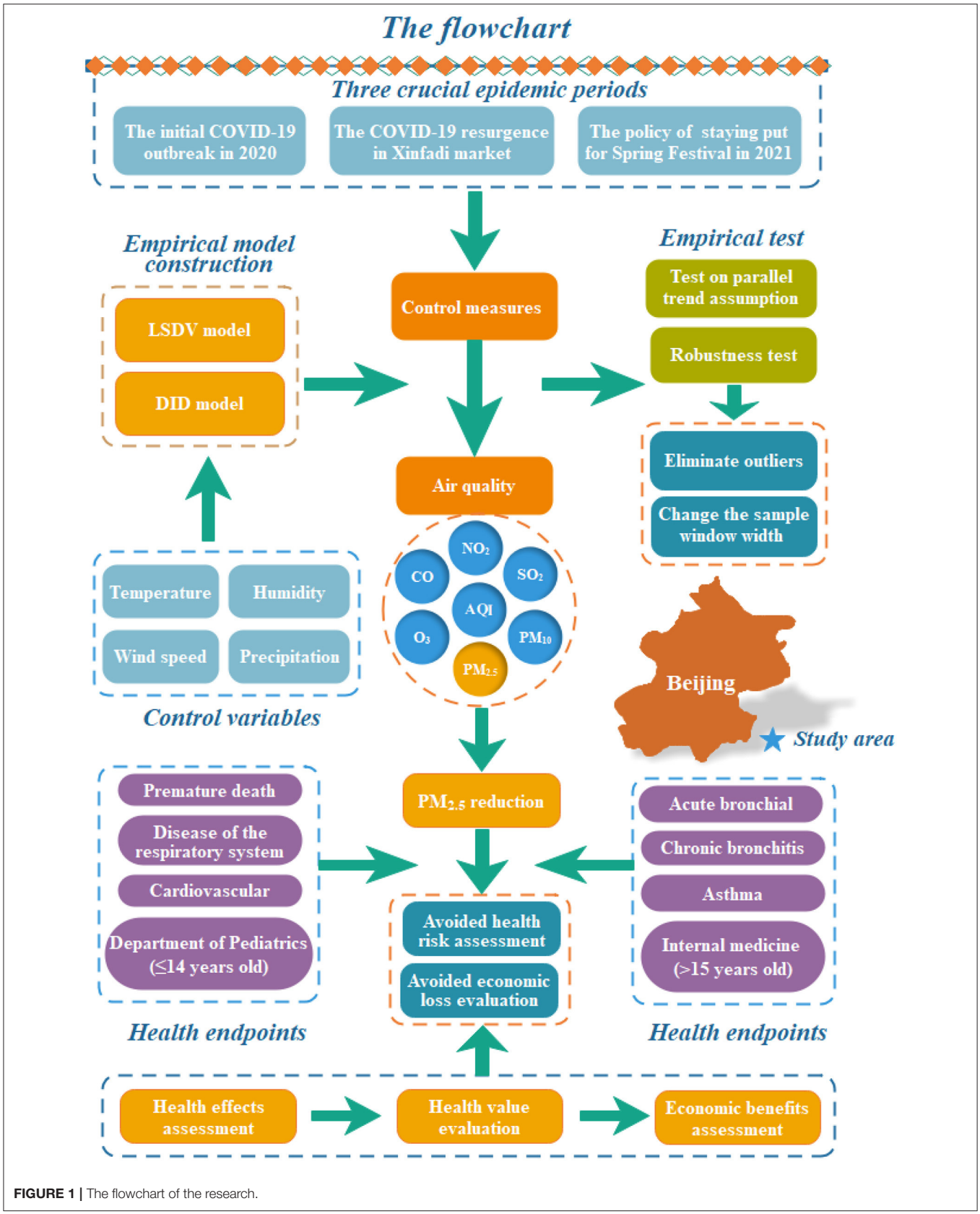


FIGURE 1 | The flowchart of the research.

market (11 June 2020–6 August 2020). "Other holidays" is a dummy variable for holidays during the epidemic outbreak in the Xinfadi market. The explanation of $ln\tilde{P}_{it}$, W_{it}, X_{it}, μ_i, π_t, and ε_{it} is the same as the equation (1).

Differences-In-Differences Model

The DID model (Jiménez and Perdiguero, 2017) has been frequently utilized to assess the causal influence of control measures taken by the government on the atmosphere and to distinguish policy effects from other impacting variables. These models may avoid uncontrolled and unexpected factors in the time leading up to and after adopting the regulations (Li and Lin, 2017).

As the policy of staying put during the Spring Festival in 2021 coincided with the Chinese Spring Festival in 2021, which influences factory production, people's travel, and entertainment, separating the impact of the Spring Festival holiday effect becomes indispensable (Fu and Gu, 2017). Consequently, this study uses the DID model to estimate the impact of a policy, which stated that people stayed in place for the Lunar New Year in 2021, on air quality. The experimental group consisted of air quality data from 2021, while the control group consisted of data from the same time in 2019 due to the unusual epidemic and complicated situation in the Chinese New Year 2020. We collected data for 10 days before and after the Chinese New Year in 2021 (1 February 2021–21 February 2021) and 2019 (25 January 2019–13 February 2019). The DID model goes like this:

$$\ln P_{it} = \delta_0 + \delta_{12} Holiday \times Treat + \delta_1 Holiday + \delta_2 Treat + \theta_3 W_{it} + \mu_i + \pi_t + \varepsilon_{it} \quad (3)$$

where "Treat" belongs to a grouping dummy variable whose value was 1 if it falls within 2021, and it took the value 0 for 2019. The value of "Holiday" was 1 when it falls after the Chinese New Year (4 February 2019 or 11 February 2021) during the sample period. The interaction term "Holiday × Treat" connotes the policy effect of staying put during the holiday of Lunar New Year's Eve in 2021.

Hausman test is applied to determine whether the model is valid, and we have corroborated that the usage of fixed effects in Equations (1, 2) is reasonable. The explained variables in all equations take the value of its logarithm to eliminate potential heteroscedasticity and decrease data fluctuations. The relative variations are easier to grasp (Bel and Holst, 2018; Lin and Zhu, 2019).

$PM_{2.5}$ Reduction-Induced Health Effects and Economic Benefits

This section is the methods employed in three steps of health benefits assessment. First, we evaluated the environmental health effect combined with the Poisson regression relative risk model and the relation coefficient β of exposure-response model. Second, we assessed the environmental health value using the value of a statistical life. Third, we evaluated the health and economic benefits according to the calculated results of each health endpoint's health effects and unit economic value in the previous two steps.

TABLE 1 | Exposure-response coefficients of $PM_{2.5}$ and the occurrence rates under health endpoints.

Diseases	Health endpoints	*E* Value (%)	*β* Value (95% CI)
Death	Early death	0.539	0.296 (0.076, 0.504)
Hospital admissions	Disease of respiratory system	1.33	0.109 (0.000, 0.221)
	Cardiovascular	0.69	0.068 (0.043, 0.093)
Outpatient service	Department of pediatrics (≤14 years old)	7.25	0.056 (0.020, 0.090)
	Internal medicine (>15 years old)	22.33	0.049 (0.027, 0.070)
Diseases	Acute bronchial	3.72	0.790 (0.270, 1.300)
	Chronic bronchitis	0.694	1.009 (0.366, 1.559)
	Asthma	0.94	0.210 (0.145, 0.274)

Sources from Kan and Chen (1989), Xie et al. (2009), and Liu et al. (2010); Beijing Municipal Bureau of Statistics; Statistical Information Center of National Health and Family Planning Commission. β represents the percentage increase (%) of the morbidity and mortality per 10 μg/m^3 of the $PM_{2.5}$ rise. CI, confidence interval.

Environmental Health Effect Assessment

Below, the estimation equation simulates the associated population health risks (Huang et al., 2012). This expression illustrates the correlations between changes in $PM_{2.5}$ concentration and changes in human health endpoints.

$$E = E_0 \times \exp[\beta(C - C_0)] \quad (4)$$

$$\Delta E = P \times (E - E_0) \quad (5)$$

where C denotes the daily $PM_{2.5}$ concentration and C_0 represents the limit of the secondary concentration of $PM_{2.5}$ pollutants. In this article, we chose 35 μg/m^3 as C_0. P is the number of exposed populations. This study selects the resident population at the end of the year to replace. β indicates a coefficient in the explosion-response relationship that connotes the percentage of the health impacts of variation for each 10 μg/m^3 rise in $PM_{2.5}$ levels, as indicated in **Table 1**. E and E_0 are the health effects under C and C_0 concentrations, respectively, and ΔE is the change of health effects. According to the existing research (Huang and Zhang, 2013; Xie et al., 2014; Li et al., 2017a), the selected health endpoints affected by $PM_{2.5}$ are premature death, respiratory disease hospitalization, cardiovascular disease hospitalization, internal medicine clinic, pediatrics clinic, acute bronchitis, chronic bronchitis, and asthma. Additionally, the variations of the health impacts (ΔHE) induced by $PM_{2.5}$ can be acquired utilizing the formula below based on the exposure population (*Pop*).

$$\Delta HE = Pop \times (E - E_0) = Pop \times E \times \left[1 - \frac{1}{\exp(\beta \times (C - C_0))}\right] \quad (6)$$

Research on exposure-response relationship coefficients was conducted to increase the reliability of air pollution damage

to residents and reduce errors as much as possible. Numerous scholars, such as Kan et al. (2004), Xie et al. (2009), Lv and Li (2016), have ultimately proposed an exposure-response relationship coefficient that is suitable for China taking into account differences in pollutant concentrations at home and abroad and also the different effects of pollutants such as $PM_{2.5}$ on various races and populations. **Table 1** shows the relationship coefficient (β) of exposure-response model and the benchmark incidence (*E* value) of corresponding health endpoints based on previous research findings.

This study adopted the achievement reference method since the relative deficiency of the latest analyzed data in China. The incidence of premature death was acquired from the Beijing Regional Statistical Yearbook (CSY., 2020) and the Statistical Bulletin of National Economic and Social Development [4] published on the Beijing Bureau of Statistics' official website. The hospitalization rates of respiratory and cardiovascular diseases were attained from the Fifth National Health Service Survey and Analysis report released by the Statistical Information Center of the National Health and Family Planning Commission [5] in 2015. Additionally, the incidence of medical and surgical outpatients was estimated from the percentage of Beijing's medical and surgical outpatients in the total number of outpatients according to the literature (Kip Viscusi et al., 1991; Xie et al., 2015; Lv and Li, 2016). Finally, the baseline incidence (*E* value) of acute and chronic bronchitis and asthma was determined concerning the research results of predecessors.

Environmental Health Value Evaluation

The value of a statistical life (VOSL) is how residents are willing to lower the risk of death by employing money. This study applied the VOSL method to assess the economic cost of premature death due to $PM_{2.5}$ based on VOSL research results in Beijing (Xie, 2011; Huang and Zhang, 2013). In recent years, Beijing residents' statistical life expectancy value has been estimated by adopting per capita gross domestic product (GDP) and the Consumer Price Index (CPI) (Matus et al., 2012; Lu et al., 2017; Giannadaki et al., 2018; Maji et al., 2018). The following is the calculation method:

$$\mathrm{VOSL_t} = \mathrm{VOSL}_k \times (1 + \%\Delta P + \%\Delta G)^{\beta 1} \tag{7}$$

where VOSL_t and VOSL_k are Beijing's statistical life value in t and k years, respectively, $\%\Delta P$ and $\%\Delta G$ are the growth rates of CPI and per capita GDP in Beijing from k to t years. β1 is the coefficient of income elasticity, and its value in this study is 0.8 (Lanzi et al., 2016).

The disease cost approach is utilized to estimate outpatient and hospitalization expenses in this study. The economic loss induced by outpatient and hospitalization expenses includes two parts: the outpatient and hospitalized medical costs per capita and the lost work time due to disease treatment. The unit cost estimation formula is as follows (Zhang et al., 2007):

$$\mathrm{EC_i} = \mathrm{EC_{i,p}} + (\mathrm{GDP_p} \times \mathrm{T_{i,L}}) \tag{8}$$

where *i* is the *i*th health endpoint, EC_i is the unit economic loss of outpatient or hospitalization, $EC_{i,p}$ is the per capita medical expense, including direct and indirect medical expenses, GDP_p represents the cost of absenteeism per capita on a daily basis, estimated by the daily per capita GDP, and $T_{i,L}$ is the number of days of absenteeism caused by treating *i*th health endpoint diseases. The time spent missing work in the outpatient clinic is calculated using a 0.5-day rule (Wei and Shi, 2018; Han et al., 2019). In addition, the outpatient expenses are taken from the China Health and Family Planning Statistical Yearbook of 2020 issued by the National Health and Family Planning Commission. The hospitalization expenses and length of stay are obtained by adopting Wei and Shi's (2018) estimation method.

Since chronic bronchitis treatment is slow, the treatment costs are difficult to calculate. Consequently, the disease cost method is not suitable for determining its economic cost. This study applied the achievement reference method to estimate chronic bronchitis according to 32% of Kip Viscusi's et al. (1991) statistical life value. Additionally, the unit economic loss of acute bronchitis was calculated by outpatient cost according to Huang and Zhang's (2013) ratio of the unit economic value of outpatient clinic to acute bronchitis. **Table 2** shows the average hospitalization or outpatient days, medical expenses, and unit economic loss cost for each health endpoint.

Health and Economic Benefits Evaluation

Eventually, the overall economic loss of health impacts induced by $PM_{2.5}$ can be evaluated utilizing the formula below (Yin et al., 2017).

$$EL = \sum \Delta HE_i \times V_i \tag{9}$$

where EL denotes the overall health and economic benefits of residents brought by $PM_{2.5}$ reduction and V_i connotes the unit economic value of the *i* th health endpoint, obtained by the equation below (Hammitt and Robinson, 2011).

$$V_{Beijing_2020} = V_{Beijing_2009} \times \left(\frac{Income_{Beijing_2020}}{Income_{Beijing_2009}} \right)^{e} \tag{10}$$

where $Income_{Beijing_2009}$ and $Income_{Beijing_2020}$ represent Beijing's income in 2009 and 2020 correspondingly, *e* is the elastic coefficient, and then the $V_{Beijing_2009}$ and $V_{Beijing_2020}$ are the values of 2009 and 2020 correspondingly in Beijing. The economic value in 2020 ($V_{Beijing_2020}$) can be acquired referring to the $Income_{Beijing_2009}$ in Hammitt and Robinson (2011) and Equation 10.

[4] *Beijing Municipal Bureau of Statistics: Statistical Bulletin of National Economic and Social Development in Beijing* (2019). Available online at: http://tjj.beijing.gov.cn/tjsj_31433/sjjd_31444/202003/t20200302_1673395.html (accessed August 18, 2020).

[5] Statistical Information Center of National Health and Family Planning Commission. (2015). *2013 Report on the Fifth National Health Service Survey and Analysis.* Beijing: China Union Medical University Press.

TABLE 2 | Days of hospitalization, outpatient service and their medical expenses for health endpoints and unit economic loss value in 2020.

Average hospitalization days			Average hospitalization cost per time		
Cardiovascular	Disease of the respiratory system	Asthma	Cardiovascular	Disease of the respiratory system	Asthma
8.1	10.2	10.7	30381.3	8150.03	11196.5
	Hospitalization				
VSL (unit: Ten thousand yuan per person)	Disease of respiratory system (unit: Yuan per person)		Cardiovascular (unit: Yuan per person)	Outpatient service (unit: Yuan per person)	
374.7	13466.4		34748.5	807.4	
Average outpatient days per case(unit: Yuan per case)	Average outpatient expenses per case (unit: Yuan per case)		Indirect cost per case		
			Hospitalization	Outpatient service	
0.5	561.4		706	20	
Acute bronchial (unit: Yuan per person)	Chronic bronchitis (unit: Yuan per person)		Asthma (unit: Yuan per person)	Average GDP per person per day (unit: Yuan)	
3915.9	119.9		16032.9	452	

The data are derived from Xie et al. (2015) and the relevant literature described in Sections Environmental Health Value Evaluation or are estimated through the formulas, and related methods in Section Environmental Health Value Evaluation of this study.

DATA

This section describes the data sources and processing in empirical analysis and health benefits assessment of $PM_{2.5}$ pollution, respectively, including air quality and meteorological data required in empirical analysis, and related data employed in health benefits assessment of $PM_{2.5}$ pollution.

Air Quality and Meteorological Data

Air quality data was taken from the Qingyue Open Environment Data Center [6]. We attained relevant air quality data, including the daily AQI, $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3 concentrations from 16 urban areas in Beijing from the following periods: (i) 1 January 2019–29 February 2020, the period containing the Chinese New Year 2020 and the initial COVID-19 outbreak; (ii) 11 June 2020–6 August 2020, the period including the outbreak of Beijing's Xinfadi market from the beginning to the end; and (iii) 1 February 2021–21 February 2021, the period containing the implementation of policy that people stayed in place for Lunar New Year in 2021. Meteorological data were sourced from Huiju data[7] and the National Meteorological Information Center[8]. We gathered daily meteorological data of three periods same as air quality information, including mean temperature, mean relative humidity, mean wind speed, and precipitation (accumulated over 8 h) on a daily basis for 16 urban districts in Beijing. **Table 3** presents the summary statistics of our important variables.

Data Related to Health Benefit Evaluation of $PM_{2.5}$ Pollution

The GDP, per capita GDP, CPI, and year-end resident population (exposed population) data in various districts of Beijing are obtained from the Beijing Regional Statistical Yearbook of 2020 (CSY., 2020) or the Beijing Bureau of Statistics (or government), the Beijing District Bureau of Statistics (or government), and the official website of the National Bureau of Statistics[9], etc. In addition, the mortality, morbidity, per capita hospitalization, and outpatient expenses were obtained or estimated from the Beijing Regional Statistical Yearbook of 2020, the Statistical Bulletin on National Economic and Social Development of Beijing in 2020 published by the Beijing Municipal Bureau of Statistics[10], the Statistical Information Center of National Health and Family Planning Commission[11], and survey data from the China Asthma Alliance.

It is challenging to acquire or estimate detailed data, such as mortality, the statistical value of life, prevalence rate, outpatient, hospitalization expenses, and length of stay, for each district in Beijing. Therefore, this study uniformly adopts the corresponding data of Beijing in the same year. Moreover, this article uses the corresponding data in the next year or the preliminary accounting data from the district people's government

[6] *The Qingyue Open Environment Data Center*. Available online at: http://dataold.epmap.org/.

[7] *Huiju Data*. Available online at: http://hz.hjhj-e.com/home/.

[8] *The National Meteorological Information Center*. Available online at: http://www.nmic.cn/.

[9] *The National Bureau of Statistics*. Available online at: http: //www.stats.gov.cn/.

[10] Beijing Municipal Bureau of Statistics. *Statistical Bulletin on National Economic and Social Development of Beijing in 2020*. Available online at: http://www.beijing.gov.cn/zhengce/gfxwj/sj/202103/t20210312_2305538.html.

[11] The Statistical Information Center of National Health and Family Planning Commission. (2020). *China Health and Family Planning Statistical Yearbook*. Beijing: China Union Medical University Press.

TABLE 3 | Summary statistics of key model variables.

Variables	Unit	Obs.	Mean	Std. Dev.	Min	Max
Panel A. 2020 sample: the initial COVID-19 outbreak						
AQI	N/A	6,800	87.81	48.35	15	433
$PM_{2.5}$	$\mu g/m^3$	6,800	45.56	39.21	3	399
PM_{10}	$\mu g/m^3$	6,800	70.74	46.71	2	457
SO_2	$\mu g/m^3$	6,800	4.516	3.245	1	37
NO_2	$\mu g/m^3$	6,800	34.95	18.24	2	116
CO	mg/m^3	6,800	0.726	0.421	0.100	3.600
O_3	$\mu g/m^3$	6,800	92.39	58.21	2	316
Temperature	°C	6,800	11.95	11.49	−8.500	32.20
Humidity	%	6,800	47.77	18.48	10	92
Wind speed	m/s	6,800	1.615	0.810	0.300	4.600
Precipitation	mm	6,800	1.060	4.166	0	37.50
Panel B. 2020 Xinfadi sample: the epidemics outbreak in Beijing's Xinfadi market						
AQI	N/A	2,544	85.90	36.50	30	209
$PM_{2.5}$	$\mu g/m^3$	2,544	33.28	21.85	3	129
PM_{10}	$\mu g/m^3$	2,544	63.23	35.14	11	323
SO_2	$\mu g/m^3$	2,544	3.395	1.925	1	17
NO_2	$\mu g/m^3$	2,544	23.57	11.05	2	71
CO	mg/m^3	2,544	0.545	0.257	0.100	1.600
O_3	$\mu g/m^3$	2,544	127.9	48.46	30	309
Temperature	°C	2,544	20.11	7.471	2.100	31.70
Humidity	%	2,544	48.91	18.78	11	90
Wind speed	m/s	2,544	2.779	2.072	0.800	8.800
Precipitation	mm	2,544	1.226	3.902	0	25.80
Panel C. CNY2021 sample: the policy period of staying in place for Chinese New Year 2021						
AQI	N/A	640	80.17	56.69	27	324
$PM_{2.5}$	$\mu g/m^3$	640	50.89	49.89	4	274
PM_{10}	$\mu g/m^3$	633	80.31	55.87	11	355
SO_2	$\mu g/m^3$	640	5.370	3.397	1	22
NO_2	$\mu g/m^3$	640	27.59	16.30	2	92
CO	mg/m^3	640	0.762	0.474	0.200	3.200
O_3	$\mu g/m^3$	640	64.05	17.49	18	152
Temperature	°C	640	0.0950	3.792	−7.300	9.200
Humidity	%	640	34.52	18.31	11	87
Wind speed	m/s	640	1.940	0.984	0.600	4.700
Precipitation	mm	640	0.0300	0.159	0	1

and the Bureau of Statistics to supplement the missing or unpublished data.

EMPIRICAL STUDY ON THE EFFECT OF CONTROL MEASURES ON AIR QUALITY

This section indicates the effect of restriction measures during the COVID-19 outbreak on the air pollutant concentrations in three significant periods. Each period has two parts: the graphical analysis and empirical model regression results.

The Initial COVID-19 Outbreak in 2020

In this part, we firstly analyzed the changes in air contaminants during the initial COVID-19 outbreak in 2020 based on the variation of pollutant concentration distribution diagrams. Secondly, we conducted a further simulation and indicated the influence of preventive measures on air pollution based on LSDV model regression results.

The Variation of Pollutant Concentration Distributions

As a visual demonstration of changes in air pollutants, the mapping tool in ArcGIS10.2 was employed to graphically portray the pollutant concentration distributions of 16 urban areas in Beijing. The work is achieved by comparing the emissions of air contaminants in each urban area before and after taking measures to stop the further spread of the COVID-19 epidemic. As shown in **Figure 2**, the temporal and spatial distribution of different air pollutants in various urban areas has obvious heterogeneity. Specifically, except for no significant change in the O_3 concentration, the AQI and other air pollutant concentrations seem to be much lower than the normal concentration during the epidemic prevention and control measures, which provides supporting evidence for the effectiveness of epidemic control efforts to decrease air pollution. In the following section, we further simulate the effect obtained by the LSDV model.

The Impact of Control Efforts on air Pollution

First, we conducted a regression employing the LSDV model on the influencing variables of air pollution from 1 January 2019 to 29 February 2020, as shown in **Table 4**. We discovered that except for the increase in O_3, which increased by 36.27%, other air pollutant concentrations declined significantly when restriction measures were implemented, indicating that they have alleviated the air pollution. The AQI dropped by 7.8%, while NO_2 reduced by 46.76%. At the same time, $PM_{2.5}$, SO_2, PM_{10}, and CO had different degrees of decline, which fell by 34.07, 37.32, 53.22, and 19.97%, respectively. Spring Festival had a great impact on the concentration of NO_2. Furthermore, other holidays reduced pollutants even if their impact on PM_{10} and O_3 was small. In addition, related weather variables explain plenty of changes in air pollutant concentrations. The time variable (month) shows the long-run temporal tendency of the monthly fluctuation of air pollutants (Wang et al., 2010) and substantially impacts air quality. These constants are statistically crucial because they effectively balance the errors that other terms in the model do not take into account and ensure that the residual's average value is zero.

Mobility restrictions and rapid reduction of pollutants emitted by vehicles and industry following the lockdowns are possible explanations for short-term air quality improvement (Dang and Trinh, 2021; Jiang et al., 2021). The decline in economic activity and traffic restrictions during the epidemic directly led to changes in China's energy consumption, resulting in a decline in carbon emissions and air pollution levels, alleviating ecological and environmental pollution (Muhammad et al., 2020). Due to the restrictions on human activities and traffic, the pollutant concentrations in China have sharply dropped in a few days, especially NO_2 and PM_{10} (Dutheil et al., 2020; Liu et al., 2020).

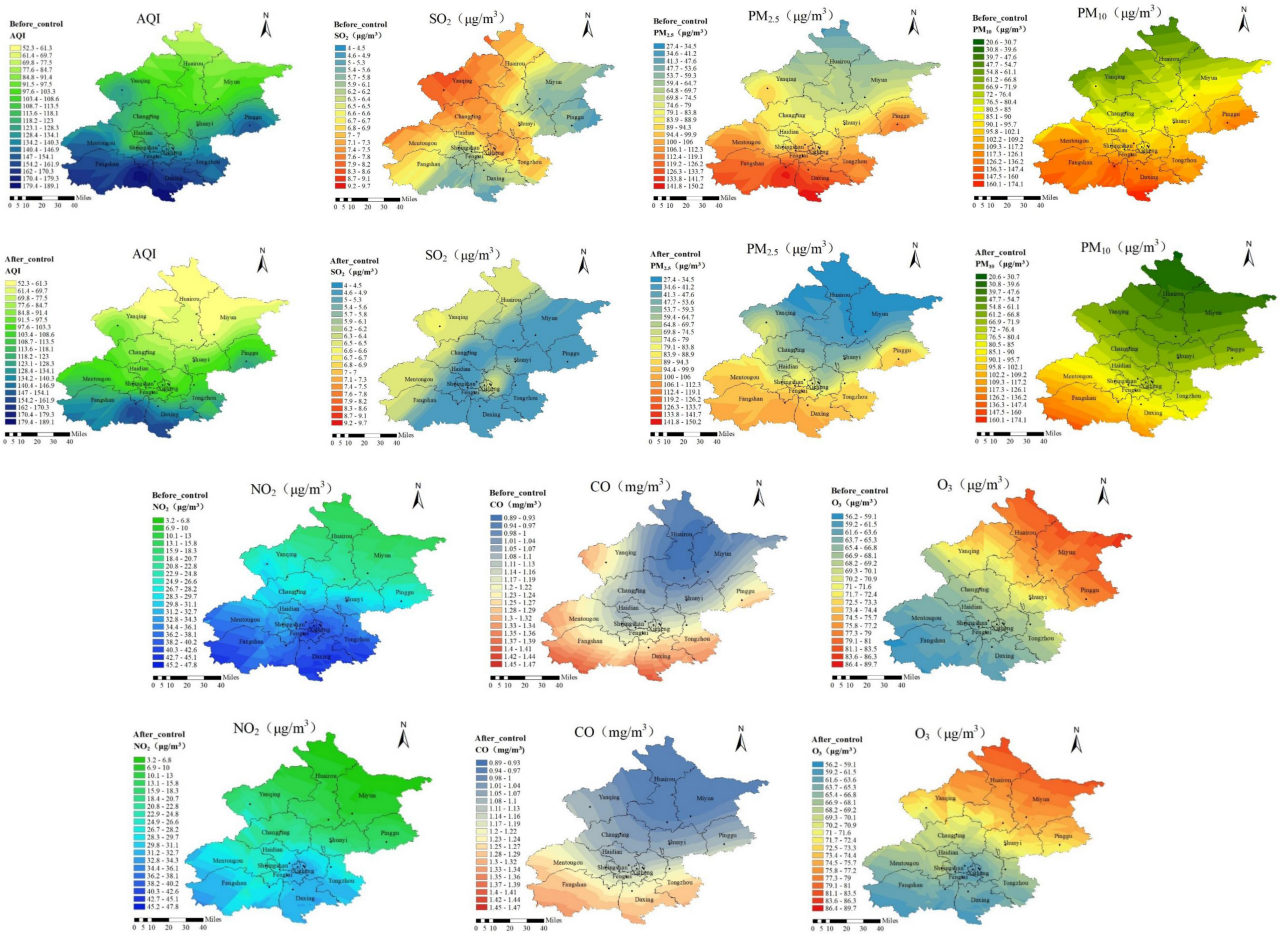


FIGURE 2 | Variations of AQI, SO_2, $PM_{2.5}$, PM_{10}, NO_2, CO and O_3 concentration distribution before and after control measures.

O_3 participates in photochemical reactions, and its concentration is often opposite to the change of emission due to the non-linear characteristics of the chain reaction (Kim et al., 2017).

The COVID-19 Resurgence in the Xinfadi Market

In this part, we firstly analyzed the variations in air quality during the cluster of COVID-19 outbreaks in Beijing's Xinfadi market based on the changing trend of pollutant concentrations diagrams. Secondly, we further simulated and revealed the relationship between control measures and air quality based on LSDV model regression results.

The Changing Trend of Pollutant Concentrations

As a visual demonstration of changing trend in pollutant concentrations, we used the mapping tools in Minitab19 to graphically describe the time trend of each pollutant. This is achieved by comparing the emissions of air pollutants in 16 urban areas of Beijing before and after the outbreak in the Xinfadi market. As illustrated in **Figure 3**, the pollutant concentrations in each urban area briefly decrease after taking measures, in which $PM_{2.5}$ is the most obvious and the decline of AQI is relatively slow. In addition, the concentrations of $PM_{2.5}$, CO, and AQI have an upward trend after a period of control measures. The time trends of different air pollutants in various urban areas are generally similar, proving evidence of the effectiveness of epidemic control measures to reduce air pollution. Below, we will further simulate the effect derived from the LSDV model.

The Effect of Restriction Actions on Air Quality

Second, we still implement a regression applying the LSDV method on the contributors impacting air pollution from 11 June 2020 to 6 August 2020 (**Table 5**). Similar to the first period, findings revealed that enacting prevention measures during Beijing's Xinfadi market outbreak significantly lowered air pollutant concentrations. The AQI dropped by 22.6% and NO_2 reduced by 34.6%. Meanwhile, $PM_{2.5}$, SO_2, PM_{10}, and CO concentrations fell by 45.1, 35.0, 46.7, and 18.5%, respectively, indicating that control measures improved air quality. PM_{10} and $PM_{2.5}$, related to vehicle exhaust emissions and industrial processes, declined most obviously. Compared with the first

TABLE 4 | Panel model regression results using least squares dummy variable (LSDV).

Variables	ln (AQI)	ln ($PM_{2.5}$)	ln (PM_{10})	ln (SO_2)	ln (NO_2)	ln (CO)	ln (O_3)
Covid	−0.078***	−0.3407***	−0.5322***	−0.3732***	−0.4676***	−0.1997***	0.3627***
	(0.0245)	(0.028)	(0.0391)	(0.0359)	(0.0284)	(0.0177)	(0.0218)
Holiday	0.1071***	0.1299***	0.209***	0.1526***	−0.4374***	−0.0326	0.3366***
	(0.0185)	(0.0185)	(0.0334)	(0.0376)	(0.0203)	(0.0242)	(0.0191)
Other holidays	−0.1501***	−0.1455***	−0.0389*	−0.1211***	−0.1701***	−0.1827***	−0.0147*
	(0.0083)	(0.0094)	(0.0194)	(0.0229)	(0.0186)	(0.0124)	(0.0079)
Temperature	0.012***	−0.0113***	−0.0008	−0.0098***	−0.0118***	−0.0123***	0.0568***
	(0.0014)	(0.0014)	(0.0013)	(0.0021)	(0.0015)	(0.0014)	(0.0016)
Humidity	0.0095***	0.0243***	0.0057***	−0.0035***	0.0014**	0.0146***	−0.006***
	(0.0002)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Wind speed	−0.1143***	−0.2325***	−0.16***	−0.2123***	−0.3506***	−0.2118***	0.0948***
	(0.011)	(0.0125)	(0.0074)	(0.0167)	(0.0083)	(0.01)	(0.0099)
Precipitation	−0.0265***	−0.0291***	−0.0196***	−0.0001	−0.0142***	−0.0117***	−0.0065***
	(0.0005)	(0.0008)	(0.0008)	(0.0011)	(0.0012)	(0.0005)	(0.0003)
Month	−0.0331***	−0.0692***	−0.0305***	−0.0527***	−0.0059**	−0.0304***	−0.0496***
	(0.0013)	(0.0014)	(0.0021)	(0.003)	(0.0023)	(0.0015)	(0.0015)
Constant	4.158***	3.3017***	4.2962***	2.2897***	4.1587***	−0.4611***	4.0006***
	(0.0441)	(0.0577)	(0.0359)	(0.0796)	(0.0333)	(0.0538)	(0.0306)
Urban fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,800	6,800	6,676	6,797	6,800	6,800	6,800
R-squared	0.2685	0.3948	0.1319	0.2831	0.4568	0.4834	0.6574

*, **, *** *represent significance at the 10, 5, and 1% levels correspondingly.*

period, the reduction of PM_{10}, SO_2, NO_2, and CO was less. The recurrence of the outbreak in the Xinfadi market was only a small-scale aggregated outbreak in Fengtai District and was brought under control in about a week. Furthermore, Industrial sectors return to work and production with partial relaxation of the epidemic's control measures to avoid the economic hardships experienced during the first lockdown. In addition, appropriate people's travel and traffic flow are allowed, which might explain these results. It is worth mentioning that the O_3 concentration has also decreased by 7.26%, which may be the reduction and dispersion of emissions in NOx and volatile organic compounds (VOCs) precursors (Yang et al., 2019). Other festivals have a greater impact on PM_{10} and $PM_{2.5}$ and less impact on NO_2, SO_2, CO, which may be due to restrictions on travel, play, and related activities since the initial outbreak of COVID-19 in 2020. Thus, NO_2, SO_2 and CO levels have few changes during the epidemic outbreak of Xinfadi market. Meteorological factors also show strong explanatory power that is coherent with the previous regression findings.

The Policy Period of Staying in Place for Lunar New Year in 2021

In this part, we first compared the air quality changes in the policy period of staying in place for Lunar New Year based on the comparison of pollutant concentrations variations between 2020 and the same period in 2019. Secondly, we further simulated and indicated the impact of control measures on air quality based on DID model regression results.

The Comparison of Pollutant Concentrations With 2019

For a visual representation of the comparison between the control and the treatment groups in air pollution, we used the mapping tools in Minitab19 to graphically describe the time trends before and after the Spring Festival in 2019 and 2021. As shown in **Figure 4**, except for O_3, the concentrations of all pollutants decreased during the Spring Festival policy in 2021, especially that of AQI, $PM_{2.5}$, and PM_{10}. In addition, for a period after the end of the policy, the concentration of pollutants in 2020 was often lower than that in 2019, which provides evidence of the effectiveness of epidemic control measures to reduce air pollution. Below, we will further simulate the effect acquired from DID model.

The Influence of Control Measures on Air Quality

Finally, we adopted the DID model to explore the impact of the policy period people stayed put during the Spring Festival in 2021 on air quality. This model can quantify the net impact of prevention efforts on air pollutants in the COVID-19 pandemic by separating restriction measures from the Chinese New Year holiday effect. The air quality data in 2019 were chosen as the control group (no controls), and the air quality data for 2021 were used as the treatment group. The DID analysis (**Table 6**) revealed that prevention efforts had improved air quality in comparison to the same time in 2019, with the most significant impact on NO_2. The pollutant was primarily related to a decline in economic development and transportation constraints, which resulted in reduced energy usage and lowered emissions (Filonchyk et al., 2020). This was

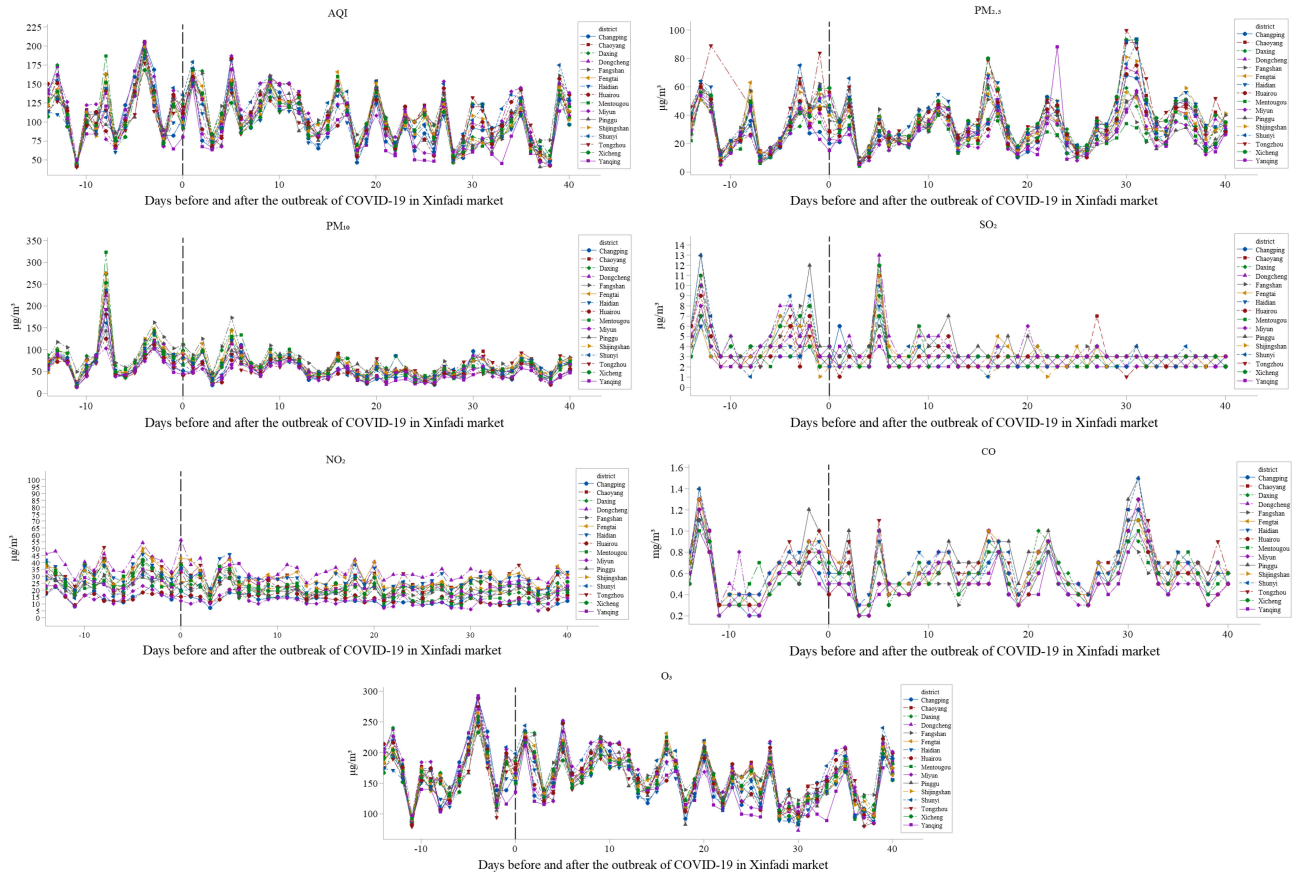


FIGURE 3 | The changing trend of pollutant concentrations before and after the COVID-19 outbreak in Beijing's Xinfadi market. The vertical open black line denotes the start of the cluster of epidemic outbreaks in the Beijing's Xinfadi market.

TABLE 5 | Panel model regression results adopting LSDV model.

Variables	ln (AQI)	ln ($PM_{2.5}$)	ln (PM_{10})	ln (SO_2)	ln (NO_2)	ln (CO)	ln (O_3)
Covid	−0.2261***	−0.4512***	−0.4669***	−0.3504***	−0.3459***	−0.1845***	−0.0726***
	(0.0241)	(0.0379)	(0.0308)	(0.0293)	(0.023)	(0.0249)	(0.019)
Other holidays	0.1462***	0.2186***	0.3192***	0.1515***	0.0687**	0.0818***	0.0641**
	(0.0283)	(0.0428)	(0.0392)	(0.0368)	(0.0321)	(0.0262)	(0.0277)
Temperature	0.0377***	0.0114***	0.0345***	0.0129***	0.0173***	0.0059***	0.0397***
	(0.0014)	(0.0022)	(0.002)	(0.0016)	(0.0016)	(0.0014)	(0.001)
Humidity	0.004***	0.024***	0.0002	0.0015***	0.0073***	0.0198***	−0.0014***
	(0.0005)	(0.0007)	(0.0007)	(0.0006)	(0.0006)	(0.0005)	(0.0004)
Wind speed	0.0201***	−0.0023	0.0099**	−0.0033	−0.021***	0.0078**	0.0169***
	(0.0035)	(0.0052)	(0.0042)	(0.005)	(0.0039)	(0.0037)	(0.0025)
Precipitation	−0.0073***	−0.0271***	−0.0114***	−0.0007	−0.0181***	−0.0108***	−0.0012
	(.0019)	(0.0025)	(0.0022)	(0.0016)	(0.0015)	(0.0018)	(0.0015)
Constant	3.4342***	2.0713***	3.477***	1.0588***	2.9335***	−1.6544***	4.0057***
	(0.0358)	(0.0644)	(0.054)	(0.0534)	(0.0535)	(0.0428)	(0.0272)
Urban fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,385	2,380	2,348	2,385	2,385	2,385	2,385
R-squared	0.3969	0.3464	0.2263	0.1121	0.3402	0.485	0.5642

, * *represent significance at the 5 and 1% levels correspondingly.*

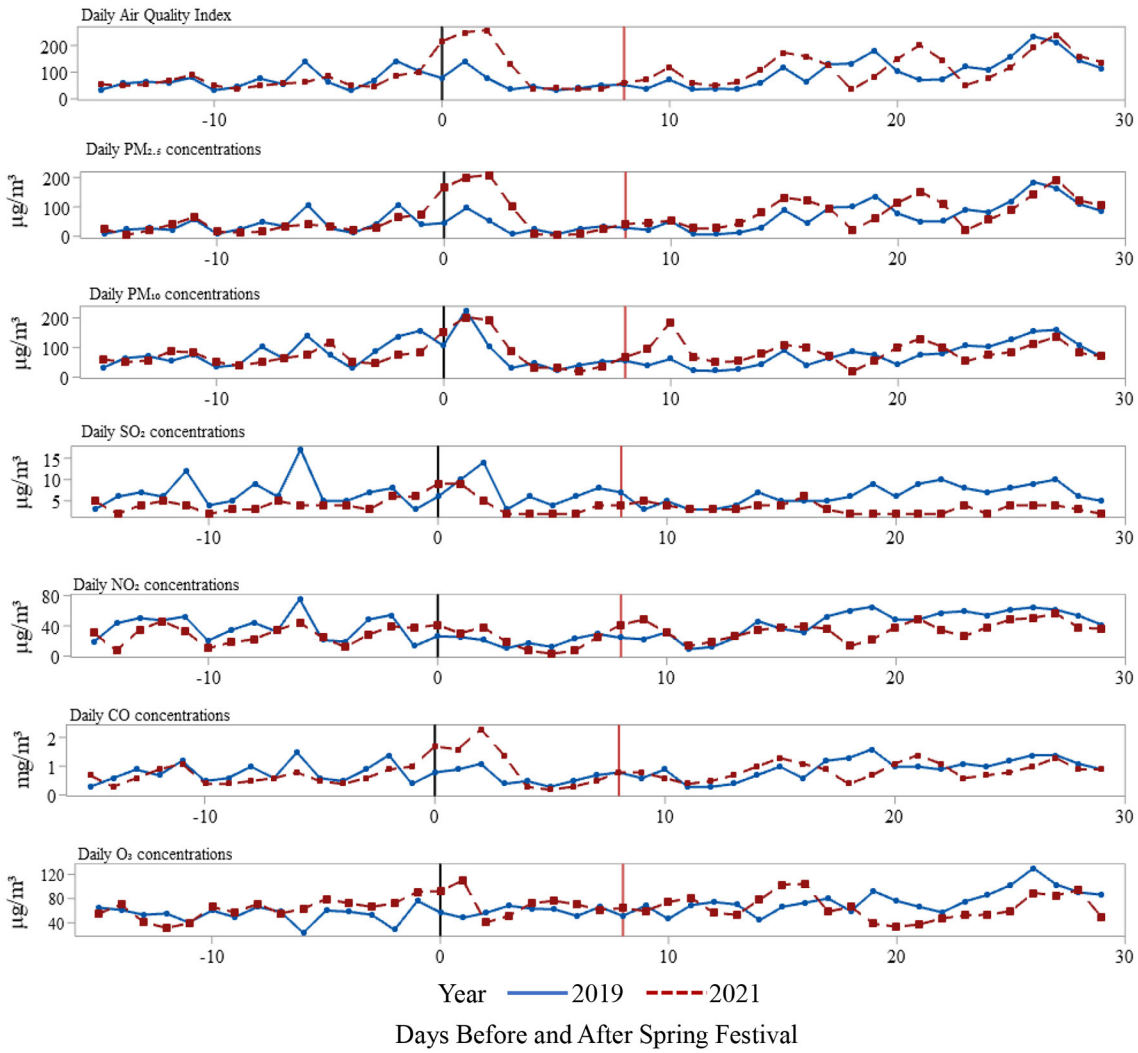


FIGURE 4 | The comparison of pollutant concentrations changes between 2020 and the same period in 2019. The vertical solid black line connotes the beginning of Chinese New Year holiday, while the vertical solid red line denotes the end of Chinese New Year holiday.

followed by PM_{10}, SO_2, and CO, which dropped by 30.35, 42.38, and 39.13%, respectively. However, AQI increased during this period compared to the same time in 2019, and no significant effect on $PM_{2.5}$ was observed.

On the one hand, some pollutant concentrations have not significantly fallen in comparison with the first period. People staying at their current residing localities during the Spring Festival in 2021 may reduce the migration index but has a relatively small impact on the urban travel intensity index. On the other hand, several major pollutant concentrations in the third period declined more than in the second period, probably due to the large flow of people during the Spring Festival. The prevention and control measures of the COVID-19 pandemic are more stringent than the second period to prevent the large-scale recurrence of the epidemic similar to the Spring Festival in 2020, which might the explanations of these results.

Test on Parallel Trend Assumption

The DID model requires consistency in the development tendency over time between the experimental and the control group (Wan et al., 2019). Therefore, we tested the parallel trend hypothesis to ascertain if the pollutant concentration trends of the control and the treatment groups are parallel before the implementation of the intervention measure. Firstly, it can be seen that before the implementation in Section The Comparison of Pollutant Concentrations With 2019 (**Figure 4**), the trend of air pollutants in the control group (2019) and the treatment group (2021) was basically the same before the intervention. Consequently, our research may be able to pass the parallel trend hypothesis test. The residual diagram of the estimated coefficients of the model equation (3) was drawn to exhibit the policy effect better. **Figure 5** shows the time trend of residuals of air quality for 10 days before and following the Spring Festival after excluding

TABLE 6 | Regression results according to the DID model.

Variables	ln (AQI)	ln ($PM_{2.5}$)	ln (PM_{10})	ln (SO_2)	ln (NO_2)	ln (CO)	ln (O_3)
Diff (Covid)	0.1766***	−0.1521	−0.3035***	−0.4238***	−0.9396***	−0.3913***	0.1014**
	(0.0579)	(0.1093)	(0.1001)	(0.1243)	(0.1647)	(0.1114)	(0.0507)
Temperature	0.0663***	0.0466***	0.0849***	0.0059	0.043***	0.034***	−0.004
	(0.0047)	(0.0066)	(0.0063)	(0.0106)	(0.0067)	(0.0086)	(0.0056)
Humidity	0.0508***	0.0533***	0.035***	−0.0199	0.0294*	0.0431***	0.0105
	(0.008)	(0.0129)	(0.0108)	(0.0161)	(0.0157)	(0.0134)	(0.0163)
Wind speed	0.4257***	−0.1883	−0.0894	−0.5503***	−0.5591***	−0.0762	0.169
	(0.0835)	(0.1188)	(0.1136)	(0.1813)	(0.1542)	(0.1203)	(0.1336)
Precipitation	−10.0712***	−7.5783**	−8.4402***	3.8658	−8.0672*	−6.7001*	−4.1689
	(2.2774)	(3.6641)	(3.2534)	(4.5746)	(4.4533)	(3.6901)	(4.6013)
Constant	1.4409***	2.0778***	2.9998***	3.1405***	4.3461***	−1.4123**	3.3558***
	(0.3834)	(0.5771)	(0.5187)	(0.8059)	(0.7164)	(0.5752)	(0.6784)
Urban fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	640	640	633	640	640	640	640
R-squared	0.9447	0.9389	0.9022	0.7681	0.8872	0.8925	0.7866

*, **, *** *represent significance at the 10, 5, and 1% levels correspondingly.*

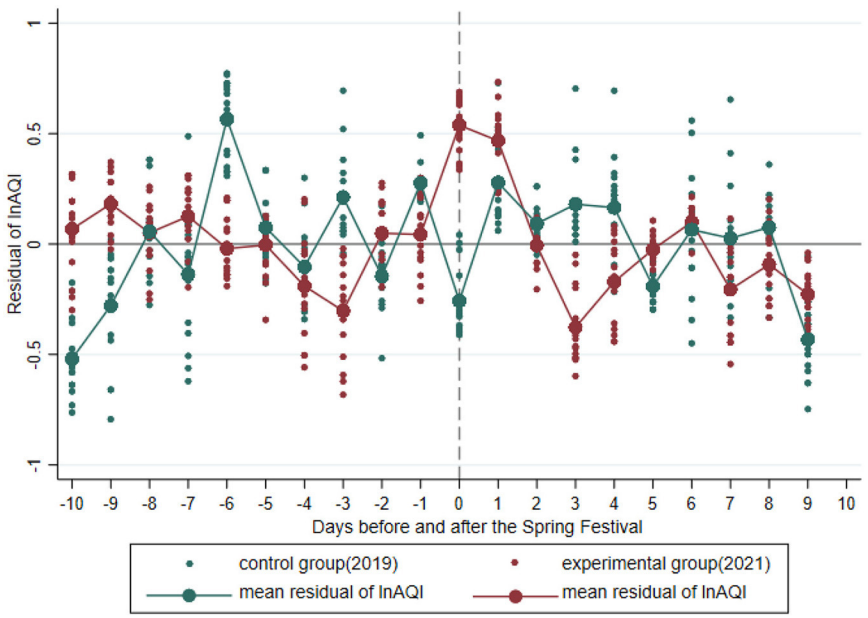


FIGURE 5 | Time trends 10 days prior to and after the Chinese New Year.

weather effects and other potential complicating variables. In both the control and treatment groups, the residual errors are standardized to zero, as can be observed (Li et al., 2017b).

We conducted further counterfactual research utilizing the model below (Equation 11) to examine the parallel trend assumption with greater rigor (Guo et al., 2020). Specifically, we included interaction terms between the grouping variable "Treat" and the temporal tendency of the 10 days before carrying out the restriction work to validate the parallel tendency of the 10 days before enacting the prevention efforts. The commencement of the intervention actions and the following 3 days were incorporated to avoid complete collinearity. The parallel tendency hypothesis is met if the interaction items of 10 days before implementing the restriction efforts exist no apparent impact on the independent variables.

$$\ln P_{it} = \gamma_0 + \sum_{d=-10}^{d=-1} \gamma_1 trend_{id} \times Treat + \sum_{d=0}^{d=2} \gamma_2 trend_{id} \times Treat + \gamma_3 Holiday + \theta_4 W_{it} + \mu_i + \pi_t + \varepsilon_{it} \quad (11)$$

TABLE 7 | Tests on parallel trends hypothesis adopting model 11.

Variables	ln (AQI)	ln ($PM_{2.5}$)	ln (PM_{10})	ln (SO_2)	ln (NO_2)	ln (CO)	ln (O_3)
day_−10	−0.895***	0.1361	0.4628	0.5495	−0.0406	−0.3404	−0.3477
	(0.244)	(0.4134)	(0.3408)	(0.4971)	(0.5079)	(0.4225)	(0.4821)
day_−9	0.3227***	0.112	0.5337***	−0.4218**	−0.2936*	0.0892	−0.076
	(0.089)	(0.1554)	(0.1323)	(0.1932)	(0.1754)	(0.1616)	(0.1173)
day_−8	−0.0489	−0.3701*	0.2559	−0.2365	−0.569**	−0.3428	0.0645
	(0.123)	(0.2232)	(0.1798)	(0.2398)	(0.2719)	(0.2406)	(0.2492)
day_−7	0.3074***	0.4467***	0.3931***	−0.007	−0.216	0.0824	−0.059
	(0.0841)	(0.1524)	(0.127)	(0.1626)	(0.1769)	(0.1462)	(0.1209)
day_−6	0.3103***	−0.1835	0.0079	−0.561**	−0.9439***	−0.2342	0.1187
	(0.1074)	(0.1775)	(0.1589)	(0.2266)	(0.2341)	(0.1839)	(0.1567)
day_−5	0.1355	−0.2747	0.2423*	−0.4357**	−1.3173***	−0.6369***	0.3181*
	(0.0916)	(0.1709)	(0.1383)	(0.1881)	(0.2238)	(0.166)	(0.1683)
day_−4	−0.1806	−0.6859***	−0.1758	−0.2721	−1.732***	−0.7923***	0.1419
	(0.1317)	(0.2324)	(0.2072)	(0.2457)	(0.282)	(0.239)	(0.25)
day_−3	−0.2265	−0.5171**	−0.3851*	−0.4362	−1.1808***	−0.5358**	0.0586
	(0.138)	(0.2406)	(0.199)	(0.2753)	(0.3001)	(0.2616)	(0.2674)
day_−2	0.0459	−0.3881	−0.2871	0.0154	−1.3744***	−0.6117*	0.1376
	(0.1729)	(0.3046)	(0.2502)	(0.3384)	(0.3835)	(0.3299)	(0.359)
day_−1	0.0151	−0.2905	−0.3306	0.1103	−1.4679***	−0.5499*	0.369
	(0.1682)	(0.2956)	(0.2422)	(0.3289)	(0.373)	(0.3114)	(0.3525)
day_0	0.0535	0.223	0.0835	1.1606***	−0.5815	−0.0927	0.1497
	(0.1947)	(0.3175)	(0.272)	(0.3876)	(0.3915)	(0.3146)	(0.4057)
day_1	−0.2204	0.1384	0.0369	1.3737***	−0.9596**	−0.3889	0.2732
	(0.2454)	(0.3923)	(0.3325)	(0.4921)	(0.4805)	(0.3824)	(0.4962)
day_2	−1.463***	−1.0606	−0.6776	1.3291	−1.3005	−1.0454	−1.1394
	(0.4561)	(0.7361)	(0.6188)	(0.9242)	(0.8973)	(0.7413)	(0.9322)
day_3	−1.8529***	−1.4064**	−1.6527***	0.5362	−1.5704*	−1.2538*	−0.7377
	(0.4242)	(0.683)	(0.6107)	(0.8507)	(0.8299)	(0.6868)	(0.8578)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Urban fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	640	640	633	640	640	640	640
R-squared	0.9447	0.9389	0.9022	0.7681	0.8872	0.8925	0.7866

*, **, *** *represent significance at the 10, 5, and 1% levels correspondingly.*

where d denotes the days since the beginning of the intervention policy, $trend_{id}$ indicates the time tendency, and γ_1 is a series of estimated coefficients for the 10 days before the preventative measures start, indicating the divergence in pollutant indexes between the control and treatment groups when making a comparison to the time before implementing the intervention policy.

Table 7 presents the estimated results of model 11. Most of the estimated results for the first 10 days before the prevention efforts are near zero. Besides, most of the coefficients are not statistically significant, as we can observe. In addition, **Figure 6** depicts the tendency of AQI estimation coefficients. We find that the coefficients are mostly close to zero before carrying out the measures, and that there is no obvious trend. The findings above indicate that the treatment and experimental groups have the same tendency before enacting the control measures (Zhang et al., 2020).

Robustness Tests

Further evidence is provided to prove the robustness of our empirical findings. First, we evaluate if our conclusions are still true if the width of the sample window changes. The initial sample window of our research includes 10 days before and after the Chinese New Year. We removed the head and tail for 1–3 days, respectively, and re-evaluated our model. **Table 8** show the findings of our research. The majority of the coefficients had similar orientations and magnitudes to our earlier results (i.e., coherent with **Table 6**).

Furthermore, considering that Chaoyang District, Haidian District, and Xicheng District intensified control efforts in the COVID-19 epidemic, we further confirmed whether the estimated results are robust by excluding Chaoyang District, Haidian District, and Xicheng District to prevent interference in economically developed areas. **Table 9** shows only the impact of the other 13 urban areas. All the estimated results are robust to this series of changes, which indicates that our research results

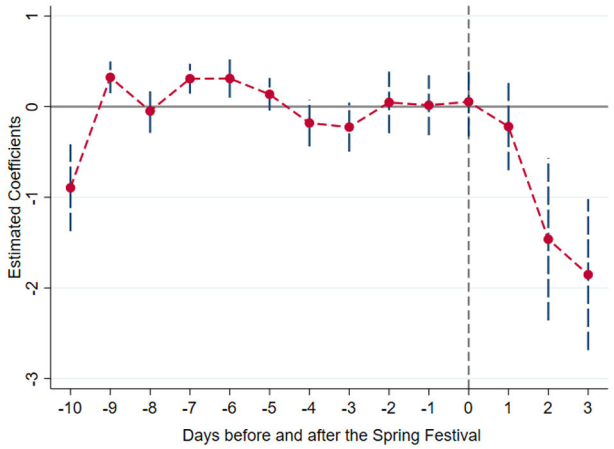


FIGURE 6 | Tests on parallel trends hypothesis. The vertical dashed gray line denotes the initiation of the policy of staying in place for Lunar New Year of 2021.

are not dominated by the super-developed urban areas most impacted by the pandemic.

$PM_{2.5}$ REDUCTION-INDUCED HEALTH BENEFIT EVALUATION

We estimated the health effects and health economic benefits of $PM_{2.5}$ reduction induced by COVID-19 epidemic control measures in three periods based on the results of previous studies. The estimation process adopts the $PM_{2.5}$ concentration, resident population, GDP data, and related calculation formulas. We utilized the secondary standard limit of the annual average value of $PM_{2.5}$ in China's Environmental Air Quality Standard (GB3095-2012) 35 $\mu g \cdot m^{-3}$. Specifically, we determined the $PM_{2.5}$ concentration in the absence of the COVID-19 epidemic (P_1) using the percentage decrease of $PM_{2.5}$ concentration (Z%) due to the COVID-19 epidemic and the actual $PM_{2.5}$ concentration data (P_2) in each district of Beijing: $P_1 = P_2/(1\text{-}Z\%)$.

First, the P_1 and exposed population data were substituted into the Poisson regression relative risk model (equations (4) and (5)), and the change of health effect (ΔE_1) when there was no COVID-19 epidemic could be estimated. The change of health effect (ΔE_2) when the COVID-19 occurred in each district of Beijing can be acquired through P_2 in the same way. Then, the difference in the change of health effect ($\Delta E_3 = \Delta E_1 - \Delta E_2$) can be obtained. The variation of health effect (ΔE_4) attributed to the $PM_{2.5}$ decrease owing to the restriction measures of COVID-19 outbreak can be obtained by dividing by the sum of days of the year (d) and multiplying by the number of days of each sample period (m):$\Delta E_4 = \Delta\ E_3 * m/d$.

Finally, ΔE_4 is brought into the model (9) to determine the monetary value of each health endpoint's economic benefits owing to the variation of $PM_{2.5}$ concentration employing the unit economic value of each health effect. This section mainly consists of two parts: estimation and analysis of the avoided health risk owing to $PM_{2.5}$ changes in Beijing during the initial epidemic outbreak in 2020, the epidemic outbreak in Beijing's Xinfadi market in 2020, and the policy period of staying in place for Lunar New Year of 2021. Besides, the health and economic benefit assessment results due to the $PM_{2.5}$ changes in Beijing during these three special periods are estimated and analyzed, as described below.

Avoided Health Risk Assessment

In this part, we evaluated the health effect of each health endpoint of $PM_{2.5}$ reduction caused by the COVID-19 epidemic. Then, we assessed the total health effects of 16 urban areas in Beijing by adding up the health effects of all health endpoints.

Health Endpoint Effects Assessment

We summarized the avoided health risks' estimation results of health endpoints in Beijing's 16 urban areas during the above three epidemic periods. Generally, the findings demonstrate that $PM_{2.5}$ reduction induced by restriction efforts has impacted each health endpoint, as observed in **Table 10**. Additionally, the degree of health endpoints affected by $PM_{2.5}$ pollution varies depending on $PM_{2.5}$ concentration, exposure population, outpatient incidence, and mortality in various urban areas. Similarly, the health effects of the same area are also various in distinct epidemic periods. Specifically, the health effects between different periods exist differences. It is estimated that the health effect of the $PM_{2.5}$ decrease induced by the outbreak in the Xinfadi market is the greatest. It has a more extended sample period. For this reason, there is a greater impact of $PM_{2.5}$ reduction than that of the other two periods, which are possible explanations. On the contrary, the policy of staying in place for the Lunar New Year in 2021 has a relatively short duration, so it brings a small influence due to $PM_{2.5}$ reduction. As a result, its health effect is smaller than the former two. Furthermore, the number of health beneficiaries in divergent regions is also heterogeneous. The three periods have the greatest impact on the health of Chaoyang District and Haidian District while having little effect on Mentougou, Huairou, and Yanqing areas. The top three health endpoints for the health benefits owing to $PM_{2.5}$ decrease are acute bronchitis, internal medicine clinic, and chronic bronchitis, accounting for about 80% of the total health effects. Cardiovascular diseases are relatively uncommon among hospitalized patients. Besides, the decline in $PM_{2.5}$ caused by control measures in the three epidemic periods probably avoided premature deaths of 1,117 cases [95% CI (328, 1,676)], 1,273 cases [95% CI (339, 2098)], and 115 cases [95% CI (33, 175)], respectively.

Total Health Effects Assessment

Table 11 shows the results of the three-period evaluation and ranking of the total health effects of $PM_{2.5}$ pollution changes in each of Beijing's urban areas. As observed, the total quantity of beneficiaries induced by the $PM_{2.5}$ decline resulting from COVID-19's epidemic control measures in the three periods are 35,968 cases [95% CI (17,238, 47,646)], 45,146 cases [95% CI (18,621, 67,876)], and 3,752 cases [95% CI (17,64, 55,035)],

TABLE 8 | Robustness test utilizing various sample windows.

Variables	ln (AQI)	ln ($PM_{2.5}$)	ln (PM_{10})	ln (SO_2)	ln (NO_2)	ln (CO)	ln (O_3)
	Panel C: 10 days before and after the spring festival						
Diff (Covid)	0.1766***	−0.1521	−0.3035***	−0.4238***	−0.9396***	−0.3913***	0.1014**
	(0.0579)	(0.1093)	(0.1001)	(0.1243)	(0.1647)	(0.1114)	(0.0507)
	Panel D: 9 days before and after the spring festival						
Diff (Covid)	0.1349	−0.2*	−0.7184***	−0.8935***	−0.7955***	−0.4306***	0.3091***
	(0.0875)	(0.1135)	(0.1327)	(0.1932)	(0.209)	(0.1313)	(0.0865)
	Panel E: 8 days before and after the spring festival						
Diff (Covid)	−0.5295***	−1.7599***	−1.1156***	−1.8004***	−1.5196***	−1.3205***	0.1807**
	(0.1191)	(0.1915)	(0.1814)	(0.2801)	(0.2083)	(0.1499)	(0.087)
	Panel F: 7 days before and after the spring festival						
Diff (Covid)	−0.6264***	−1.7269***	−0.4495***	−1.1221***	−1.058***	−1.1542***	0.3554***
	(0.1149)	(0.2015)	(0.1576)	(0.2217)	(0.1337)	(0.1204)	(0.0663)

*, **, *** *represent significance at the 10, 5, and 1% levels correspondingly.*

TABLE 9 | Robustness test of 13 urbans, excluding Chaoyang, Haidian, and Xicheng.

Variables	ln (AQI)	ln ($PM_{2.5}$)	ln (PM_{10})	ln (SO_2)	ln (NO_2)	ln (CO)	ln (O_3)
Diff (Covid)	0.1441**	−0.1627	−0.3032**	−0.4151***	−1.0631***	−0.4677***	0.0927
	(0.0641)	(0.1333)	(0.1183)	(0.1435)	(0.1909)	(0.1333)	(0.058)
Temperature	0.0655***	0.0479***	0.0866***	0.0134	0.0375***	0.0268***	−0.0005
	(0.0052)	(0.0078)	(0.0078)	(0.0123)	(0.0078)	(0.0095)	(0.0065)
Humidity	0.0525***	0.0582***	0.042***	−0.0088	0.0247	0.0452***	0.0145
	(0.0092)	(0.0158)	(0.0125)	(0.0183)	(0.0181)	(0.0149)	(0.0191)
Wind speed	0.4547***	−0.1362	−0.0195	−0.4293**	−0.6584***	−0.0835	0.1917
	(0.0976)	(0.1424)	(0.1331)	(0.2087)	(0.1775)	(0.1378)	(0.1575)
Precipitation	−10.3249***	−8.7487*	−11.4966***	0.7674	−6.3787	−6.9345*	−5.3797
	(2.6346)	(4.4574)	(3.7204)	(5.1923)	(5.1419)	(4.1308)	(5.424)
Constant	1.3622***	1.8495***	2.6752***	2.5886***	4.7425***	−1.4029**	3.2085***
	(0.4459)	(0.6966)	(0.605)	(0.9244)	(0.8281)	(0.6528)	(0.7999)
Urban fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	520	520	514	520	520	520	520
R-squared	0.9397	0.9322	0.898	0.7499	0.8752	0.8816	0.7567

*, **, *** *indicate significance at the 10, 5, and 1% levels, respectively.*

respectively, and this is the total number of avoided premature death, respiratory disease hospitalization, cardiovascular disease hospitalization, internal medicine clinic, pediatrics clinic, acute bronchitis, chronic bronchitis, and asthma. In terms of different urban areas, the total number of health beneficiaries brought about by $PM_{2.5}$ pollution changes is highest in Chaoyang District and lowest in Huairou, Miyun, Yanqing, and Mentougou. For example, at the beginning of the outbreak in 2020, the total health benefits of Chaoyang reached 5,598 cases [95% CI (2,657, 7,462)]. The second is Haidian with 5,213 cases [95% CI (2,508, 6,879)], followed by Daxing with 3,520 cases [95% CI (1,753, 4,530)], and Yanqing with 532 cases [95% CI (246, 723)]. The main reasons are as follows: on the one hand, the $PM_{2.5}$ concentrations are higher in urban areas with more beneficiaries. On the other hand, the exposed population in these urban areas is relatively large. For example, the exposed population of Chaoyang district in 2018 is about nine times that of Huairou. The resident distribution is fairly dense and Huairou's per capita occupation area (5,128 m^2) is about 39 times that of Chaoyang (131 m^2).

Avoided Economic Loss Evaluation

In this part, we assessed the health and economic benefit of various health endpoints and summarized the total health endpoint benefits of $PM_{2.5}$ reduction due to the COVID-19 epidemic in various urban of Beijing.

Economic Benefits Evaluation of Health Endpoints

This section uses the methods in Section PM2.5 Reduction-Induced Health Effects and Economic Benefits to estimate the initial outbreak of COVID-19 in 2020, the outbreak of the Xinfadi market in Beijing, and the policy period of celebrating the Chinese New Year 2021 in place based on the evaluation

TABLE 10 | Evaluated amount of health effects induced by $PM_{2.5}$ reduction caused by the COVID-19 epidemic of 16 districts of Beijing in three periods ($C_0 = 35\ \mu g/m^3$).

Period	Districts	Death	In hospital		Outpatient service		Diseases			Total estimated number
			The respiratory system	Cardiovascular	Department of pediatrics (≤14 years old)	Internal medicine (>15 years old)	Acute bronchial	Chronic bronchitis	Asthma	
The initial outbreak of COVID-19 in 2020	Dongcheng	34 (10, 52)	36 (0, 68)	12 (8, 16)	103 (37, 162)	278 (155, 392)	503 (227, 634)	107 (54, 124)	46 (33, 58)	1118 (524, 1507)
	Xicheng	57 (17, 85)	60 (0, 114)	20 (13, 27)	174 (64, 274)	472 (263, 665)	805 (378, 974)	168 (90, 187)	77 (55, 97)	1833 (880, 2422)
	Chaoyang	173 (50, 261)	181 (0, 345)	60 (39, 81)	524 (191, 825)	1417 (791, 2000)	2487 (1145, 3064)	523 (274, 593)	233 (167, 293)	5598 (2657, 7462)
	Haidian	162 (48, 243)	172 (0, 325)	57 (37, 77)	497 (181, 782)	1345 (752, 1897)	2284 (1076, 2753)	476 (256, 527)	220 (158, 275)	5213 (2508, 6879)
	Fengtai	108 (32, 160)	115 (0, 217)	38 (25, 52)	334 (122, 525)	905 (506, 1275)	1487 (717, 1753)	307 (170, 332)	147 (106, 183)	3442 (1678, 4497)
	Shijingshan	30 (9, 44)	32 (0, 60)	11 (7, 14)	92 (34, 145)	249 (139, 351)	416 (198, 496)	86 (47, 94)	41 (29, 51)	956 (463, 1256)
	Mentougou	19 (5, 29)	20 (0, 38)	7 (4, 9)	57 (21, 90)	155 (86, 219)	280 (126, 351)	59 (30, 69)	26 (18, 32)	622 (292, 837)
	Fangshan	76 (23, 109)	82 (0, 153)	27 (18, 37)	241 (88, 376)	652 (365, 916)	983 (503, 1094)	198 (118, 201)	104 (75, 129)	2363 (1191, 3016)
	Tongzhou	96 (28, 144)	102 (0, 193)	34 (22, 46)	296 (108, 465)	800 (447, 1128)	1346 (638, 1611)	280 (152, 307)	130 (94, 163)	3084 (1489, 4057)
	Shunyi	63 (18, 96)	65 (0, 124)	22 (14, 29)	187 (68, 295)	505 (282, 714)	932 (415, 1188)	199 (100, 234)	84 (60, 106)	2057 (956, 2787)
	Daxing	112 (34, 163)	121 (0, 226)	40 (26, 54)	352 (129, 552)	954 (534, 1342)	1486 (744, 1689)	302 (175, 314)	153 (111, 190)	3520 (1753, 4530)
	Changping	106 (30, 163)	110 (0, 210)	36 (23, 49)	315 (115, 498)	852 (475, 1205)	1589 (701, 2040)	340 (169, 403)	142 (101, 179)	3491 (1615, 4747)
	Pinggu	25 (7, 36)	26 (0, 49)	9 (6, 12)	76 (28, 119)	206 (115, 290)	337 (163, 396)	69 (39, 75)	33 (24, 42)	781 (381, 1020)
	Huairou	18 (5, 29)	18 (0, 36)	6 (4, 8)	52 (19, 83)	142 (79, 201)	290 (120, 397)	64 (29, 81)	24 (17, 31)	615 (273, 865)
	Miyun	22 (6, 35)	22 (0, 43)	7 (5, 10)	64 (23, 101)	173 (96, 245)	350 (146, 475)	77 (35, 96)	29 (21, 37)	744 (332, 1043)
	Yanqing	16 (5, 25)	17 (0, 32)	6 (4, 7)	48 (17, 76)	130 (72, 183)	242 (107, 311)	52 (26, 61)	22 (15, 27)	532 (246, 723)
	Beijing	1117 (328, 1676)	1179 (0, 2233)	391 (251, 527)	3412 (1246, 5371)	9235 (5158, 13024)	15817 (7404, 19227)	3307 (1766, 3696)	1509 (1085, 1894)	35968 (17238, 47646)
The outbreak in the Xinfadi Market in 2020	Dongcheng	45 (12, 74)	43 (0, 86)	14 (9, 19)	121 (44, 194)	327 (181, 466)	796 (292, 1226)	184 (73, 265)	58 (40, 75)	1589 (650, 2405)
	Xicheng	82 (22, 133)	80 (0, 159)	26 (17, 36)	227 (82, 363)	613 (339, 872)	1409 (537, 2089)	321 (133, 444)	107 (75, 138)	2865 (1205, 4233)
	Chaoyang	250 (67, 409)	244 (0, 485)	80 (51, 109)	691 (249, 1104)	1866 (1033, 2655)	4334 (1640, 6471)	990 (407, 1379)	326 (228, 420)	8782 (3674, 13030)
	Haidian	228 (61, 373)	223 (0, 443)	73 (46, 99)	633 (228, 1010)	1708 (945, 2429)	3953 (1499, 5888)	902 (372, 1253)	298 (209, 384)	8018 (3360, 11880)
	Fengtai	95 (25, 159)	91 (0, 183)	30 (19, 40)	257 (92, 411)	692 (382, 987)	1727 (622, 2704)	403 (156, 590)	123 (86, 159)	3417 (1381, 5233)
	Shijingshan	39 (10, 64)	38 (0, 75)	12 (8, 17)	107 (38, 170)	287 (159, 409)	685 (255, 1041)	158 (63, 224)	51 (35, 65)	1376 (569, 2066)
	Mentougou	7 (2, 12)	7 (0, 14)	2 (2, 3)	20 (7, 31)	53 (29, 75)	137 (48, 221)	32 (12, 49)	10 (7, 12)	268 (106, 419)
	Fangshan	67 (18, 111)	64 (0, 128)	21 (13, 29)	181 (65, 289)	487 (269, 694)	1207 (436, 1882)	281 (109, 410)	86 (60, 112)	2393 (970, 3655)
	Tongzhou	133 (36, 218)	130 (0, 259)	43 (27, 58)	369 (133, 589)	996 (551, 1416)	2312 (875, 3450)	528 (217, 735)	174 (122, 224)	4684 (1960, 6949)
	Shunyi	78 (21, 129)	75 (0, 150)	24 (16, 33)	211 (76, 338)	570 (315, 812)	1396 (509, 2159)	324 (127, 468)	101 (70, 131)	2779 (1133, 4220)
	Daxing	87 (23, 146)	84 (0, 168)	27 (17, 37)	235 (84, 377)	634 (350, 903)	1589 (570, 2497)	371 (143, 546)	113 (78, 146)	3140 (1266, 4819)
	Changping	99 (26, 165)	95 (0, 190)	31 (19, 42)	266 (95, 426)	717 (396, 1021)	1798 (644, 2825)	420 (162, 618)	127 (88, 165)	3551 (1431, 5452)
	Pinggu	18 (5, 30)	17 (0, 34)	6 (4, 8)	48 (17, 77)	130 (72, 185)	327 (117, 517)	77 (29, 113)	23 (16, 30)	646 (259, 994)
	Huairou	21 (6, 36)	20 (0, 41)	7 (4, 9)	57 (21, 92)	155 (86, 221)	386 (139, 603)	90 (35, 132)	28 (19, 36)	764 (309, 1169)
	Miyun	12 (3, 20)	12 (0, 23)	4 (2, 5)	32 (12, 52)	87 (48, 124)	225 (79, 362)	53 (20, 80)	16 (11, 20)	441 (175, 687)
	Yanqing	12 (3, 20)	11 (0, 23)	4 (2, 5)	32 (12, 51)	86 (48, 123)	219 (78, 347)	51 (20, 76)	15 (11, 20)	431 (173, 665)
	Beijing	1273 (339, 2098)	1235 (0, 2459)	403 (256, 549)	3487 (1253, 5574)	9407 (5202, 13393)	22500 (8338, 34283)	5186 (2078, 7383)	1655 (1155, 2137)	45146 (18621, 67876)

(Continued)

TABLE 10 | Continued

Period	Districts	Death	In hospital		Outpatient service		Diseases			Total estimated number
			The respiratory system	Cardiovascular	Department of pediatrics (≤14 years old)	Internal medicine (>15 years old)	Acute bronchial	Chronic bronchitis	Asthma	
The policy of staying put during Spring Festival in 2021	Dongcheng	4 (1, 6)	4 (0, 8)	1 (1, 2)	12 (4, 19)	32 (18, 45)	55 (26, 68)	12 (6, 13)	5 (4, 7)	125 (60, 166)
	Xicheng	6 (2, 9)	6 (0, 12)	2 (1, 3)	18 (7, 29)	49 (27, 70)	86 (40, 106)	18 (10, 20)	8 (6, 10)	194 (92, 258)
	Chaoyang	18 (5, 28)	19 (0, 37)	6 (4, 9)	56 (20, 88)	151 (84, 213)	268 (122, 332)	56 (29, 64)	25 (18, 31)	599 (283, 801)
	Haidian	17 (5, 25)	17 (0, 33)	6 (4, 8)	50 (18, 79)	136 (76, 192)	242 (110, 301)	51 (26, 58)	22 (16, 28)	542 (255, 725)
	Fengtai	11 (3, 16)	11 (0, 22)	4 (2, 5)	33 (12, 52)	89 (50, 126)	157 (72, 194)	33 (17, 37)	15 (10, 18)	353 (167, 470)
	Shijingshan	3 (1, 5)	3 (0, 6)	1 (1, 1)	9 (3, 14)	25 (14, 35)	44 (20, 55)	9 (5, 11)	4 (3, 5)	98 (46, 131)
	Mentougou	2 (1, 3)	2 (0, 4)	1 (0, 1)	5 (2, 9)	15 (8, 21)	29 (12, 38)	6 (3, 8)	2 (2, 3)	62 (28, 86)
	Fangshan	6 (2, 10)	6 (0, 12)	2 (1, 3)	18 (7, 29)	49 (27, 69)	96 (41, 128)	21 (10, 26)	8 (6, 10)	207 (93, 286)
	Tongzhou	10 (3, 15)	11 (0, 21)	4 (2, 5)	32 (12, 50)	86 (48, 121)	145 (68, 174)	30 (16, 33)	14 (10, 18)	331 (160, 436)
	Shunyi	7 (2, 11)	8 (0, 14)	3 (2, 3)	22 (8, 34)	59 (33, 83)	103 (48, 127)	22 (11, 24)	10 (7, 12)	233 (111, 310)
	Daxing	11 (3, 16)	11 (0, 21)	4 (2, 5)	33 (12, 51)	88 (49, 125)	155 (71, 191)	33 (17, 37)	15 (10, 18)	349 (166, 465)
	Changping	11 (3, 17)	11 (0, 21)	4 (2, 5)	31 (11, 49)	84 (47, 119)	166 (70, 221)	36 (17, 44)	14 (10, 18)	356 (160, 494)
	Pinggu	3 (1, 4)	3 (0, 5)	1 (1, 1)	8 (3, 13)	22 (12, 31)	36 (18, 43)	8 (4, 8)	4 (3, 5)	84 (41, 110)
	Huairou	2 (1, 3)	2 (0, 4)	1 (0, 1)	7 (2, 10)	18 (10, 25)	33 (15, 43)	7 (4, 8)	3 (2, 4)	73 (33, 99)
	Miyun	3 (1, 4)	3 (0, 6)	1 (1, 1)	9 (3, 14)	24 (13, 34)	41 (19, 50)	9 (5, 10)	4 (3, 5)	93 (45, 124)
	Yanqing	2 (0, 2)	2 (0, 3)	1 (0, 1)	4 (2, 7)	12 (7, 17)	25 (10, 34)	5 (2, 7)	2 (1, 3)	52 (23, 73)
	Beijing	115 (33, 175)	120 (0, 229)	40 (26, 54)	346 (126, 546)	937 (522, 1323)	1682 (762, 2103)	356 (183, 409)	155 (111, 195)	3752 (1764, 5035)

It's the value of 95% CI in parentheses; the actual average annual concentration of $PM_{2.5}$ is less than or equal to 35 $\mu g/m^3$, and its health effect is 0.

results of health effects in **Table 10** and the unit economic value of each health endpoint in **Table 2**. **Table 12** shows the estimations of the health endpoint benefits of residents under the condition of $PM_{2.5}$ reduction induced by prevention efforts. Generally, the corresponding economic benefits of health endpoints in Beijing and its 16 urban areas have the same characteristics as the number of health effects in terms of different epidemic periods. Additionally, the economic benefits of various health endpoints caused by the $PM_{2.5}$ reduction are diverse due to differences in $PM_{2.5}$ concentration, benchmark incidence, economic value per unit of health endpoints, and exposure population. For example, the health and economic benefit of the $PM_{2.5}$ decrease in Beijing due to early death was 4,183.79 million yuan [95% CI (1,228.81, 6,278.53)]. In comparison, the health and economic benefit of hospitalization was 29.47 million yuan [95% CI (8.73, 48.38)] during the initial COVID outbreak in 2020. Furthermore, the health benefit of chronic bronchitis was the greatest, followed by early death, acute bronchitis, asthma, and hospitalization, while the health benefit of the outpatient clinic was the smallest.

Total Economic Benefits Evaluation

Table 13 summarizes the health and economic benefits of $PM_{2.5}$ pollution changes caused by three-period epidemic prevention and control measures in Beijing and various regions. The total health and economic benefits induced by changes in $PM_{2.5}$ pollution in Beijing and various districts over three periods are heterogeneous according to the assessed results in **Table 13**. Additionally, the per capita health economic benefits and total health benefits of various districts in Beijing are dissimilar or even quite divergent in the three periods, mainly due to divergences in resident population and $PM_{2.5}$ concentration changes among Beijing's urban areas. Specifically, the avoided total health and economic loss induced by the $PM_{2.5}$ reductions owing to the control measures in three epidemic periods are 82,747.65 million yuan [95% CI (3,406.4,10,879.1)], 11,143.71 million yuan [95% CI (3,826.43,16,949.1)], and 871.65 million yuan [95% CI (350.54, 1,165.95)], which accounted for 0.23% [95% CI (0.09%, 0.3%)], 0.31% [95% CI (0.11%, 0.47%),] 0.02% [95% CI (0.1%, 0.03%)] of the GDP of Beijing that year, respectively. Furthermore, Chaoyang, Haidian, Tongzhou, and Fengtai are still the areas that benefit the most from the changes in $PM_{2.5}$ concentration in terms of health and economy. For instance, the health and economic benefits of the four urban areas were 2,154.15 million yuan [95% CI (751.53, 3,229.36)], 1,965.49 million yuan [95% CI (687.41, 2,942.39)], 1,148.71 million yuan [95% CI (400.8, 1,722.54)], 850.61 million yuan [95% CI (285.52, 1,319.71)], while the health and economic benefits of Yanqing, Mentougou, Huairou, and other districts are relatively low during the outbreak of the Beijing Xinfadi market.

DISCUSSIONS ON RECOMMENDATIONS FOR METICULOUS CONTROL OF AIR POLLUTION

China has enacted various stringent controls to prevent the epidemic of COVID-19 from spreading, such as closing factories and restricting traffic. However, Beijing still suffers from severe air pollution in extreme weather. People hold divergent perspectives on how to prevent and control air pollution effectively. Rigorous empirical research is required to quantify the causal impact of control actions on air pollution in the COVID-19 epidemic. Firstly, the methods and findings of this research provide a valuable natural experiment for exploring the causal effects of blockade efforts on air pollution. In our study, real-time observation data are used to demonstrate the changing air quality trend before and after the outbreak of COVID-19. Subsequently, LSDV and DID models are constructed to determine the causal impact of epidemic restriction efforts on air pollution and keep weather, holidays, and other influential factors from interfering. The execution of regulatory efforts improved air quality and provided empirical evidence for ascertaining the causal relation of blockade actions to air pollution in the COVID-19 pandemic. Furthermore, we could carry out targeted control of air pollution by comparing the influence of the pandemic on air pollutants in various periods and analyzing the causes of the distinction. What deserves our attention is that adopting lower-scale data is beneficial for the meticulous prevention, control, and governance of air pollution.

The study's two conclusions are critical for future air pollution mitigation. Firstly, traffic restrictions are crucial for controlling NO_2 pollution in densely populated urban areas, which indicates an essential direction for NO_2 pollution prevention. The management of traffic activities should focus on NO_2 pollution control in the future. With the gradual lifting of traffic restrictions at the end of February 2020, epidemic prevention actions for public transport, such as buses and subways, continue to be strict. This may lead to an increase in the utilization of personal automobiles (Lee et al., 2020). Therefore, calling on urban daily travel to return to green safety is the government's primary concern. In this case, walking and cycling are advisable for personal travel. The infrastructure for non-motorized transportation, such as walking and cycling, is supposed to be optimized regularly, which will have dual benefits for pandemic control and air pollution mitigation in the long run. Additionally, the index system of travel intensity in the city can be established to accomplish precise management and control through the real-time observation of the big data system. Meantime, proper management, and motivations can be constituted to lower the strength of urban traffic, such as tail number restrictions, rising oil prices, public transportation subsidies, and others.

Secondly, the health and economic benefits of changes in $PM_{2.5}$ concentrations are quantified in this study. The health effects were related to the exposure population, the change in $PM_{2.5}$ concentration, the standard concentration threshold, the exposure-response relationship coefficient, and the baseline incidence. For each urban area of Beijing, the concentration of $PM_{2.5}$ in the urban areas with more health benefits is higher, and the base of exposed population in these urban areas is larger. For example, the exposed population in Chaoyang District in 2018 is about nine times that of Huairou, and the distribution of residents is relatively dense. Therefore, the control of $PM_{2.5}$ pollution

TABLE 11 | Health effects and ranking of $PM_{2.5}$ reduction caused by the COVID-19 epidemic of 16 districts of Beijing.

Districts	The initial outbreak of COVID-19 in 2020		The outbreak in the Xinfadi market in 2020		The policy of staying put during the Spring Festival in 2021	
	Total health effect	Order	Total health effect	Order	Total health effect	Order
Dongcheng	1,118 (524, 1,507)	10	1,589 (650, 2,405)	10	125 (60, 166)	10
Xicheng	1,833 (880, 2,422)	9	2,865 (1,205, 4,233)	7	194 (92, 258)	9
Chaoyang	5,598 (2,657, 7,462)	1	8,782 (3,674, 13,030)	1	599 (283, 801)	1
Haidian	5,213 (2,508, 6,879)	2	8,018 (3,360, 11,880)	2	542 (255, 725)	2
Fengtai	3,442 (1,678, 4,497)	5	3,417 (1,381, 5,233)	5	353 (167, 470)	4
Shijingshan	956 (463, 1,256)	11	1,376 (569, 2,066)	11	98 (46, 131)	11
Mentougou	622 (292, 837)	14	268 (106, 419)	16	62 (28, 86)	15
Fangshan	2,363 (1,191, 3,016)	7	2,393 (970, 3,655)	9	207 (93, 286)	8
Tongzhou	3,084 (1,489, 4,057)	6	4,684 (1,960, 6,949)	3	331 (160, 436)	6
Shunyi	2,057 (956, 2,787)	8	2,779 (1,133, 4,220)	8	233 (111, 310)	7
Daxing	3,520 (1,753, 4,530)	3	3,140 (1,266, 4,819)	6	349 (166, 465)	5
Changping	3,491 (1,615, 4,747)	4	3,551 (1,431, 5,452)	4	356 (160, 494)	3
Pinggu	781 (381, 1,020)	12	646 (259, 994)	13	84 (41, 110)	13
Huairou	615 (273, 865)	15	764 (309, 1,169)	12	73 (33, 99)	14
Miyun	744 (332, 1,043)	13	441 (175, 687)	14	93 (45, 124)	12
Yanqing	532 (246, 723)	16	431 (173, 665)	15	52 (23, 73)	16
Beijing	35,968 (17,238, 47,646)	—	45,146 (18,621, 67,876)	—	3,752 (1,764, 5,035)	—

It's the value of 95% CI in parentheses.

concentration and the health and safety protection of residents in these urban areas will undoubtedly have great potential for health effects in the future. The health effects and health economic benefits owing to the $PM_{2.5}$ decrease, along with the corresponding 95% confidence interval, provide a reference value for Beijing to meet the air quality standards, control severe ambient air pollution, and implement health early warning system.

Finally, the experimental importance of control efforts in reducing severe air pollution is also discussed in our study. The COVID-19 pandemic is a special public health event as it is particular and uncertain, providing an exceedingly unusual natural experiment for controlling the social and economic activities that impact air quality. Large-scale pollution outbreaks, such as haze pollution in the winter, have occurred in northern China, particularly in the Beijing area. Extreme air pollution is harmful to people's health and affects the regular operation of the social economy, particularly in regions with high population density and developed economies (Archer-Nicholls et al., 2016; Vu et al., 2019). Consequently, high-intensity limitations on population, transportation, and economic activities should be adopted to lower their damage and cope with serious urban pollution. For instance, in the event of an extreme air pollution incident, the remote work and online teaching system emerging during COVID-19's blockade will be immediately employed. This study provides significant evidence for the development of contingency designs for comprehensive socio-economic governance strategies in cases of severe pollution.

CONCLUSIONS

The COVID-19 epidemic has been around for a long time. However, it is not yet over and is entering the normalization stage. Outbreaks still occur in different regions. Therefore, it is still worthy of further study and remains an important issue to determine the causal impact of restriction efforts on air quality in the ecological environment. Here, by establishing LSDV and DID models, we quantitatively determine and compare the causal impact of control and blockade measures on air quality across three significant periods in China. We find that restriction measures have a significant positive influence on improving the air pollution in the ecological environment, and the effects of the three periods are different. As expected, during the initial outbreak in 2020, except for the increase in O_3 concentration, the execution of control actions decreased the AQI by 7.8% and NO_2, and SO_2 by 46.76% and 37.32%. At the same time, the level of $PM_{2.5}$, PM_{10}, and CO had different degrees of decline, which fell by 34.07, 53.22, and 19.97%, respectively. Additionally, during the outbreak of the Xinfadi market in Beijing, the air pollutant concentrations, including O_3, decreased significantly, of which O_3 decreased by 7.26% and AQI and $PM_{2.5}$ dropped by 22.61 and 45.12%, respectively, compared with the first period. During the policy period of staying in place for the Lunar New Year of 2021, except for slight decline in AQI and O_3 levels, other pollutants decreased compared with the same period in 2019, among which NO2 decreased most obviously.

$PM_{2.5}$ concentrations in the ecological environment beyond a particular threshold will raise the risk of cardiovascular and

TABLE 12 | Avoided economic loss of the health effects induced by $PM_{2.5}$ reduction caused by the COVID-19 epidemic of 16 districts of Beijing in three periods ($C_0 = 35\ \mu g/m^3$, unit: million yuan).

Period	Districts	Death	In hospital		Outpatient service		Diseases			Total economic loss (million yuan)
			The respiratory system	Cardiovascular	Department of pediatrics (≤14 years old)	Internal medicine (>15 years old)	Acute bronchial	Chronic bronchitis	Asthma	
The initial COVID-19 outbreak in 2020	**Dongcheng**	128.41 (37.12,195.84)	0.48 (0,0.92)	0.41 (0.26,0.55)	0.08 (0.03,0.13)	0.22 (0.13,0.32)	1.97 (0.89,2.48)	128.24 (65.33,148.6)	0.74 (0.53,0.93)	260.55 (104.29,349.77)
	Xicheng	213.64 (62.5,320.08)	0.81 (0,1.53)	0.69 (0.45,0.93)	0.14 (0.05,0.22)	0.38 (0.21,0.54)	3.15 (1.48,3.81)	201.33 (108,223.75)	1.24 (0.89,1.55)	421.38 (173.58,552.43)
	Chaoyang	647.74 (189.02,978.81)	2.44 (0,4.64)	2.09 (1.34,2.81)	0.42 (0.15,0.67)	1.14 (0.64,1.62)	9.74 (4.48,12)	627.14 (328.6,710.54)	3.73 (2.68,4.7)	1294.46 (526.91,1715.78)
	Haidian	607.97 (178.79,910.62)	2.31 (0,4.37)	1.98 (1.28,2.67)	0.4 (0.15,0.63)	1.09 (0.61,1.53)	8.95 (4.21,10.78)	570.66 (307.51,631.38)	3.52 (2.53,4.41)	1196.87 (495.08,1566.41)
	Fengtai	404.55 (120.46,600.01)	1.55 (0,2.92)	1.33 (0.86,1.79)	0.27 (0.1,0.42)	0.73 (0.41,1.03)	5.82 (2.81,6.87)	367.99 (203.88,397.57)	2.35 (1.7,2.94)	784.61 (330.2,1013.55)
	Shijingshan	112.12 (33.33,166.67)	0.43 (0,0.81)	0.37 (0.24,0.5)	0.07 (0.03,0.12)	0.2 (0.11,0.28)	1.63 (0.78,1.94)	103.51 (56.61,113.09)	0.65 (0.47,0.81)	218.99 (91.56,284.22)
	Mentougou	71.21 (20.45,108.71)	0.27 (0,0.51)	0.23 (0.15,0.31)	0.05 (0.02,0.07)	0.13 (0.07,0.18)	1.09 (0.49,1.37)	71.15 (36.36,82.18)	0.41 (0.29,0.52)	144.54 (57.84,193.85)
	Fangshan	282.96 (86.37,410.24)	1.11 (0,2.06)	0.96 (0.62,1.29)	0.19 (0.07,0.3)	0.53 (0.29,0.74)	3.85 (1.97,4.29)	237.33 (141.57,241.09)	1.66 (1.21,2.06)	528.59 (232.1,662.06)
	Tongzhou	360.61 (106.44,538.27)	1.37 (0,2.59)	1.18 (0.76,1.58)	0.24 (0.09,0.38)	0.65 (0.36,0.91)	5.27 (2.5,6.31)	335.27 (182.06,368.24)	2.09 (1.5,2.62)	706.68 (293.7,920.9)
	Shunyi	234.85 (67.43,359.86)	0.87 (0,1.67)	0.75 (0.48,1)	0.15 (0.05,0.24)	0.41 (0.23,0.58)	3.65 (1.62,4.65)	238.54 (119.76,280.24)	1.35 (0.96,1.7)	480.57 (190.53,649.94)
	Daxing	418.95 (126.52,612.13)	1.63 (0,3.04)	1.4 (0.9,1.88)	0.28 (0.1,0.45)	0.77 (0.43,1.08)	5.82 (2.91,6.61)	362.06 (210.06,376.24)	2.45 (1.77,3.05)	793.36 (342.7,1004.48)
	Changping	397.73 (113.64,610.62)	1.48 (0,2.83)	1.26 (0.8,1.7)	0.25 (0.09,0.4)	0.69 (0.38,0.97)	6.22 (2.75,7.99)	407.87 (202.78,482.9)	2.28 (1.62,2.88)	817.78 (322.07,1110.29)
	Pinggu	92.05 (27.27,135.99)	0.35 (0,0.66)	0.3 (0.19,0.41)	0.06 (0.02,0.1)	0.17 (0.09,0.23)	1.32 (0.64,1.55)	83.27 (46.3,89.82)	0.53 (0.39,0.67)	178.06 (74.91,229.42)
	Huairou	68.18 (18.94,107.58)	0.25 (0,0.48)	0.21 (0.13,0.28)	0.04 (0.02,0.07)	0.11 (0.06,0.16)	1.14 (0.47,1.55)	76.48 (35.03,96.85)	0.39 (0.27,0.49)	146.8 (54.92,207.47)
	Miyun	82.58 (23.11,130.31)	0.3 (0,0.59)	0.25 (0.16,0.34)	0.05 (0.02,0.08)	0.14 (0.08,0.2)	1.37 (0.57,1.86)	92 (42.54,115.63)	0.47 (0.33,0.6)	177.16 (66.81,249.61)
	Yanqing	60.23 (17.42,92.8)	0.22 (0,0.43)	0.19 (0.12,0.26)	0.04 (0.01,0.06)	0.1 (0.06,0.15)	0.95 (0.42,1.22)	62.18 (30.91,73.57)	0.35 (0.25,0.44)	124.27 (49.2,168.93)
	Beijing	4183.79 (1228.81,6278.53)	15.88 (0,30.07)	13.59 (8.73,18.31)	2.76 (1.01,4.34)	7.46 (4.16,10.52)	61.94 (28.99,75.29)	3965.03 (2117.3,4431.69)	24.2 (17.39,30.36)	8274.65 (3406.4,10879.1)
The outbreak in the Xinfadi market in 2020	**Dongcheng**	45 (12, 74)	43 (0, 86)	14 (9, 19)	121 (44, 194)	327 (181, 466)	796 (292, 1226)	184 (73, 265)	58 (40, 75)	1589 (650, 2405)
	Xicheng	82 (22, 133)	80 (0, 159)	26 (17, 36)	227 (82, 363)	613 (339, 872)	1409 (537, 2089)	321 (133, 444)	107 (75, 138)	2865 (1205, 4233)
	Chaoyang	250 (67, 409)	244 (0, 485)	80 (51, 109)	691 (249, 1104)	1866 (1033, 2655)	4334 (1640, 6471)	990 (407, 1379)	326 (228, 420)	8782 (3674, 13030)

(Continued)

TABLE 12 | Continued

Period	Districts	Death	In hospital		Outpatient service		Diseases			Total economic loss (million yuan)
			The respiratory system	Cardiovascular	Department of pediatrics (≤14 years old)	Internal medicine (>15 years old)	Acute bronchial	Chronic bronchitis	Asthma	
	Haidian	228 (61, 373)	223 (0, 443)	73 (46, 99)	633 (228, 1010)	1708 (945, 2429)	3953 (1499, 5888)	902 (372, 1253)	298 (209, 384)	8018 (3360, 11880)
	Fengtai	95 (25, 159)	91 (0, 183)	30 (19, 40)	257 (92, 411)	692 (382, 987)	1727 (622, 2704)	403 (156, 590)	123 (86, 159)	3417 (1381, 5233)
	Shijingshan	39 (10, 64)	38 (0, 75)	12 (8, 17)	107 (38, 170)	287 (159, 409)	685 (255, 1041)	158 (63, 224)	51 (35, 65)	1376 (569, 2066)
	Mentougou	7 (2, 12)	7 (0, 14)	2 (2, 3)	20 (7, 31)	53 (29, 75)	137 (48, 221)	32 (12, 49)	10 (7, 12)	268 (106, 419)
	Fangshan	67 (18, 111)	64 (0, 128)	21 (13, 29)	181 (65, 289)	487 (269, 694)	1207 (436, 1882)	281 (109, 410)	86 (60, 112)	2393 (970, 3655)
	Tongzhou	133 (36, 218)	130 (0, 259)	43 (27, 58)	369 (133, 589)	996 (551, 1416)	2312 (875, 3450)	528 (217, 735)	174 (122, 224)	4684 (1960, 6949)
	Shunyi	78 (21, 129)	75 (0, 150)	24 (16, 33)	211 (76, 338)	570 (315, 812)	1396 (509, 2159)	324 (127, 468)	101 (70, 131)	2779 (1133, 4220)
	Daxing	87 (23, 146)	84 (0, 168)	27 (17, 37)	235 (84, 377)	634 (350, 903)	1589 (570, 2497)	371 (143, 546)	113 (78, 146)	3140 (1266, 4819)
	Changping	99 (26, 165)	95 (0, 190)	31 (19, 42)	266 (95, 426)	717 (396, 1021)	1798 (644, 2825)	420 (162, 618)	127 (88, 165)	3551 (1431, 5452)
	Pinggu	18 (5, 30)	17 (0, 34)	6 (4, 8)	48 (17, 77)	130 (72, 185)	327 (117, 517)	77 (29, 113)	23 (16, 30)	646 (259, 994)
	Huairou	21 (6, 36)	20 (0, 41)	7 (4, 9)	57 (21, 92)	155 (86, 221)	386 (139, 603)	90 (35, 132)	28 (19, 36)	764 (309, 1169)
	Miyun	12 (3, 20)	12 (0, 23)	4 (2, 5)	32 (12, 52)	87 (48, 124)	225 (79, 362)	53 (20, 80)	16 (11, 20)	441 (175, 687)
	Yanqing	12 (3, 20)	11 (0, 23)	4 (2, 5)	32 (12, 51)	86 (48, 123)	219 (78, 347)	51 (20, 76)	15 (11, 20)	431 (173, 665)
	Beijing	1273 (339, 2098)	1235 (0, 2459)	403 (256, 549)	3487 (1253, 5574)	9407 (5202, 13393)	22500 (8338, 34283)	5186 (2078, 7383)	1655 (1155, 2137)	45146 (18621, 67876)
The policy of staying put during Spring Festival in 2021	**Dongcheng**	4 (1, 6)	4 (0, 8)	1 (1, 2)	12 (4, 19)	32 (18, 45)	55 (26, 68)	12 (6, 13)	5 (4, 7)	125 (60, 166)
	Xicheng	6 (2, 9)	6 (0, 12)	2 (1, 3)	18 (7, 29)	49 (27, 70)	86 (40, 106)	18 (10, 20)	8 (6, 10)	194 (92, 258)
	Chaoyang	18 (5, 28)	19 (0, 37)	6 (4, 9)	56 (20, 88)	151 (84, 213)	268 (122, 332)	56 (29, 64)	25 (18, 31)	599 (283, 801)
	Haidian	17 (5, 25)	17 (0, 33)	6 (4, 8)	50 (18, 79)	136 (76, 192)	242 (110, 301)	51 (26, 58)	22 (16, 28)	542 (255, 725)
	Fengtai	11 (3, 16)	11 (0, 22)	4 (2, 5)	33 (12, 52)	89 (50, 126)	157 (72, 194)	33 (17, 37)	15 (10, 18)	353 (167, 470)
	Shijingshan	3 (1, 5)	3 (0, 6)	1 (1, 1)	9 (3, 14)	25 (14, 35)	44 (20, 55)	9 (5, 11)	4 (3, 5)	98 (46, 131)
	Mentougou	2 (1, 3)	2 (0, 4)	1 (0, 1)	5 (2, 9)	15 (8, 21)	29 (12, 38)	6 (3, 8)	2 (2, 3)	62 (28, 86)
	Fangshan	6 (2, 10)	6 (0, 12)	2 (1, 3)	18 (7, 29)	49 (27, 69)	96 (41, 128)	21 (10, 26)	8 (6, 10)	207 (93, 286)
	Tongzhou	10 (3, 15)	11 (0, 21)	4 (2, 5)	32 (12, 50)	86 (48, 121)	145 (68, 174)	30 (16, 33)	14 (10, 18)	331 (160, 436)
	Shunyi	7 (2, 11)	8 (0, 14)	3 (2, 3)	22 (8, 34)	59 (33, 83)	103 (48, 127)	22 (11, 24)	10 (7, 12)	233 (111, 310)
	Daxing	11 (3, 16)	11 (0, 21)	4 (2, 5)	33 (12, 51)	88 (49, 125)	155 (71, 191)	33 (17, 37)	15 (10, 18)	349 (166, 465)
	Changping	11 (3, 17)	11 (0, 21)	4 (2, 5)	31 (11, 49)	84 (47, 119)	166 (70, 221)	36 (17, 44)	14 (10, 18)	356 (160, 494)
	Pinggu	3 (1, 4)	3 (0, 5)	1 (1, 1)	8 (3, 13)	22 (12, 31)	36 (18, 43)	8 (4, 8)	4 (3, 5)	84 (41, 110)
	Huairou	2 (1, 3)	2 (0, 4)	1 (0, 1)	7 (2, 10)	18 (10, 25)	33 (15, 43)	7 (4, 8)	3 (2, 4)	73 (33, 99)
	Miyun	3 (1, 4)	3 (0, 6)	1 (1, 1)	9 (3, 14)	24 (13, 34)	41 (19, 50)	9 (5, 10)	4 (3, 5)	93 (45, 124)
	Yanqing	2 (0, 2)	2 (0, 3)	1 (0, 1)	4 (2, 7)	12 (7, 17)	25 (10, 34)	5 (2, 7)	2 (1, 3)	52 (23, 73)
	Beijing	115 (33, 175)	120 (0, 229)	40 (26, 54)	346 (126, 546)	937 (522, 1323)	1682 (762, 2103)	356 (183, 409)	155 (111, 195)	3752 (1764, 5035)

It's the value of 95% CI in parentheses; the actual average annual concentration of $PM_{2.5}$ is less than or equal to 35 $\mu g/m^3$, and its health effect is 0.

TABLE 13 | Health benefits of $PM_{2.5}$ reduction caused by the COVID-19 epidemic of 16 districts of Beijing in three periods.

Districts	The initial COVID-19 outbreak in 2020			The outbreak in the Xinfadi market in 2020			The policy of staying put during the Spring Festival in 2021		
	Total health benefit (unit: million yuan)	Proportion of GDP (unit: %)	Per capita health benefit (unit: yuan)	Total health benefit (unit: million yuan)	Proportion of GDP (unit: %)	Per capita health benefit (unit: yuan)	Total health benefit (unit: million yuan)	Proportion of GDP (unit: %)	Per capita health benefit (unit: yuan)
Dongcheng	260.55 (104.29,349.77)	0.09 (0.04,0.12)	367.6 (147.13,493.47)	393.46 (134.02,603.14)	0.13 (0.05,0.2)	555.1 (189.08,850.94)	28.84 (11.78,38.06)	0.01 (0,0.01)	40.69 (16.61,53.69)
Xicheng	421.38 (173.58,552.43)	0.08 (0.03,0.11)	380.93 (156.91,499.39)	701.11 (246.74,1046.85)	0.14 (0.05,0.21)	633.8 (223.05,946.35)	44.91 (18.32,59.33)	0.01 (0,0.01)	40.6 (16.57,53.64)
Chaoyang	1294.46 (526.91,1715.78)	0.18 (0.07,0.24)	374.93 (152.62,496.97)	2154.15 (751.53,3229.36)	0.31 (0.11,0.46)	623.94 (217.68,935.37)	138.96 (56.26,184.65)	0.02 (0.01,0.03)	40.25 (16.3,53.48)
Haidian	1196.87 (495.08,1566.41)	0.14 (0.06,0.18)	381.9 (157.97,499.81)	1965.49 (687.41,2942.39)	0.23 (0.08,0.35)	627.15 (219.34,938.86)	125.7 (50.69,167.52)	0.01 (0.01,0.02)	40.11 (16.18,53.45)
Fengtai	784.61 (330.2,1013.55)	0.42 (0.18,0.55)	388.42 (163.47,501.76)	850.61 (285.52,1319.71)	0.46 (0.15,0.71)	421.09 (141.35,653.32)	81.6 (33.16,108.24)	0.04 (0.02,0.06)	40.4 (16.41,53.58)
Shijingshan	218.99 (91.56,284.22)	0.26 (0.11,0.33)	385.54 (161.2,500.38)	339.21 (116.51,515.54)	0.4 (0.14,0.6)	597.2 (205.12,907.64)	22.77 (9.2,30.43)	0.03 (0.01,0.04)	40.09 (16.19,53.57)
Mentougou	144.54 (57.84,193.85)	0.58 (0.23,0.77)	368.15 (147.33,493.77)	67.77 (22.53,106.95)	0.27 (0.09,0.43)	172.63 (57.38,272.42)	14.58 (5.6,20.41)	0.06 (0.02,0.08)	37.14 (14.25,51.99)
Fangshan	528.59 (232.1,662.06)	0.65 (0.29,0.82)	402.64 (176.8,504.31)	595.04 (200.24,920.4)	0.73 (0.25,1.14)	453.26 (152.53,701.1)	48.98 (18.69,68.03)	0.06 (0.02,0.08)	37.31 (14.23,51.82)
Tongzhou	706.68 (293.7,920.9)	0.64 (0.27,0.83)	384 (159.6,500.41)	1148.71 (400.8,1722.54)	1.04 (0.36,1.56)	624.2 (217.79,936.01)	75.96 (31.5,99.26)	0.07 (0.03,0.09)	41.28 (17.12,53.94)
Shunyi	480.57 (190.53,649.94)	0.26 (0.1,0.35)	362.97 (143.9,490.89)	689.56 (233.56,1059.32)	0.37 (0.12,0.57)	520.81 (176.41,800.09)	53.83 (21.99,71.09)	0.03 (0.01,0.04)	40.66 (16.61,53.69)
Daxing	793.36 (342.7,1004.48)	0.85 (0.37,1.08)	397.95 (171.9,503.85)	782.76 (261.64,1216.8)	0.84 (0.28,1.3)	392.63 (131.24,610.36)	80.69 (32.78,106.78)	0.09 (0.04,0.11)	40.47 (16.44,53.56)
Changping	817.78 (322.07,1110.29)	0.71 (0.28,0.97)	360.34 (141.91,489.22)	885.21 (295.71,1376.46)	0.77 (0.26,1.2)	390.05 (130.3,606.5)	84.43 (32.22,117.37)	0.07 (0.03,0.1)	37.2 (14.2,51.72)
Pinggu	178.06 (74.91,229.42)	0.63 (0.26,0.81)	389.36 (163.81,501.69)	161.2 (53.71,251.43)	0.57 (0.19,0.88)	352.51 (117.45,549.81)	19.25 (8.15,24.7)	0.07 (0.03,0.09)	42.1 (17.82,54)
Huairou	146.8 (54.92,207.47)	0.37 (0.14,0.52)	332.89 (124.54,470.45)	190.51 (63.92,294.89)	0.48 (0.16,0.74)	431.99 (144.94,668.69)	17.14 (6.68,23.16)	0.04 (0.02,0.06)	38.87 (15.15,52.52)
Miyun	177.16 (66.81,249.61)	0.52 (0.2,0.74)	335.72 (126.61,473.01)	110.89 (36.77,174.8)	0.33 (0.11,0.52)	210.14 (69.69,331.25)	21.49 (8.82,28.36)	0.06 (0.03,0.08)	40.72 (16.72,53.74)
Yanqing	124.27 (49.2,168.93)	0.64 (0.25,0.87)	359.46 (142.31,488.67)	108.04 (35.8,168.5)	0.56 (0.18,0.87)	312.52 (103.57,487.41)	12.5 (4.71,17.57)	0.06 (0.02,0.09)	36.17 (13.63,50.84)
Beijing	8274.65 (3406.4,10879.1)	0.23 (0.09,0.3)	377.96 (155.59,496.92)	11143.71 (3826.43,16949.1)	0.31 (0.11,0.47)	509.01 (174.78,774.18)	871.65 (350.54,1164.95)	0.02 (0.01,0.03)	39.81 (16.01,53.21)

It's the value of 95% CI in parentheses; "-" means that Chaoyang GDP in 2019 has not been released and cannot be calculated for the time being. If the actual annual $PM_{2.5}$ concentration is less than or equal to 35 $\mu g/m^3$, the health benefit is set to 0.

respiratory disorders, pose a potential threat to human life, and impose a significant economic cost on society. To this end, we also studied the health benefits of $PM_{2.5}$ reductions due to control measures in three periods. Firstly, we discovered that the health effects of the same area are various in distinct epidemic periods, and that the number of health beneficiaries in different regions is also heterogeneous. Chaoyang, Haidian, Tongzhou, and Fengtai are the urban areas that have benefited most from the $PM_{2.5}$ reduction, as evidenced by the COVID-19 outbreak in Beijing's Xinfadi market. Secondly, acute bronchitis, internal medicine clinics, and chronic bronchitis are the top three health endpoints for health benefits owing to the $PM_{2.5}$ decrease, accounting for about 80% of the total health effects. Meanwhile, the decline in $PM_{2.5}$ caused by control measures in the three epidemic periods probably avoided premature deaths of 1,117 cases [95% CI (328, 1,676)], 1,273 cases [95% CI (339, 2,098)], and 115 cases [95% CI (33, 175)], respectively. Furthermore, the total health and economic benefits owing to the $PM_{2.5}$ reduction affected by the restriction actions in three pandemic periods were 82,747.65 million yuan [95% CI (3,406.4, 10,879.1)], 11,143.71 million yuan [95% CI (3,826.43, 16,949.1)], and 871.65 million yuan [95% CI (350.54, 1,165.95)], respectively.

It is necessary to indicate two fields that require further study. Although these control measures have led to unprecedented improvements in air quality, air pollution in the ecological environment remained high during the blockade. Other factors contributing to air pollution, such as coal-fired winter heating systems and unfavorable weather conditions, might be responsible for air pollutant concentrations (Chen et al., 2013; Ebenstein et al., 2017). Second, the positive impact on air quality is transient since the study noticed that restriction efforts in the COVID-19 pandemic only reduced China's air pollutant concentrations in the short term. Nevertheless, in the long run, when COVID-19 control measures are lifted, large amounts of energy consumption and industrial activities may bring about more severe air pollution (Wang and Su, 2020). Maintaining this improvement in air quality remains a major challenge. Finally, as air pollution is consistent with human life and economic activities, the improvement of air quality belongs to a prolonged fight. Despite the strong restriction actions on dealing with public health crises, such as the COVID-19 epidemic, which have alleviated air pollution in the ecological environment, they have led to great impairment on society and the economy. Consequently, an essential task for improving the air quality in Beijing is the establishment of sustainable development strategies that consider the economic, social, and ecological environment factors (Wang and Watanabe, 2019; Lee et al., 2020).

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/s.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by the Major Program of the National Social Science Foundation of China (Grant No. 14ZDB151) and Jiangxi Provincial Natural Science Foundation (Grant No. 20212ACB211003).

REFERENCES

Ali, I., and Alharbi, O. M. L. (2020). COVID-19: disease, management, treatment, and social impact. *Sci. Total Environ.* 728, 138861. doi: 10.1016/j.scitotenv.2020.138861

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., and Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* 395, 931–934. doi: 10.1016/S0140-6736(20)30567-5

Archer-Nicholls, S., Carter, E., Kumar, R., Xiao, Q., Liu, Y., Frostad, J., et al. (2016). The regional impacts of cooking and heating emissions on ambient air quality and disease burden in China. *Environ. Sci. Technol.* 50, 9416–9423. doi: 10.1021/acs.est.6b02533

Bai, H., Gao, W., Zhang, Y., and Wang, L. (2022). Assessment of health benefit of $PM_{2.5}$ reduction during COVID-19 lockdown in China and separating contributions from anthropogenic emissions and meteorology. *J. Environ. Sci.* 115, 422–431. doi: 10.1016/j.jes.2021.01.022

Bel, G., and Holst, M. (2018). Evaluation of the impact of Bus Rapid Transit on air pollution in Mexico City. *Transp. Policy* 63, 209–220. doi: 10.1016/j.tranpol.2018.01.001

Berkowitz, S. A., and Basu, S. (2021). Unmet social needs and worse mental health after expiration of COVID-19 federal pandemic unemployment compensation. *Health Aff.* 40, 426–434. doi: 10.1377/hlthaff.2020.01990

Briz-Redón, Á., Belenguer-Sapiña, C., and Serrano-Aroca, Á. (2021). Changes in air pollution during COVID-19 lockdown in Spain: a multi-city study. *J. Environ. Sci.* 101, 16–26. doi: 10.1016/j.jes.2020.07.029

Brodeur, A., Gray, D., Islam, A., and Bhuiyan, S. (2021). A literature review of the economics of COVID-19. *J. Econ. Surv.* 35, 1007–1044. doi: 10.1111/joes.12423

Chen, R., Chen, B., and Kan, H. (2010). A health-based economic assessment of particulate air pollution in 113 Chinese cities. *China Environ. Sci.* 30, 410–415.

Chen, Y., Ebenstein, A., Greenstone, M., and Li, H. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proc. Natl Acad. Sci. U.S.A.* 110, 12936–12941. doi: 10.1073/pnas.1300018110

Chiaramonti, D., and Maniatis, K. (2020). Security of supply, strategic storage and Covid19: Which lessons learnt for renewable and recycled carbon fuels, and their future role in decarbonizing transport? *Appl. Energy* 271, 115216. doi: 10.1016/j.apenergy.2020.115216

CSY. (2020). *Beijing Statistical Yearbook*. Beijing: China Statistical Publishing House.

Dang, H. H., and Trinh, T. A. (2021). Does the COVID-19 lockdown improve global air quality? New cross-national evidence on its unintended consequences. *J. Environ. Econ. Manag.* 105, 102401. doi: 10.1016/j.jeem.2020.102401

Dutheil, F., Baker, J. S., and Navel, V. (2020). COVID-19 as a factor influencing air pollution? *Environ. Pollut.* 263, 114466. doi: 10.1016/j.envpol.2020.114466

Ebenstein, A., Fan, M., Greenstone, M., He, G., and Zhou, M. (2017). New evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River Policy. *Proc. Natl Acad. Sci. U.S.A.* 114, 10384–10389. doi: 10.1073/pnas.1616784114

Filonchyk, M., Hurynovich, V., Yan, H., Gusev, A., and Shpilevskaya, N. (2020). Impact assessment of COVID-19 on variations of SO_2, NO_2,

CO and AOD over East China. *Aerosol Air Qual. Res*. 20, 1530–1540. doi: 10.4209/aaqr.2020.05.0226

Fu, S., and Gu, Y. (2017). Highway toll and air pollution: evidence from Chinese cities. *J. Environ. Econ. Manag*. 83, 32–49. doi: 10.1016/j.jeem.2016.11.007

Giannadaki, D., Giannakis, E., Pozzer, A., and Lelieveld, J. (2018). Estimating health and economic benefits of reductions in air pollution from agriculture. *Sci. Total Environ*. 622–623, 1304–1316. doi: 10.1016/j.scitotenv.2017.12.064

Guo, X., Xiao, B., and Song, L. (2020). Emission reduction and energy-intensity enhancement: the expected and unexpected consequences of China's coal consumption constraint policy. *J. Clean. Prod*. 271, 122691. doi: 10.1016/j.jclepro.2020.122691

Gupta, A., Bherwani, H., Gautam, S., Anjum, S., Musugu, K., Kumar, N., et al. (2021). Air pollution aggravating COVID-19 lethality? Exploration in Asian cities using statistical models. *Environ. Dev. Sustain*. 23, 6408–6417. doi: 10.1007/s10668-020-00878-9

Hammitt, J. K., and Robinson, L. A. (2011). The income elasticity of the value per statistical life: transferring estimates between high and low income populations. *J. Benefit Cost Anal*. 2, 1–29. doi: 10.2202/2152-2812.1009

Han, S. J., Wang, J., Yan, Q. S., Yang, L. M., Zhao, Q. Y., Wang, S. B., et al. (2019). Health benefit evaluation for PM_{10} and $PM_{2.5}$ pollution control in Zhengzhou, China, 2014–2016. *Environmental Science*. 40, 2565–2571. doi: 10.13227/j.hjkx.201810255

Hao, Y., Niu, X., and Wang, J. (2021). Impacts of haze pollution on China's tourism industry: a system of economic loss analysis. *J. Environ. Manage*. 295, 113051. doi: 10.1016/j.jenvman.2021.113051

He, G., Pan, Y., and Tanaka, T. (2020). The short-term impacts of COVID-19 lockdown on urban air pollution in China. *Nat. Sustain*. 3, 1005–1011. doi: 10.1038/s41893-020-0581-y

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506. doi: 10.1016/S0140-6736(20)30183-5

Huang, D., Xu, J., and Zhang, S. (2012). Valuing the health risks of particulate air pollution in the Pearl River Delta, China. *Environ. Sci. Policy*. 15, 38–47. doi: 10.1016/j.envsci.2011.09.007

Huang, D., and Zhang, S. Q. (2013). Health benefit evaluation for $PM_{2.5}$ pollution control in Beijing-Tianjin-Hebei region of China. *China Environ. Sci*. 33, 166–174. doi: 10.3969/j.issn.1000-6923.2013.01.024

Hwang, H., and Lee, J. Y. (2022). Impacts of COVID-19 on air quality through traffic reduction. *Int. J. Environ. Res. Public Health*. 19, 1718. doi: 10.3390/ijerph19031718

Jiang, P., Fu, X., Van Fan, Y. V., Kleme,š, J. J., Chen, P., Ma, S., et al. (2021). Spatial-temporal potential exposure risk analytics and urban sustainability impacts related to COVID-19 mitigation: a perspective from car mobility behaviour. *J. Clean. Prod*. 279. 123673. doi: 10.1016/j.jclepro.2020.123673

Jiménez, J. L., and Perdiguero, J. (2017). *Difference-Indifference*. New York, NY: Springer. doi: 10.1007/978-1-4614-7883-6_664-1

Jones, E. (2020). COVID-19 and the EU economy: try again, fail better. *Survival* 62, 81–100. doi: 10.1080/00396338.2020.1792124

Ju, M. J., Oh, J., and Choi, Y. H. (2021). Changes in air pollution levels after COVID-19 outbreak in Korea. *Sci. Total Environ*. 750, 141521. doi: 10.1016/j.scitotenv.2020.141521

Kan, H., and Chen, B. (1989). Analysis of exposure–response relationships of air particulate matter and adverse health outcomes in China. *J. Environ. Health* 19, 422–424.

Kan, H. D., Chen, B. H., and Wang, H. (2004). Economic assessment of health impact due to particulate air pollution in urban areas of shanghai. Chin. *Health Econ*. 23, 8–11. doi: 10.3969/j.issn.1003-0743.2004.02.003

Kassa, M. D., and Grace, J. M. (2020). Race against death or starvation? COVID-19 and its impact on African populations. *Public Health Rev*. 41, 30. doi: 10.1186/s40985-020-00139-0

Kim, E., Kim, B. U., Kim, H. C., and Kim, S. (2017). The variability of ozone sensitivity to anthropogenic emissions with biogenic emissions modeled by MEGAN and BEIS3. *Atmosphere* 8, 187. doi: 10.3390/atmos8100187

Kip Viscusi, W. K., Magat, W. A., and Huber, J. (1991). Pricing environmental health risks: survey assessments of risk-risk and risk-dollar trade-offs for chronic bronchitis. *J. Environ. Econ. Manag*. 21, 32–51. doi: 10.1016/0095-0696(91)90003-2

Kraemer, M. U. G., Yang, C. H., Gutierrez, B., Wu, C. H., Klein, B., Pigott, D. M., et al. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* 368, 493–497. doi: 10.1126/science.abb4218

Lam, Y. F., Chang, J. M., Loo, B. P., Zhang, H. S., Leung, K. K., and Axhausen, K. W. (2022). Screening approach for short-term $PM_{2.5}$. Atmosphere (CO: health communications)-benefits: a case study from 15 metropolitan cities around the world during the COVID-19 pandemic. *Atmosphere* 13, 18. doi: 10.3390/atmos13010018

Lanzi, E., Dellink, R., Chateau, J., Holland, M., and Klimont, Z. (2016). *The Economic Consequences of Outdoor Air Pollution*. Paris: OECD.

Lee, H., Park, S. J., Lee, G. R., Kim, J. E., Lee, J. H., Jung, Y., et al. (2020). The relationship between trends in COVID-19 prevalence and traffic levels in South Korea. *Int. J. Infect. Dis*. 96, 399–407. doi: 10.1016/j.ijid.2020.05.031

Li, J., and Lin, B. (2017). Environmental impact of electricity relocation: a quasi-natural experiment from interregional electricity transmission. *Environ. Impact Assess. Rev*. 66, 151–161. doi: 10.1016/j.eiar.2017.07.002

Li, L., Lei, Y., Wu, S., Chen, J., and Yan, D. (2017a). The health economic loss of fine particulate matter ($PM_{2.5}$) in Beijing. *J. Cleaner Prod*. 161, 1153–1161. doi: 10.1016/j.jclepro.2017.05.029

Li, X., Qiao, Y., Zhu, J., Shi, L., and Wang, Y. (2017b). The "APEC blue" endeavor: causal effects of air pollution regulation on air quality in China. *J. Cleaner Prod*. 168, 1381–1388. doi: 10.1016/j.jclepro.2017.08.164

Lin, B., and Zhu, J. (2019). Impact of energy saving and emission reduction policy on urban sustainable development: empirical evidence from China. *Appl. Energy* 239, 12–22. doi: 10.1016/j.apenergy.2019.01.166

Liu, F., Page, A., Strode, S. A., Yoshida, Y., Choi, S., Zheng, B., et al. (2020). Abrupt decline in tropospheric nitrogen dioxide over China after the outbreak of COVID-19. *Sci. Adv*. 6, eabc2992. doi: 10.1126/sciadv.abc2992

Liu, F., Wang, M., and Zheng, M. (2021). Effects of COVID-19 lockdown on global air quality and health. *Sci. Total Environ*. 755, 142533. doi: 10.1016/j.scitotenv.2020.142533

Liu, X., Xie, P., Liu, Z., Li, T., Zhong, L., and Xiang, Y. (2010). Economic assessment of acute health impact due to inhalable particulate air pollution in the Pearl River Delta. *Acta Sci. Nat*. 46, 829–834. doi: 10.13209/j.0479-8023.2010.118

Lu, D., Zhang, J., Xue, C., Zuo, P., Chen, Z., Zhang, L., et al. (2021). COVID-19-induced lockdowns indicate the short-term control effect of air pollutant emission in 174 cities in China. *Environ. Sci. Technol*. 55, 4094–4102. doi: 10.1021/acs.est.0c07170

Lu, X., Lin, C., Li, Y., Yao, T., Fung, J. C., and Lau, A. K. (2017). Assessment of health burden caused by particulate matter in southern China using high-resolution satellite observation. *Environ. Int*. 98, 160–170. doi: 10.1016/j.envint.2016.11.002

Lv, L. Y., and Li, H. Y. (2016). Health economic evaluation of PM_{10} and $PM_{2.5}$ pollution in Beijing-Tianjin-Hebei region of China. *Acta Sci. Nat. Univ. Nankaiensis* 49, 69–77.

Mahato, S., and Pal, S. (2022). Revisiting air quality during lockdown persuaded by second surge of COVID-19 of mega city Delhi, India. *Urban Clim*. 41. 101082. doi: 10.1016/j.uclim.2021.101082

Maji, K. J., Ye, W. F., Arora, M., and Shiva Nagendra, S. M. (2018). PM2. 5-related health and economic loss assessment for 338 Chinese cities. *Environ. Int*. 121, 392–403. doi: 10.1016/j.envint.2018.09.024

Matus, K., Nam, K. M., Selin, N. E., Lamsal, L. N., Reilly, J. M., and Paltsev, S. (2012). Health damages from air pollution in China. *Glob. Environ. Change* 22, 55–66. doi: 10.1016/j.gloenvcha.2011.08.006

Meo, S. A., Al-Khlaiwi, T., Usmani, A. M., Meo, A. S., Klonoff, D. C., and Hoang, T. D. (2020). Biological and epidemiological trends in the prevalence and mortality due to outbreaks of novel coronavirus COVID-19. *J. King Saud Univ. Sci*. 32, 2495–2499. doi: 10.1016/j.jksus.2020.04.004

Muhammad, S., Long, X., and Salman, M. (2020). COVID-19 pandemic and environmental pollution: a blessing in disguise? *Sci. Total Environ*. 728, 138820. doi: 10.1016/j.scitotenv.2020.138820

Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., et al. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): a review. *Int. J. Surg*. 78, 185–193. doi: 10.1016/j.ijsu.2020.04.018

Querol, X., Massagu,é, J., Alastuey, A., Moreno, T., Gangoiti, G., Mantilla, E., et al. (2021). Lessons from the COVID-19 air pollution decrease in Spain: Now what? *Sci. Total Environ.* 779, 146380. doi: 10.1016/j.scitotenv.2021.146380

Rasul, G., Nepal, A. K., Hussain, A., Maharjan, A., Joshi, S., Lama, A., et al. (2021). Socio-economic implications of COVID-19 pandemic in South Asia: emerging risks and growing challenges. *Front. Sociol.* 6, 629693. doi: 10.3389/fsoc.2021.629693

Ren, J., Ma, Y., Panteha, F., and Sana, U. (2022). Managing transitions for sustainable economic development in post-COVID world: do fiscal and monetary support matter? *Econ. Res. Ekon. IstraŽivanja* 1–14. doi: 10.1080/1331677X.2022.2028176

Seo, J. H., Kim, J. S., Yang, J., Yun, H., Roh, M., Kim, J. W., et al. (2020). Changes in air quality during the COVID-19 pandemic and associated health benefits in Korea. *Appl. Sci.* 10, 8720. doi: 10.3390/app10238720

Shi, W., Bi, J., Liu, R., Liu, M., and Ma, Z. (2021). Decrease in the chronic health effects from PM2. 5 during the 13th five-year plan in China: impacts of air pollution control policies. *J. Clean. Prod.* 317, 128433. doi: 10.1016/j.jclepro.2021.128433

Vu, T. V., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., et al. (2019). Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Atmos. Chem. Phys.* 19, 11303–11314. doi: 10.5194/acp-19-11303-2019

Wan, Z., Zhou, X., Zhang, Q., and Chen, J. (2019). Do ship emission control areas in China reduce sulfur dioxide concentrations in local air? A study on causal effect using the difference-in-difference model. *Mar. Pollut. Bull.* 149, 110506. doi: 10.1016/j.marpolbul.2019.110506

Wang, C., Horby, P. W., Hayden, F. G., and Gao, G. F. (2020a). A novel coronavirus outbreak of global health concern. *Lancet* 395, 470–473. doi: 10.1016/S0140-6736(20)30185-9

Wang, L., and Watanabe, T. (2019). Effects of environmental policy on public risk perceptions of haze in Tianjin city: a difference-in-differences analysis. *Renew. Sustain. Energy Rev.* 109, 199–212. doi: 10.1016/j.rser.2019.04.017

Wang, M., Liu, F., and Zheng, M. (2021). Air quality improvement from COVID-19 lockdown: evidence from China. *Air Qual. Atmos. Health.* 14, 591–604. doi: 10.1007/s11869-020-00963-y

Wang, P., Chen, K., Zhu, S., Wang, P., and Zhang, H. (2020b). Severe air pollution events not avoided by reduced anthropogenic activities during COVID-19 outbreak. *Resour. Conserv. Recy.* 158, 104814. doi: 10.1016/j.resconrec.2020.104814

Wang, Q., and Su, M. (2020). A preliminary assessment of the impact of COVID-19 on environment - a case study of China. *Sci. Total Environ.* 728, 138915. doi: 10.1016/j.scitotenv.2020.138915

Wang, S., Zhao, M., Xing, J., Wu, Y., Zhou, Y., Lei, Y., et al. (2010). Quantifying the air pollutants emission reduction during the 2008 Olympic Games in Beijing. *Environ. Sci. Technol.* 44, 2490–2496. doi: 10.1021/es9028167

Wang, Y., Liu, H., Mao, G., Zuo, J., and Ma, J. (2017). Inter-regional and sectoral linkage analysis of air pollution in Beijing–Tianjin–Hebei (Jing-Jin-Ji) urban agglomeration of China. *J. Clean. Prod.* 165, 1436–1444. doi: 10.1016/j.jclepro.2017.07.210

Wei, G. R., and Shi, X. M. (2018). Evaluation the extent of health damage caused by $PM_{2.5}$ particulate in Xi'an city. *Environmental Science.* 39, 3014–3021. doi: 10.13227/j.hjkx.201710174

Xie, H., Zhu, T. Y., Liu, M. C., Song, D., Liu, J., and Zhang, J. L. (2015). Investigation and analysis of hospitalized patients of ACOS with asthma and COPD at the same time. *Prog. Mod. Biomed.* 15, 6279–6282. doi: 10.13241/j.cnki.pmb.2015.32.021

Xie, P., Liu, X., Liu, Z., Li, T., and Bai, Y. (2009). Exposure–response functions for health effects of ambient participate matter pollution applicable for China. *China Environ. Sci.* 29, 1034–1040. doi: 10.3321/j.issn:1000-6923.2009.10.006

Xie, X. X. (2011). *The Value of Health: Environmental Benefit Assessment Method and Control Strategy of Urban Air Pollution*. Beijing: Peking University.

Xie, Y. B., Chen, J., and Li, W. (2014). An assessment of PM2. 5 related health risks and impaired values of Beijing residents in a consecutive high-level exposure during heavy haze days. *Environmental Science.* 35, 1–8. doi: 10.13227/j.hjkx.2014.01.057

Yang, Y., Liu, X., Zheng, J., Tan, Q., Feng, M., Qu, Y., et al. (2019). Characteristics of one-year observation of VOCs, NOx, and O_3 at an urban site in Wuhan, China. *J. Environ. Sci.* 79, 297–310. doi: 10.1016/j.jes.2018.12.002

Yin, H., Pizzol, M., and Xu, L. (2017). External costs of $PM_{2.5}$ pollution in Beijing, China: uncertainty analysis of multiple health impacts and costs. *Environ. Pollut.* 226, 356–369. doi: 10.1016/j.envpol.2017.02.029

Zeng, J., and Wang, C. (2022). Temporal characteristics and spatial heterogeneity of air quality changes due to the COVID-19 lockdown in China. *Resour. Conserv. Recy.* 181, 106223. doi: 10.1016/j.resconrec.2022.106223

Zhang, F., Xu, J., Zhang, Z., Meng, H., Wang, L., Lu, J., et al. (2015). Ambient air quality and the effects of air pollutants on otolaryngology in Beijing. *Environ. Monit. Assess.* 187, 495. doi: 10.1007/s10661-015-4711-3

Zhang, J., Li, H., Lei, M., and Zhang, L. (2021). The impact of the COVID-19 outbreak on the air quality in China: evidence from a quasi-natural experiment. *J. Clean. Prod.* 296, 126475. doi: 10.1016/j.jclepro.2021.126475

Zhang, M., Song, Y., and Cai, X. (2007). A health-based assessment of particulate air pollution in urban areas of Beijing in 2000–2004. *Sci. Total Environ.* 376, 100–108. doi: 10.1016/j.scitotenv.2007.01.085

Zhang, Y., Li, W., and Wu, F. (2020). Does energy transition improve air quality? Evidence derived from China's winter clean heating pilot (WCHP) project. *Energy* 206, 118130. doi: 10.1016/j.energy.2020.118130

Zhou, M., Hu, T., Zhang, W., Wang, Q., Kong, L., Zhou, M., et al. (2022). COVID-19 pandemic: impacts on air quality and economy before, during and after lockdown in China in 2020. *Environ. Technol.* 1–11. doi: 10.1080/09593330.2022.2049894

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.

Copyright © 2022 Cai, Yin and Hao. 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 Environmental Science

BRIEF RESEARCH REPORT
published: 29 June 2022
doi: 10.3389/fenvs.2022.893824

Analysis of Ambient Air PM10-Bound Pollutants Surrounding an Industrial Site and Their Prediction Using Artificial Neural Network

M.G. Bonelli[1], M. Cerasa[2], E. Guerriero[2], A. Manni[3], S. Mosca[2], M. Perilli[2] and G. Rossetti[3]*

[1]Programming and Grant Office Unit, Italian National Research Council (CNR), Monterotondo, Rome, Italy, [2]Institute for Atmospheric Pollution Research (IIA), Italian National Research Council (CNR), Monterotondo, Rome, Italy, [3]Chemical Research 2000 Srl, Rome, Italy

The 2030 Agenda dictated the Sustainable Development Goals. It states the waste reduction needs through their reuse, i.e., considering them as secondary raw materials (Objective 12.5). Bottom ashes from municipal or industrial incinerators can be reused as partial cement replacement in concrete after preventive physical processes such as ferrous metals removal (magnetic separation) and nonferrous metals removal (Eddy current separation). Net of the principal pollutant containment systems, diffusive emissions of fine particles from these processes, coupled with several screening steps and a final long-time open-air residues stabilization, could impact the surrounding environment due to the chemical composition of the particulate matter itself (inorganic and organic pollutants). Moreover, the particulate may also arise from transporting the raw bottom ashes to the pre-treatment plant (point source). The present work aims to predict the concentration of the PM10-bound organic contaminants that are usually sampled weekly (PCDD/Fs, PCBs, PAHs) from the concentration of the daily analyzed inorganic pollutants in the surrounding area of an municipal solid waste slag treatment plant, using Artificial Neural Networks (ANNs) as a forecasting tool. Moreover, ANNs have also been used as a clustering tool to evaluate the plant's environmental impact on the surrounding area with respect to other additional emission sources.

Keywords: artificial neural network, organic micropollutants forecasting, data clusterization, PM10 characterization, MSWI slag

OPEN ACCESS

Edited by:
Wendong Yang,
Shandong University of Finance and Economics, China

Reviewed by:
Paolo Viotti,
Sapienza University of Rome, Italy
Erol Eğrioğlu,
Giresun University, Turkey

****Correspondence:***
M.G. Bonelli
mariagrazia.bonelli@cnr.it

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Environmental Science

Received: *11 March 2022*
Accepted: *13 June 2022*
Published: *29 June 2022*

Citation:
Bonelli MG, Cerasa M, Guerriero E, Manni A, Mosca S, Perilli M and Rossetti G (2022) Analysis of Ambient Air PM10-Bound Pollutants Surrounding an Industrial Site and Their Prediction Using Artificial Neural Network. Front. Environ. Sci. 10:893824. doi: 10.3389/fenvs.2022.893824

INTRODUCTION

A combustion process creates heat that is recycled and reused or converted to electrical energy. The fate of the residues (fly and bottom ashes) depends on their characteristics. The thermal treatment plants' residues from power production and municipal or industrial wastes show pozzolanic properties. They can be used as secondary raw materials for cement and building material production. (Giergiczny 1991; Kumar and Singh 2021; Mafalda Matos and Sousa-Coutinho 2022). This statement agrees with the request of the 2030 Agenda to protect the planet from degradation by minimizing waste generation through prevention, reduction, recycling, and reuse, as mentioned in its objective 12.5 (United Nations 2015). After incineration, bottom ashes are mainly composed of slag, synthetic ceramics, minerals, ferrous and nonferrous metals, unburned organic matter, glass, porcelain, and soluble salts such as hydroxides and chlorides. They are preliminarily

screened to remove bigger particles, after which they are subjected to physical treatments such as magnetic separation to extract ferrous materials and Eddy current separation to remove nonferrous metal constituents. They undergo a final long-time open-air stabilization. This can last up to 6–9 months, and it is used to weather/oxidize the components mainly by the action of O_2, CO_2, and water. The pH of the bottom ashes bulk will decrease, allowing the constituents' modification from hydroxides to sulfate than to carbonates, decreasing their leachability and dramatically contributing to the heavy metals leaching reduction (Chimenos et al., 2003). Finally, bottom ashes are sieved with different meshes to create aggregates showing several physical and mechanical characteristics such as density, compressive strength, and flexural strength, making them suitable for various uses in construction either directly or as an aggregate in other materials (Saffarzadeh et al., 2011; Spreadbury et al., 2021; Y.; Kim and Lee 2002; Youcai 2017; Astrup et al., 2016; Koksal et al., 2021). Prior to any other considered use, their environmental impact must be proven. It means sampling, analysis, and data processing. The focus is on the content of metals and persistent organic pollutants (POPs) due to the toxic aspects (Kim et al., 2004; Wei et al., 2021). The analysis of POPs (i.e., PCDD/Fs, PCBs, and PAHs) requires multiple days of sampling (due to the low concentrations in air, an enrichment of the sample is necessary to have an amount greater than the instrumental limit of quantification) and the laboratory analysis (extraction, analysis and data processing) requires at least 2 days. Usually, airborne metals are collected daily on filters and then determined by Inductively Coupled Plasma–Optical Emission Spectrometer (ICP-OES), Atomic Emission Spectrometer (AES), Mass Spectrometer (MS), and X-ray fluorescence (XRF) (Suvarapu and Baek, 2017).

Artificial neural networks (ANNs) are computational methodologies that perform multifactorial analyses. Inspired by biological neuron processes, the concept was introduced in 1943 by McCulloch and Pitts, simulating how the human brain processes information through the nerve cells, or neurons, connected to each other in a complex network within a computational model (McCulloch and Pitts, 1943). ANNs can model complicated and non-linear relationships. Moreover, from a modeling perspective, it works as a black box (Mjalli et al., 2007): it can approximate any function, studying its structure, but it cannot give any insights about the structure of the function being approximated. Therefore, ANNs can process the available data (input) and produce a prediction of the target value (output), identifying and learning the effects of an unknown complex cause-effect relationship between input and output through a training process. A neural network can approximate a wide range of statistical models without hypothesizing in advance any relationships between the dependent and independent variables. Instead, the form of the relationships is determined during the learning process. If a linear relationship between the dependent and independent variables is appropriate, the neural network results should closely approximate those of the linear regression model. If a non-linear relationship is more appropriate, the neural network will automatically match the "correct" model structure. The neuron (node) is the basic processing unit in neural networks. Neural networks impose minimal demands on model structure and assumptions. Still, it is necessary to choose the general network architecture correctly, consisting of multiple layers of nodes in a directed graph. Each layer is fully connected to the next one.

ANNs are a useful statistical tool for solving classifications, clustering, regression, pattern recognition, dimension reduction, structured prediction, machine translation, anomaly detection, decision making, visualization, and computer vision problems. They are often used as alternative forecasting methods in many fields, such as marketing, meteorology, and finance, where a significant amount of data is challenging to manage. In environmental sciences, they have been recently used in the prediction of sorption/desorption of chemicals from soil (Silva et al., 2019), delineation of soil contaminant plumes (Tao et al., 2019), risk assessment, and spatial modeling of heavy metals (Abbaszadeh et al., 2020), soil infiltration in furrow irrigation (Nazli et al., 2019), determination of principal components affecting soil infiltration (Alipour et al., 2021), forecasting the change in organic agricultural output (Doan 2021), investigating in PAHs bioremediation (Bao et al., 2019) and investigating the atmospheric sciences (Gardner and Dorling 1998). Many researches were performed using ANNs for air pollutant time series modeling and air pollutant concentrations forecasting, describing this method as good training, validation, and testing techniques and discussing measurements of performance and reliability (Prachi and Matta, 2011). An ANNs model has also been used to forecast short and middle long-term concentration levels of well-known air pollutants (Viotti et al., 2002). The method has shown outstanding performances for the short forecasts. For the medium and long-term forecasts, the results are better than the usual deterministic models in terms of mean square error (MSE), introducing hypotheses about the values of the meteorological and traffic parameters. Other studies have compared the predictive ability of the ANN models (non-linear method) for forecasting concentrations of air pollutants with the Multi-Linear Regression (MLR), proving that MLR is better than ANNs except in a few cases (Cakir and Moro 2020).

A multilayer perceptron (MLP) is a class feedforward artificial neural network. It consists of three or more layers (an input and an output layer with one or more hidden layers) of nonlinearly activating nodes: it is a function of predictors (also called inputs or independent variables) that minimize the prediction error of target variables (also called outputs). An example of MLP's architecture is shown in **Supplementary Material S1**.

The MLP model is an example of a feedforward neural network, referring to a fully connected network with three or more layers (an input and an output layer with one or several hidden layers) of nonlinearly activating nodes. The connections are unidirectional, and there are no cycles or loops in the network; thus, each neuron is linked only to neurons in the next layer. Each layer is connected to the adjacent neurons through an activation function, and all connections have their weights.

The MLP's Learning Process Occurs in the Following Consecutive Phases

a) Training (or calibration) phase: the original input set is divided into three subsets: training set, test set, and holdout set (Riad et al., 2004). The MLP reads the input and output variables of the training set and optimizes the prediction error of the output.
b) Testing (or verification) phase: the model accuracy is estimated by error indicators such as the Coefficient of determination (R^2) and the Root Mean Square Error (RMSE) of prediction calculated for the holdout set. The minimum RMSE and the maximum R^2 are often used to select the "better" neural network (Afan et al., 2015). RMSE is calculated in both the training and test set. Comparing both values, if they are of the same order of magnitude, the neural network provides reliable predictions (Chaloulakou et al., 2003).

As previously stated, the analysis of POPs is time- and cost-consuming. In this study, an MLP model has been used to predict the concentration of PM10-bound organic micropollutants (PCDD/Fs, dl-PCBs, and PAHs) from the concentration of daily airborne metals in an area where a municipal solid waste (MSW) bottom slag recovery plant is present, also considering whether it is possible to apply the neural networks for identifying the different emission sources. This work aims to underline that ANNs can be a helpful tool for predicting the concentrations of persistent pollutants and as a support tool for the plant manager to reduce the fallout of its emissions on the ground.

MATERIALS AND METHODS

Sampling Area and Strategy

This section provides information related to ambient air sample data collection. The industrial plant, focus of this work, bottom slag recovery plant. It is a mechanical slag treatment in an industrial-covered shed. There is an aspiration system for collecting and treating dust emissions generated by processing incoming waste and bag filters for air filtration. The plant is in a mainly periurban area, where other productive settlements are located (concrete production, semi-finished food products, carpentry, welding). Ambient air samples were collected at three sites, named A, B, and C, selected based on the position of the industrial plant (distance, wind direction). In detail, site A can be considered representative of the maximum fallout of the plant, as it is located within the perimeter of the plant itself; site B is located 4 km West of the plant, in a suburban site, on a moderately high-traffic road; site C is located 3 km East of the plant, in an urban park.

Ambient air samples were collected in two experimental campaigns in summer and winter, lasting 3 weeks each. Air samples were collected using a high-volume sampler (Echo PUF high volume sampler, TCR Tecora, Milan, Italy), equipped with a quartz fiber filter (QFF) and a polyurethane foam (PUF), allowing simultaneous sampling of particulates and gases at a flow rate of 200 L/min for organic micropollutants, and with a SkyPost PM10 sampler for the collection of particulate matter on which the subsequent analysis of metals was carried out. The organic micropollutants were collected weekly (18 samples), whereas the particulate matter was collected daily.

Chemical Analysis

Once collected, the samples were sent to the laboratory for analysis. Each QFF + PUF sample was spiked with standard solutions (Wellington Lab, Canada) containing PCDD/Fs (EN-1948 ES) and dl-PCBs (WP-LCS) prior to the extraction process (36 h Soxhlet extraction with toluene). The extract was concentrated and divided into two fractions - one for PAHs and one for separating PCDD/Fs and PCBs. A subsequent clean-up followed (Mosca et al., 2010) prior to the instrumental GC/MS analysis. The analysis of metals was based on the extraction of each filter in an ultrasonic bath, extraction of the residue via acid digestion, and the subsequent ICP/MS analysis of both fractions, according to Canepari et al. (2006).

Statistical Analysis

In this work, preliminary correlation analysis has been performed, investigating the possible linearity of the relationships between the considered variables employing Pearson's correlation matrix. If there is a linear dependence between organic, PM10, and inorganic variables, it could be assumed that the same emission source is present. Then, a Principal Component Analysis (PCA) is processed to increase the interpretability of the variable's relationships. Finally, a Multilayer Perceptron algorithm has been carried out, considering PM10 and 27 metals as input variables and the organic contaminants (TCDD, $\sum$ PCDD/F, $\sum$ PCB, BaP, and $\sum$ PAHs as the output to be predicted. Daily data for PM10 and metals were aggregated on a weekly basis to carry out a consistent analysis.

All statistical applications were performed using the software package SPSS v. 27.

RESULTS AND DISCUSSION

Preliminary Statistical Analysis

As mentioned above, a preliminary analysis for the study of linear dependence was performed. PM10, 27 metals (independent variables or *input*), and organic micropollutants (dependent variables or *output*) were analyzed from 18 air samples. Among micropollutants, 2.3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) and benzo(a)pyrene (BaP) were considered, along with the sum of PAHs, PCDD/Fs, and dioxin-like PCBs, due to their toxicological aspect. Pearson's correlation matrix with the significance level at $p < 0.05$ is shown in the **Supplementary Material S1**.

Correlation analysis shows the presence of multicollinearity: two or more of the predictors (or *input* variables) are moderately or highly correlated with one another. This occurs, for instance,

TABLE 1 | Principal Component Analysis (PCA).Rotated component matrix and variance explained.

	PC1	PC2	PC3	PC4	PC5	PC6
PM10	0.057	–0.045	0.269	–0.692	–0.088	0.489
As	0.872	0.321	–0.074	–0.163	–0.013	–0.042
Ba	0.901	0.111	–0.058	–0.002	–0.227	0.116
Ca	0.224	–0.709	0.587	0.166	–0.158	–0.114
Cd	–0.315	0.259	0.589	–0.426	0.243	–0.431
Ce	0.871	0.393	–0.196	0.114	0.054	0.084
Co.	0.916	0.329	-0.058	0.015	0.054	-0.012
Cr	0.446	0.813	–0.153	0.236	0.082	0.144
Cs	–0.197	0.651	–0.246	0.345	0.451	0.202
Cu	0.529	–0.109	0.817	–0.026	0.029	0.000
Fe	0.532	0.763	–0.065	0.285	0.148	0.142
K	–0.592	–0.347	0.536	0.018	–0.276	0.234
Li	0.774	0.297	0.232	0.443	0.218	–0.048
Mg	0.551	–0.712	0.141	–0.075	0.14	–0.185
Mn	0.66	–0.182	0.647	0.188	0.047	0.04
Mo	-0.01	0.419	0.802	0.054	–0.176	0.323
Na	–0.422	–0.538	0.11	0.416	0.217	0.173
Ni	0.737	–0.244	0.182	–0.421	–0.012	–0.113
Pb	–0.303	0.465	0.555	–0.377	0.383	–0.296
Rb	–0.618	0.427	0.435	0.209	–0.104	0.385
Sb	0.269	0.105	0.84	0.000	0.002	0.398
Sn	0.058	0.694	0.553	–0.244	0.268	–0.193
Sr	0.691	–0.419	0.383	0.346	–0.032	–0.203
Ti	0.656	0.64	–0.044	–0.18	0.242	0.003
Tl	–0.888	0.095	0.078	–0.031	0.31	0.133
V	0.829	–0.308	0.092	0.131	0.026	-0.11
W	0.464	0.353	–0.108	–0.668	–0.275	0.208
Zn	0.067	–0.2	0.889	0.243	0.067	–0.056
TCDD	–0.43	0.667	0.238	0.115	–0.397	–0.259
PCDD_F	–0.318	0.611	0.107	0.072	–0.466	-0.3
PCB	0.327	0.609	0.026	0.266	–0.604	–0.196
BaP	–0.828	0.395	0.244	0.118	0.014	–0.077
PAH	–0.777	0.419	0.243	0.072	–0.025	–0.186
% variance	26.2	22.6	15.8	10.5	8.9	7.6
% cumulative variance	26.2	48.8	64.6	75.1	84.0	91.6

for As and Co., whose Pearson's correlation coefficient r^2, at a significance level of 0.05, is equal to 0.93, or for Fe and Cr, with r^2 equal to 0.97.

Some important relationships are also observed between predictors and dependent variables. For example, TCDD is negatively correlated with Mg, with r^2 equal to -0.65; ∑ PCB has a linear correlation with Cr and Fe (r^2 equal to 0,63 and 0.60, respectively); BaP and ∑ PAHs are positively correlated with Tl, with r^2 as 0.76 and 0.70 respectively, and negatively correlated with Ba (r^2 as -0.75 and -0.70, respectively), Mg (r^2 as -0.70 and -0.64, respectively), Ni (r^2 as -0.68 and -0.61, respectively) and V (r^2 as -0.73 and -0.68, respectively). At the same time, ∑ PCDD/Fs do not correlate with any metal. The presence of linear relationships allows us to hypothesize for organic and inorganic contaminants with the same emission source and to perform forecasts through traditional methods, such as the Multi Linear Regression. The next step was the application of the Principal Component Analysis (PCA) to find out if there were any latent relationships between the variables ∑ PCDD/Fs and metals. In other words, PCA was used as data clustering to identify if and–eventually - which metal influences the presence of PCDD/Fs (and the other

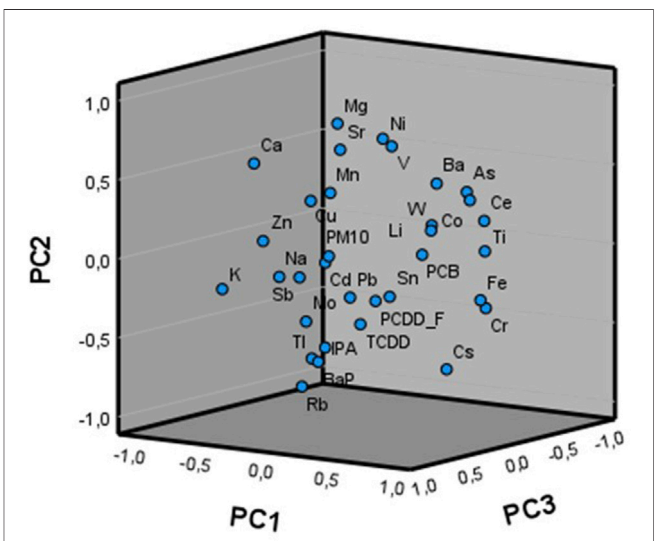


FIGURE 1 | The three-dimensional component plot in rotated space.

TABLE 2 | ANNs performance evaluation.

	R^2	$RMSE_{training}$	$RMSE_{test}$
TCDD			
Zone T	0.66	0.36	0.01
Zone E	0.93	0.39	0.02
PCDD_F			
Zone T	0.92	0.07	0.01
Zone E	0.81	0.022	0.09
PCB			
Zone T	0.83	0.17	0.00
Zone E	0.95	0.03	0.00
BaP			
Zone T	0.95	0.03	0.32
Zone E	0.96	0.05	0.00
PAH			
Zone T	0.91	0.05	0.04
Zone E	0.96	0.03	0.00

organic pollutants) to be referred to as the same contamination source.

The PCA's process computes six principal components by varimax orthogonal rotation criterium, as described in **Table 1**.

The first six PCs accounted for 91.6% of the total variation in the dataset: a six-dimensional space is supposed to be an excellent approximation to the original scatterplot of 33 variables, but it is not graphically representable. Furthermore, the first three PCs accounted only for 64.6% of the cumulative variance, and any cluster within the component plot is difficult to define (**Figure 1**).

It has thus been shown that PCA cannot clearly identify the sources of pollution, justifying the concomitant emission of PM10 and some metals with organic contaminants. Therefore, it was decided to apply an additional statistical clustering method to study the potential aggregations between variables and to explain the simultaneous presence of the variables considered in the same site.

The idea behind it all was to investigate the cause/effect relationships between all pollutants, in the three sites, by predicting the concentration of each organic component (output) resulting from PM10 and metals (input) through the development of an MLP model. The MLP, being a non-parametric technique, has been preferred to any other predictive method because the net can provide reliable results without hypothesizing in advance any relationships between the dependent and independent variables.

Artificial Neural Networks Results

Generally, ANNs models are considered a fundamental tool for collecting information about an extensive data system, but they can process small datasets. Referring to five output variables and 28 input variables (PM10 and the 27 metals), two MLP models have been performed for each organic contaminant (TCDD, PCDD/F, PCB, BaP, and PAHs), first considering all the 18 samples collected at the three sampling points, called Zone T (Sites A, B, and C), and then considering only the 12 samples placed outside the treatment plant, called Zone E (Sites B and C).

Each net has an MLP architecture 28-nine to one, where "28" is referred to the input variables (PM10 and 27 metals), "9" refers to the hidden variables (in one hidden layer), and "1" is referred to the output variables (organic contaminants). MLP architecture is shown in **Supplementary Material S1**.

The training set was used to train the network and the test set to evaluate the prediction performance of the ten models. R^2 and RMSE of the training and test set values are displayed in **Table 2**.

All values of R^2 are over 0.80, except for TCDD in zone T (0.66). Furthermore, the RMSE values in the training and test set are all in the same order. Therefore, the MLP models in Zone T and E provide reliable predictions.

The results for Zone T and Zone E models do not differ significantly, except for the TCDD variable, as the network could not "read" any relationships between TCDD and metals in samples from Site A (treatment plant). ANN predictive capability is higher for external samples (Zone E) than for the total of the samples (Zone T = Zone E + treatment plant) except for the variable PCDD/Fs. Among the external samples, the Urban site (site B) is more distant from the plant and influenced by ordinary traffic and trucks that go back and forth from the plant. This peculiarity impacts PCDD/F values and seems to affect the predictive capacity of the network adversely. The source of dioxin emission at this site is unknown and will be further investigated. It is also noted that anomalous values are recorded at the same site in the case of PCB-126 (3.3′,4.4′,5-PentaCB) and PCB-169 (3.3′,4′,5.5′-HexaCB).

A *T-test* for two independent groups has been applied to confirm if the performance of the Zone T and Zone E models provides similar analytical results for the other variables (Johnson and Wickern, 2014). The outcome of this test is the acceptance or rejection of the null hypothesis (H_0) within a predefined confidence level, generally at 95%. The null hypothesis states

TABLE 3 | *T-test* results.

TCDD	t_{exp}	Accptance region		H_0
		Inf	Sup	
R^2	5.88	-0.92	2.51	Reject
$RMSE_{training}$	22.74	0.16	0.58	Reject
$RMSE_{test}$	2.39	-0.05	0.07	Reject
PCDD_F	t_{exp}	Accptance Region		H0
		inf	sup	
R^2	15.87	0.17	1.55	Reject
$RMSE_{training}$	1.85	-0.26	0.35	Reject
$RMSE_{test}$	1.39	-0.41	0.51	Reject
PCB	t_{exp}	Accptance Region		H0
		inf	sup	
R^2	15.00	0.14	1.65	Reject
$RMSE_{training}$	1.45	-0.78	0.98	Reject
$RMSE_{test}$	1.00	-0.01	0.02	Reject
BaP	t_{exp}	Accptance Region		H0
		inf	sup	
R^2	424.51	0.93	0.98	Reject
$RMSE_{training}$	4.96	–0.06	0.14	Reject
$RMSE_{test}$	7.8	0.00	0.01	Reject
PAH	t_{exp}	Accptance Region		H0
		inf	sup	
R^2	42.94	0.66	1.21	Reject
$RMSE_{training}$	3.26	-0.12	0.2	Reject
$RMSE_{test}$	1.28	-0.19	0.24	Reject

that any differences or outlying results are purely due to random and not systematic errors. The alternative hypothesis (H_1) states precisely the opposite. Even though it is true, an erroneous rejection of H_0 constitutes a "type 1 error" or *p-value*. A smaller *p*-value means stronger evidence in favor of the alternative hypothesis. The most commonly used *p*-value is 0.05. To accept or reject H_0, the observed t-statistic text has to fall within the acceptance region (AR). The AR boundaries depend on the significance level of the test (the probability of erroneously rejecting the null hypothesis). Then they are calculated as a function of the *p*-value.

In this study, H_0 is the hypothesis that the predictive capability of both Zone T and Zone E models is the same for each output variable. *T-test* results are explained in the following **Table 3**.

For each organic pollutant, the test leads to the rejection of the null hypothesis: the predictive capability of the models is very different not only for TCDD but also for other variables despite the agreement between R^2 and RMSE values.

CONCLUSION

This study analyzed the concentration data of persistent organic micro-pollutants (PCDD/Fs, PCBs, and PAHs), PM10, and metals potentially emitted from an MSW residual treatment plant in ambient air. The traditional statistical approach could not clearly identify the sources of contaminants proving the same release of PM10, metals, and organic pollutants. ANN via MLP models was then applied to the dataset (concentration in three sampling sites), considering PM10 and metals as *input* and organic pollutants as *output*. As the first goal of this study, the contribution of the plant's emissions to the surrounding air was evaluated by differentiating the data analysis of the sampling sites: "T zone," including all three sites, and "E zone," including only the two most distant sites from the plant, thus excluding the concentrations in the site of maximum relapse.

An assessment of the predictive capability of the models (R^2 and RMSE) in both areas (inside and outside the plant) identified that the emission sources of external and internal samples were different. Therefore, the network's performance was higher for TCDD, PCB, BaP, and PAHs when only external samples were considered (even if the sample numbers are lower) since the model relationships were "contaminated" by the pollution sources within the treatment plant.

According to the R^2 values, the E model (external sites) for TCDD, PCB, BaP, and PAH provides more reliable predictions than the T Model (all sites) though with fewer samples, as if the stationary emission source due to the plant was "clouding" the relationships between the different pollutants. Conversely, in Zone E, the ANNs can better interpret the relationships. For PCDD/F, T Model is better than E Model: the relationships between the contaminants in the three sampling sites are more straightforward and allow the network to "learn" more.

Given the correspondence between the input and output data, it is possible to control the emission of micropollutants by monitoring the concentration of PM10 and metals (input). Furthermore, from an analytical point of view, it is easier and cheaper to obtain PM10 and metals data than POPs. This means that anomalous data of PM10 and/or metals (a daily event) and a higher concentration of POPs could be associated. In this case, it would be possible to promptly start the weekly sampling, thus reducing the costs of air quality analysis. Moreover, since pollutants are emitted from multiple sources, stationary and mobile, the application of ANN as a predictive tool can even support the plant manager (stationary source), acting on operative parameters (i.e., feeding, abatement systems, . . .) to control polluting emissions. In this way, the contribution to the total concentration of organic micropollutants in ambient air in the surrounding area can be monitored and eventually minimized almost in real-time.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because they are part of private ownership. Requests to access the datasets should be directed silvia.mosca@iia.cnr.it.

AUTHOR CONTRIBUTIONS

Conceptualization: MB, AM, and SM; data curation, MC, EG, SM, MP, and GR; formal analysis: MB; investigation, MB; methodology: MB; project administration: AM and SM; resources, AM, EG, SM, and GR; supervision, MB, AM, and SM; validation, MB; visualization, MB and AM; writing—original draft, MB, A.M. and SM; writing—review and editing, MB, AM, and SM All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

We want to thank Adriano Genovese for his careful revision of the text of the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenvs.2022.893824/full#supplementary-material

REFERENCES

Abbaszadeh, Maliheh, Mirzaei, Rouhollah, and Bakhtiari, Asra (2020). Risk Assessment and Spatial Modeling of Heavy Metals Contamination in Topsoil Around Venarj Manganese Mine by Artificial Neural Networks Method. *J. Environ. Health Eng.* 0 (0), 24–44. doi:10.29252/JEHE.0.24

Afan, H. A., El-Shafie, A., Yaseen, Z. M., Hameed, M. M., Wan Mohtar, W. H. M., and Hussain, A. (2015). ANN Based Sediment Prediction Model Utilizing

Different Input Scenarios. *Water Resour. Manage* 29 (4), 1231–1245. doi:10.1007/s11269-014-0870-1

Alipour, Nazli, Nasseri, Abolfazl, Mohammdi Torkashvand, Ali, and Pazira, Ebrahim (2021). Detecting the Principal Components Affecting Soil Infiltration Using Artificial Neural Networks. *Soil Environ.* 40 (1), 9–16. doi:10.25252/SE/2021/112007

Alipour, N., Nasseri, A., Ali, Mohammdi Torkashvand, and Ebrahim, Pazira (2019). Interpolation of Soil Infiltration in Furrow Irrigation: Comparison of Kriging, Inverse Distance Weighting, Multilayer Perceptron and Principal Component Analysis Methods. *pjss* 52 (1), 59–73. doi:10.17951/pjss/2019.52.1.5910.17951/pjss.2019.52.1.59

Astrup, T., Muntoni, A., Polettini, A., Pomi, R., Van Gerven, T., and Van Zomeren, A. (2016). Treatment and Reuse of Incineration Bottom Ash. *Environ. Mater. Waste Resour. Recovery Pollut. Prev. January*, 607–645. doi:10.1016/B978-0-12-803837-6.00024-X

Bao, H., Wang, J., Li, J., Zhang, H., and Wu, F. (2019). Effects of Corn Straw on Dissipation of Polycyclic Aromatic Hydrocarbons and Potential Application of Backpropagation Artificial Neural Network Prediction Model for PAHs Bioremediation. *Ecotoxicol. Environ. Saf.* 186 (December), 109745. doi:10.1016/J.ECOENV.2019.109745

Cakir, S., and Sita, M. (2020). Evaluating the Performance of ANN in Predicting the Concentrations of Ambient Air Pollutants in Nicosia. *Atmos. Pollut. Res.* 11 (12), 2327–2334. doi:10.1016/J.APR.2020.06.011

Canepari, S., Cardarelli, E., Pietrodangelo, A., and Strincone, M. (2006). Determination of Metals, Metalloids and Non-volatile Ions in Airborne Particulate Matter by a New Two-step Sequential Leaching procedurePart B: Validation on Equivalent Real Samples. *Talanta* 69 (3), 588–595. doi:10.1016/J.TALANTA.2005.10.024

Chaloulakou, A., Saisana, M., and Spyrellis, N. (2003). Comparative Assessment of Neural Networks and Regression Models for Forecasting Summertime Ozone in Athens. *Sci. Total Environ.* 313 (1–3), 1–13. doi:10.1016/S0048-9697(03)00335-8

Chimenos, J. M., Fernández, A. I., Miralles, L., Segarra, M., and Espiell, F. (2003). Short-Term Natural Weathering of MSWI Bottom Ash as a Function of Particle Size. *Waste Manag.* 23 (10), 887–895. doi:10.1016/S0956-053X(03)00074-6

Doan, Huy Quang (2021). Forecasting the Change in Organic Agricultural Output in the Northern Midland and Mountainous Provinces of Vietnam Based on Artificial Neural Network Model. *Int. J. Bus.* 26 (4), 119–132.

Gardner, M. W., and Dorling, S. R. (1998). Artificial Neural Networks (The Multilayer Perceptron)—A Review of Applications in the Atmospheric Sciences. *Atmos. Environ.* 32 (14–15), 2627–2636. doi:10.1016/S1352-2310(97)00447-0

Giergiczny, Z. (1991). The Wastes from Power Plants as Substitute of Natural Raw Materials. *Stud. Environ. Sci.* 48, 619–620. doi:10.1016/S0166-1116(08)70454-0

Johnson, R. A., and Wichern, D. W. (2014). *Applied Multivariate Statistical Analysis*. 6th Edn. Edinbugh: Pearson Educational Ltd.

Kim, K.-H., Seo, Y.-C., Nam, H., Joung, H.-T., You, J.-C., Kim, D.-J., et al. (2005). Characteristics of Major Dioxin/furan Congeners in Melted Slag of Ash from Municipal Solid Waste Incinerators. *Microchem. J.* 80, 171–181. doi:10.1016/j.microc.2004.07.022

Kim, Y., and Lee, D. (2002). Solubility Enhancement of PCDD/F in the Presence of Dissolved Humic Matter. *J. Hazard Mater* 91 (1–3), 113–127. doi:10.1016/S0304-3894(01)00364-8

Koksal, F., Gencel, O., Sahin, Y., and Okur, O. (2021). Recycling Bottom Ash in Production of Eco-Friendly Interlocking Concrete Paving Blocks. *J. Mater Cycles Waste Manag.* 23, 985–1001. doi:10.1007/s10163-021-01186-8

Kumar, S., and Singh, D. (2021). Municipal Solid Waste Incineration Bottom Ash: A Competent Raw Material with New Possibilities. *Innov. Infrastruct. Solut.* 6, 201. doi:10.1007/s41062-021-00567-0

Mafalda Matos, A., and Sousa-Coutinho, J. (2022). Municipal Solid Waste Incineration Bottom Ash Recycling in Concrete: Preliminary Approach with Oporto Wastes. *Constr. Build. Mater.* 323 (March), 126548. doi:10.1016/J.CONBUILDMAT.2022.126548

McCulloch, W. S., and Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. Math. Biophysics* 5, 115–133. doi:10.1007/bf02478259

Mjalli, F. S., Al-Asheh, S., and Alfadala, H. E. (2007). Use of Artificial Neural Network Black-Box Modeling for the Prediction of Wastewater Treatment Plants Performance. *J. Environ. Manag.* 83 (3), 329–338. doi:10.1016/J.JENVMAN.2006.03.004

Mosca, S., Torelli, G. N., Guerriero, E., Tramontana, G., Pomponio, S., Rossetti, G., et al. (2010). Evaluation of a Simultaneous Sampling Method of PAHs, PCDD/Fs and Dl-PCBs in Ambient Air. *J. Environ. Monit.* 12 (5), 1092–1099. doi:10.1039/b927004c

Prachi, Nishant Kumar, and Matta, Gagan (2011). Artificial Neural Network Applications in Air Quality Monitoring and Management. *Int. J. Environ. Rehabilatation Conservation* 2 (1), 1631–1638.

Riad, S., Mania, J., Bouchaou, L., and Najjar, Y. (2004). Predicting Catchment Flow in a Semi-arid Region via an Artificial Neural Network Technique. *Hydrol. Process.* 18 (13), 2387–2393. doi:10.1002/hyp.1469

Saffarzadeh, A., Shimaoka, T., Wei, Y., Gardner, K. H., and Musselman, C. N. (2011). Impacts of Natural Weathering on the Transformation/Neoformation Processes in Landfilled MSWI Bottom Ash: A Geoenvironmental Perspective. *Waste Manag.* 31 (12), 2440–2454. doi:10.1016/J.WASMAN.2011.07.017

Silva, T. S., de Freitas Souza, M., Maria da Silva Teófilo, T., Silva dos Santos, M., Formiga Porto, M. A., Martins Souza, C. M., et al. (2019). Use of Neural Networks to Estimate the Sorption and Desorption Coefficients of Herbicides: A Case Study of Diuron, Hexazinone, and Sulfometuron-Methyl in Brazil. *Chemosphere* 236 (December), 124333. doi:10.1016/J.CHEMOSPHERE.2019.07.064

Spreadbury, C. J., McVay, M., Laux, S. J., and Townsend, T. G. (2021). A Field-Scale Evaluation of Municipal Solid Waste Incineration Bottom Ash as a Road Base Material: Considerations for Reuse Practices. *Resour. Conservation Recycl.* 168 (May), 105264. doi:10.1016/J.RESCONREC.2020.105264

Suvarapu, L. N., and Baek, S.-O. (2017). Determination of Heavy Metals in the Ambient Atmosphere. *Toxicol. Ind. Health* 33 (January), 79–96. doi:10.1177/0748233716654827

Tao, H., Liao, X., Zhao, D., Gong, X., and Cassidy, D. P. (2019). Delineation of Soil Contaminant Plumes at a Co-contaminated Site Using BP Neural Networks and Geostatistics. *Geoderma* 354 (November), 113878. doi:10.1016/J.GEODERMA.2019.07.036

United Nations (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development*. Available at: https://sdgs.un.org/2030agenda.

Viotti, P., Liuti, G., and Di Genova, P. (2002). Atmospheric Urban Pollution: Applications of an Artificial Neural Network (ANN) to the City of Perugia. *Ecol. Model.* 148 (1), 27–46. doi:10.1016/S0304-3800(01)00434-3

Wei, J., Li, H., and Liu, J. (2021). Fate of Dioxins in a Municipal Solid Waste Incinerator with State-Of-The-Art Air Pollution Control Devices in China. *Environ. Pollut.* 289 (November), 117798. doi:10.1016/J.ENVPOL.2021.117798

Youcai, Z. (2017). Characterization and Recycling of Bottom Ash. *Pollut. Control Resour. Recovery Munic. Solid Wastes Incineration*, 61–82. doi:10.1016/B978-0-12-812165-8.00002-0

Conflict of Interest: Authors MA and RG are employed by Chemical Research 2000 Srl.

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.

Copyright © 2022 Bonelli, Cerasa, Guerriero, Manni, Mosca, Perilli and Rossetti. 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 Environmental Science

ORIGINAL RESEARCH
published: 30 June 2022
doi: 10.3389/fenvs.2022.941405

Research and Application of the Mode Decomposition-Recombination Technique Based on Sample-Fuzzy Entropy and K-Means for Air Pollution Forecasting

Chunhua Niu[1], *Ziqi Niu*[2], *Zongxi Qu*[1]*, *Lili Wei*[1] and *Yutong Li*[1]

[1]School of Management, Lanzhou University, Lanzhou, China, [2]School of Mathematics and Statistics, Lanzhou University, Lanzhou, China

Practical forecasting of air pollution components is important for monitoring and providing early warning. The accurate prediction of pollutant concentrations remains a challenging issue owing to the inherent complexity and volatility of pollutant series. In this study, a novel hybrid forecasting method for hourly pollutant concentration prediction that comprises a mode decomposition-recombination technique and a deep learning approach was designed. First, a Hampel filter was used to remove outliers from the original data. Subsequently, complete ensemble empirical mode decomposition adaptive noise (CEEMDAN) is employed to divide the original pollution data into a finite set of intrinsic mode function (IMF) components. Further, a feature extraction method based on sample-fuzzy entropy and K-means is proposed to reconstruct the main features of IMFs. In conclusion, a deterministic forecasting model based on long short-term memory (LSTM) was established for pollutant prediction. The empirical results of six-hourly pollutant concentrations from Baoding illustrate that the proposed decomposition-recombination technique can effectively handle nonlinear and highly volatile pollution data. The developed hybrid model is significantly better than other comparative models, which is promising for early air quality warning systems.

Keywords: hourly pollutants forecasting, decomposition-recombination technique, sample-fuzzy entropy, k-means, long short-term memory

OPEN ACCESS

Edited by:
Wendong Yang,
Shandong University of Finance and Economics, China

Reviewed by:
Qingbin Song,
Macau University of Science and Technology, Macau SAR, China
Feng Chen,
University of New South Wales, Australia

***Correspondence:**
Zongxi Qu
quzx@lzu.edu.cn

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: *11 May 2022*
Accepted: *13 June 2022*
Published: *30 June 2022*

Citation:
Niu C, Niu Z, Qu Z, Wei L and Li Y (2022) Research and Application of the Mode Decomposition-Recombination Technique Based on Sample-Fuzzy Entropy and K-Means for Air Pollution Forecasting. Front. Environ. Sci. 10:941405. doi: 10.3389/fenvs.2022.941405

INTRODUCTION

Following rapid industrialization and urbanization, various air pollution problems have occurred frequently. Air pollution has serious effects on human health and causes significant economic losses (Tang et al., 2010; Liu et al., 2011; Pandey et al., 2021). Therefore, establishing high-precision monitoring and prediction models is necessary to support governmental decision-making, environmental protection, and medical diagnosis.

Up to now, enormous amount of studies contributed to predicting future trends of air pollutants. In summary, most works modeled for prediction from three perspectives: mathematical and physical techniques, statistical prediction models, and machine learning models. First, mathematical and physical techniques have long been widely used in the field of air pollutant prediction. For instance, Huang et al. (2018) developed a random forest model, including gap-filled aerosol optical depth

(AOD), Modern-Era Retrospective analysis for Research and Applications Version 2 (MERRA-2) simulations, meteorological parameters, and land cover as predictors to estimate monthly $PM_{2.5}$ concentrations in North China. Tessum et al. (2015) used Weather Research and Forecasting with Chemistry Meteorological (WRF-Chem) and a chemical transport model (CTM) to simulate air pollution in adjacent areas of the United States for 12 months in 2005 with a horizontal resolution of 12 km and evaluated the simulation results. Atmospheric environment diffusion model techniques such as CTM and WRF can predict the pollutant concentration by solving the corresponding differential equation, which makes the prediction more deterministic (Yahya et al., 2014). Pollutants can further be predicted using the statistical forecasting model. Statistical methods are used to predict pollutants by mining time-series data for characteristic information. Zhang et al. (2018) used the autoregressive integrated moving average model (ARIMA) to predict the $PM_{2.5}$ concentration and compared it with other pollutant concentrations and meteorological parameters. Further, Wang P. et al. (2017) proposed a novel hybrid generalized autoregressive conditional heteroskedasticity (GARCH) method by combining ARIMA and support vector machine (SVM) forecasting models. Additionally, some improved statistical models, such as multiple linear regression (Elbayoumi et al., 2015; Yuchi et al., 2019; Yan and Enhua, 2020) and gray models (Chen & Pai, 2015; Wu & Zhao, 2019), are proposed for better prediction of PM2.5. Machine learning models such as artificial neural networks and support vector algorithms have recently become more prominent in pollution prediction. Various machine learning methods have been used in previous air pollution prediction studies. These include the following: backpropagation neural network (BPNN) (Bai et al., 2016); generalized regression neural network (GRNN) (Zhou et al., 2014); extreme learning machine (ELM) (Shang et al., 2019); random forest (Huang et al., 2018); support vector regression (SVR) (Zhu et al., 2018); long short-term memory (LSTM) (Qi et al., 2019; Yan et al., 2021) Zhang et al. (Zhang et al., 2019), integrated a multiple objectives model with five algorithms—BPNN, ARIMA, cuckoo search (CS), holt winters (HW) and online extreme learning machine (OELM)—for wind speed prediction. A constructed function comprising a three-objective combined model was optimized using a non-dominated sorting genetic algorithm. Liu et al. (Liu et al., 2018), constructed a combined model was constructed using a nonlinear neural network and statistical linear algorithm. Compared with several integrated models, it is more reliable and results in high accuracy.

However, it is hardly possible for a single prediction method to elaborately capture all complex features in pollution series which locate in a high dimension space. To this end, data preprocessing by outlier removing and series decomposition is efficient way for model construction at first. Various data preprocessing methods are developed for pollution data with nonlinearity and volatility present in. Data preprocessing approaches and optimization strategies have been extensively researched for pollutant prediction to increase the efficiency and accuracy of the prediction performance (Li & Zhu, 2018). Researchers usually propose suitable data preprocessing methods and process them according to study requirements. Several existing data preprocessing methods are relevant to the study of environmental contaminants. Empirical Mode Decomposition (EMD) (Huang et al., 1998) is a well-known algorithm for series decomposition. This algorithm projects a time series onto a set of intrinsic mode function (IMF) acting as bases because the project coefficients show good shapes via the Hilbert transform. These bases are derived from the phenomena of oscillations in the physical time domain. Owing to the poor performance of the subjective intervention for the intermittence test, EEMD (Wu and Huang, 2009) is proposed using noise-assisted data analysis (NADA) to construct a set of IMF. To increase the scales of a series at high frequency via the transformation of the IMF, an ensemble of white noise is incorporated for the designed trials because its scales are distributed uniformly in both the time and frequency domains. The true signal is estimated using the average of the ensemble in which the random white noise is canceled out, and only the persistent part of the signal remains. For example, (Zhou et al., 2014), suggested a hybrid ensemble empirical mode decomposition-generalized regression neural network (EEMD-GRNN) model that integrates the EEMD and a generalized regression neural network (GRNN) as a strategy for forecasting $PM_{2.5}$. Wang (Wang D. et al., 2017) developed a new hybrid model based on a two-phase decomposition technique and modified ELM to improve the forecasting accuracy of the air quality index. Xu et al. (Xu et al., 2017) developed a hybrid model based on Improved Complementary Ensemble Empirical Mode Decomposition, Whale Optimization Algorithm and Support Vector Machine (ICEEMD-WOA-SVM) to predict major pollutants, in which the data preprocessing part follows a "decomposition and integration" strategy. The raw series of each pollutant concentration was decomposed into several IMFs that were individually decomposed using a data preprocessing technique. The Hampel filter is an offline frequency-domain filtering method for eliminating spectral outliers (Allen, 2009). The advantage of the Hampel filter is that there is no prior need to know the outliers where the disturbance occurs. Moreover, the processed data series will not be distorted. Li et al. (Li et al., 2019) developed a new analysis and prediction system for air quality index prediction. Outliers in the air quality index series were eliminated using Hampel filter. Liu et al. (Liu and Chen, 2020) proposed a three-stage hybrid neural network model for outdoor $PM_{2.5}$ forecasting. K-means is an iterative clustering analysis algorithm used in pollutant data analysis. Riches et al. (Riches et al., 2022) employed the K-means cluster to analyze five concentrations. They further examined the patterns of association between $PM_{2.5}$, $PM_{10,}$ CO, NO_2, O_3, and SO_2 measurements and variations in annual diabetes incidence at the county level in the United States.

The data preprocessing methods mentioned above provided qualified data for later analysis with prediction models. However, most current studies only use a single data preprocessing technology which cannot offer well present data suitable for further modeling. For example, in some studies, the EEMD technique was the only method used to decompose the original data into numerous IMF components for reducing the

prediction complexity. The effective extraction of features from IMFs is difficult because features with diversity in frequency domain might be caused by outliers, which introduce disturbance into prediction. Therefore, in this study, the original data were first filtered by Hampel to eliminate the outliers in the data. The data were then decomposed into several IMFs using complete ensemble empirical mode decomposition adaptive noise (CEEMDAN). The complexity characteristics of the different sequences were obtained by calculating the fuzzy entropy and information entropy of each IMF signal. Subsequently, similar IMFs are recombined using the K-means clustering method based on fuzzy entropy and information entropy. After that, a prediction model was established using LSTM to conduct an empirical study of the six pollutants.

RELATED METHODOLOGIES

Data Preprocessing Method

Six data preprocessing methods—the Hampel filter, CEEMDAN, Sample entropy (SE), Fuzzy entropy (FE), K-means, and LSTM prediction methods—were applied in this study to better predict the concentrations of six pollutants: $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, and CO. First, the original data were filtered by Hampel filter to eliminate the outliers in the data. Then, the data was decomposed into several IMFs using CEEMDAN. The complicated characteristics of the different sequences were obtained by calculating the fuzzy entropy and information entropy of each IMF signal. Finally, to sum up similar IMFs in each group clustered with the K-means in terms of fuzzy entropy and information entropy.

Hampel Filter

The Hampel filter is an offline frequency-domain filtering method used to eliminate spectral outliers that are difficult to represent elaborately using prediction models. By representing the sequence with a one-dimensional vector, the method generates a local window around each element of the vector and calculates the median of all elements in that window. The standard deviation of each sample was further estimated using the absolute value of the median. The absolute difference between the sample and median shorted in the MAD can be a direct measurement for outlier detection. Mathematically, the Hampel filter detects elements as outliers in a vector using **Eq. 1**:

$$\begin{cases} S_j = 1.4286 median\left\{\left|x_j - x_j^*\right|\right\}, \\ x_j = x_j^*, \quad \text{if}\left|x_j - x_j^*\right| > tS_j, \quad \text{for} \quad j = 1, \ldots, N, \end{cases} \tag{1}$$

where t is the threshold, and N is the length of the vector. The variables x_j^* and S_j are the median and standard deviation of the window centered at element x_j, respectively. The deviation S_j is estimated adaptively by multiplying MAD and a constant. Element x_j is further replaced with x_j^* when the MAD is t times larger than S_j, the standard deviation.

Algorithm of CEEMDAN

EEMD can obtain better IMF than EMD. However, it does not result in exact decomposition because the white noise drives the generation on new modes that hide within the mixed IMFs. Furthermore, the IMF might not be orthogonal so that the energy of the added white noise is not similar to that when the polluted series are expanded by the IMF. To overcome this problem, CEEMDAN first defines a residual between the series and variation IMF from EEMD and then applies the step in EMD, which extracts the most IMF of the residual. The above steps were repeated until the residual energy was small. The last residual is defined as the last mode, which is why this algorithm is considered complete.

Let $E_1(\bullet)$ denote the operator that decomposes the first mode from a series, defined as in EMD, and let IMF_j^{V} denote the variation of the $j-$th IMF. Assuming M, the number of trials in EEMD, the procedure of CEEMDAN is described in detail as follows:

Step 1: The first variation IMF is the same as the first in EEMD.

$$IMF_1^{\mathrm{v}}[n] = \frac{1}{M}\sum_{i=1}^{M} IMF_1^i[n] = \overline{IMF_1}[n].$$

Step 2: The following is the first residual off the decomposed series.

$$r_1[n] = x[n] - IMF_1^{\mathrm{v}}[n].$$

Step 3: Let the second mode be the mean of the decompositions of the residuals enhanced by adaptive noise with $E_1(\bullet)$ in an ensemble of trials.

$$IMF_2^{\mathrm{v}}[n] = \frac{1}{M}\sum_{i=1}^{M} E_1\left(r_1[n] + \varepsilon_1 E_1\left(\omega^i[n]\right)\right).$$

Step 4: Similar to step 2, define the $k-$th residual off the $(k-1)-$th residual.

$$r_k[n] = r_{(k-1)}[n] - IMF_k^{\mathrm{v}}[n].$$

Step 5: Extract the $(k+1)-$th mode, $IMF_{k+1}^{\mathrm{V}}[n]$, from the enhanced $k-$th residual by an adaptive noise,

$$IMF_{(k+1)}^{\mathrm{v}}[n] = \frac{1}{M}\sum_{i=1}^{M} E_1\left(r_k[n] + \varepsilon_k E_k\left(\omega^i[n]\right)\right).$$

Step 6: After the number of K decompositions, the last residual is given as follows:

$$R[n] = x[n] - \sum_{k=1}^{K} IMF_k^{\mathrm{v}},$$

and we have the exact decomposition

$$x[n] = \sum_{k=1}^{K} IMF_k^{\mathrm{v}} + R[n].$$

Sample Entropy

Sample entropy is a new measure of time-series complexity proposed by Richman and Moornan (Richman et al., 2000), which aims to reduce the error of the approximate entropy algorithm with higher accuracy. Sample entropy was calculated as follows:

From a time series $X = [x_1, x_2, \cdots, x_n]$, and a tolerance r,
Step 1:Generate a group of vector by rolling on time,

$$X_i^m = [x_i, x_{i+1}, \cdots, x_{i+m-1}],\ i = 1, 2, ..., n-m+1.$$

Step 2: Define the distance between X_i^m and X_j^m with the maximum norm of $X_i^m - X_j^m$,

$$d_{ij}^m = d\left(X_i^m, X_j^m\right) = \max\left\{\left|x_{i+k} - x_{j+k}\right|, k = 0, 1, \cdots, m-1\right\}.$$

Step 3:For each i, $i = 1, \ldots, n-m+1$, calculate $P_i^m(r)$ by the sum of indicate function $I(\bullet)$, $P_i^m(r) = \frac{1}{n-m}\sum_{j\neq i} I(d_{ij}^m < r)$, and $\phi^m(r) = \frac{1}{n-m+1}\sum_{i=1}^{n-m+1} P_i^m(r)$.
Step 4:Let $m \triangleq m+1$, and repeat steps 1–3, get $\phi^{m+1}(r)$, then the sample entropy is defined as

$$SE^m(r) = -\ln\left(\frac{\phi^{m+1}(r)}{\phi^m(r)}\right) \tag{2}$$

Fuzzy Entropy

The concept of fuzzy sets was first introduced by Zadeh (Zadeh, 1965), which resulted in the formation of fuzzy entropy with further research on fuzzy sets. The statistical measure of fuzzy entropy was further developed by Chen et al. (Chen et al., 2009) to characterize the degree of fuzziness of fuzzy sets. As a measure of complexity, there is less bias, and continuity is achieved as well as free parameter selection and greater robustness against noise.

There are many definitions of fuzzy entropy as long as the definition satisfies the four rules described in this study (Zadeh, 1965). In this study, fuzzy entropy is similarly formulated as part of the sample entropy. The first difference is that a constant c is subtracted from all the elements of the vector X_i^m generated by rolling on time:

$$X_i^m = [x_i, x_{i+1}, \cdots, x_{i+m-1}] - c,\ i = 1, 2, ..., n-m+1,$$

where $c = \frac{1}{m}\sum_{k=i}^{i+m-1} k$. The second difference is the definition of distance for two vectors,

$$d_{ij}^m = \exp\left\{-\left(\frac{\left|X_i^m - X_j^m\right|_{\max}}{r}\right)^2\right\}.$$

The other steps are similar to those of the sample entropy. In conclusion, the fuzzy entropy is defined as

$$FuzzyEn(t) = \ln\phi^m(t) - \ln\phi^{m+1}(t). \tag{3}$$

K-Means

The K-means algorithm is a classic method for clustering points in high-dimensional space, as proposed by Macqueen (Macqueen, 1967). Based on the criterion of similarity between two points, usually measured by the Euclidean distance, a point is determined to belong to the class whose center is closest to it. The centers of all groups are updated after all points in the dataset are set. The algorithm stops when the cluster measurement function converges, which means that there are no changes for all centers in the updating.

The distance and similarity between points X_i and C_j, the center of the *jth* class, can be calculated as follows:

$$d\left(x_i, c_j\right) = \sqrt{\left(x_i^1 - c_j^1\right)^2 + ... + \left(x_i^k - c_j^k\right)^2 + ... + \left(x_i^d - c_j^d\right)^2},$$
$$s\left(x_i, c_j\right) = 1/d\left(x_i, c_j\right).$$

For each updated class, a new cluster center is calculated. Assuming that the samples in the *j*th class are $\{x_{j1}, x_{j2}, ..., x_{jn_j}\}$ and the cluster center is $c_j = (c_j^1, c_j^2, ..., c_j^k, ..., c_j^d)$, the *kth* attribute of class center C_j is represented as CK_j, which is

$$c_j^k = \frac{x_{j1}^k + x_{j2}^k + ... + x_{jn_j}^k}{n_j}.$$

The above process is repeated until the standard measure function converges. The conventional clustering measure function is usually the mean-square deviation, which is expressed as

$$J = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_j}\left(x_{ij} - c_i\right)^2}{n-1}}. \tag{4}$$

LSTM

LSTM is a type of recurrent neural network (RNN), which was originally established by Hochreiter and Schmidhuber (Aksoy et al., 2018) and was refined and popularized by many others in subsequent work. RNN are sensitive to short-term information. However, they always have a problem of long-term reliance. As an improvement to RNN, LSTM solves this problem by introducing a cell state in which the long-term state is saved. In this neural network, there are some LSTM blocks, which are regarded as intelligent net cells in certain studies. In several versions of LSTM, the most important LSTM cell is "forget gate." There are four neural network layers, each of which interacts in a unique manner.

The first stage of LSTM is to identify the information from the cell state that should be forgotten or rejected. The "forget gate layer," formulated by a sigmoid layer, makes this judgment. For input h_{t-1} and x_t, a number between zero and one is the output for each number in the cell state C_{t-1}. Output 1 means "keep this completely," while 0 means "forget this completely." This step is formulated as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right). \tag{5}$$

where h_{t-1} is the output of the previous input value, representing the effect on the current input value. Matrix w_f is the weight of the input value, and b_f is the deviation of the input value. Output f_t is the result of function σ, a 0–1 output function.

We further determine which part of the information needs to be stored in the cell state. The "input gate layer" i_t first determines which values will be updated. Denoting $C_t^{\sim}$ as the "tanh layer," a vector is generated for new candidate values. These two layers are further integrated to update the state.

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$
$$C_t^{\sim} = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right)$$

where tanh is a function of hyperbolic arctangent. The other parameters are similar to those in **Eq. 5**.

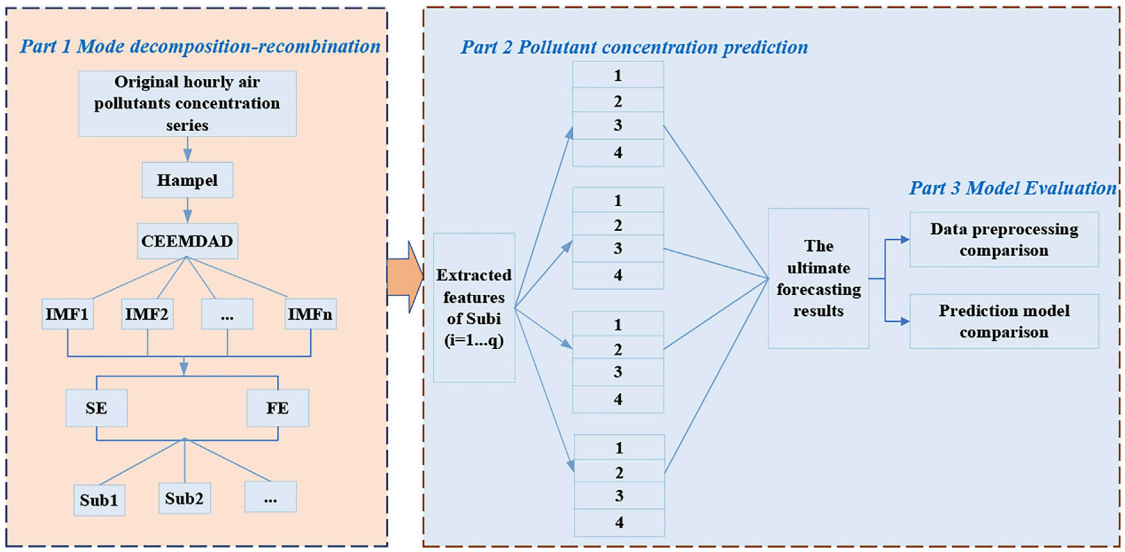


FIGURE 1 | The framework diagram of the hybrid model.

The next step is to multiply the old state by f_t, forgetting the information that should be rejected earlier. This is implemented using the following operation:

$$C_t = f_t * C_t + i_t * \tilde{C}_t$$

In conclusion, the unit state O_t is determined by the output of the sigmoid layer. Subsequently, the cell state is transformed by the tanh function, which outputs a value between 1 and -1. For the next recurrent, h_t is the multiplication of the unit state and tanh transform of the cell state. This is formulated as

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o),$$
$$h_t = o_t * \tanh(C_t). \quad (6)$$

HYBRID MODEL ARCHITECTURE

This section introduces the proposed hybrid model architecture, which includes the following three parts: data preprocessing, pollutant concentration prediction, and model evaluation. **Figure 1** shows a framework diagram of the proposed model.

Part 1 Mode Decomposition-Recombination

Step 1: Original data were filtered using Hampel filtering to eliminate outliers.

Step 2: The filtered data were decomposed into several IMF component sequences using CEEMDAN.

Step 3: Calculated the information entropy and fuzzy entropy of each IMF component into a two-dimensional vector.

Step 4: Based on the calculation results of the information entropy and fuzzy entropy, K-means was used to cluster the IMF components to achieve feature extraction.

Part 2 Pollutant Concentration Prediction

Step 1: For each data group obtained from clustering, the 4-fold cross-validation method was used for training.

Step 2: Setting up the LSTM model structure, the hidden layer was selected as a 2-layer LSTM structure, the number of neuron nodes in the first layer was 64, the number of neuron nodes in the second layer was 32, and the output layer reduced the results to the original data format.

Step 3: The mean absolute error (MAE) was chosen as the loss function, the Adam algorithm was used to generate optimization parameters for each node learning, and the error was reduced by iterating and adjusting the weights until convergence.

Step 4: To obtain the final prediction results, the prediction results of each group were superimposed.

Part 3 Model Evaluation

Step 1: Designed model evaluation experiments. In this study, we designed two sets of evaluation experiments: 1) data preprocessing comparison; 2) prediction model comparison.

Step 2: In the data preprocessing comparison experiments, we chose three comparison models: 1) the LSTM model without data preprocessing; 2) Hampel integrated with the LSTM model; 3) the CEMMDAN integrated LSTM model with our proposed SFE-K-Means integrated LSTM model for the comparison experiments.

Step 3: For the prediction model comparison experiments, we chose the backpropagation neural network (BPNN), evolutionary neural network (ENN), and Extreme Learning Machine (ELM), which are the three benchmark comparison models for comparison with the LSTM model and our proposed model.

TABLE 1 | Evaluation criteria.

Criteria	Interpretation	Equation
MAE	The average absolute error of n times forecast results	$\frac{1}{N}\sum_{n=1}^{N}\lvert y_n - \hat{y}_n\rvert$
MSE	The mean-square forecast error	$(\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2)$
MAPE	The mean absolute precent error of forecasting results	$\frac{1}{N}\sum_{n=1}^{N}\lvert\frac{y_n - \hat{y}_n}{y_n}\rvert \times 100\%$

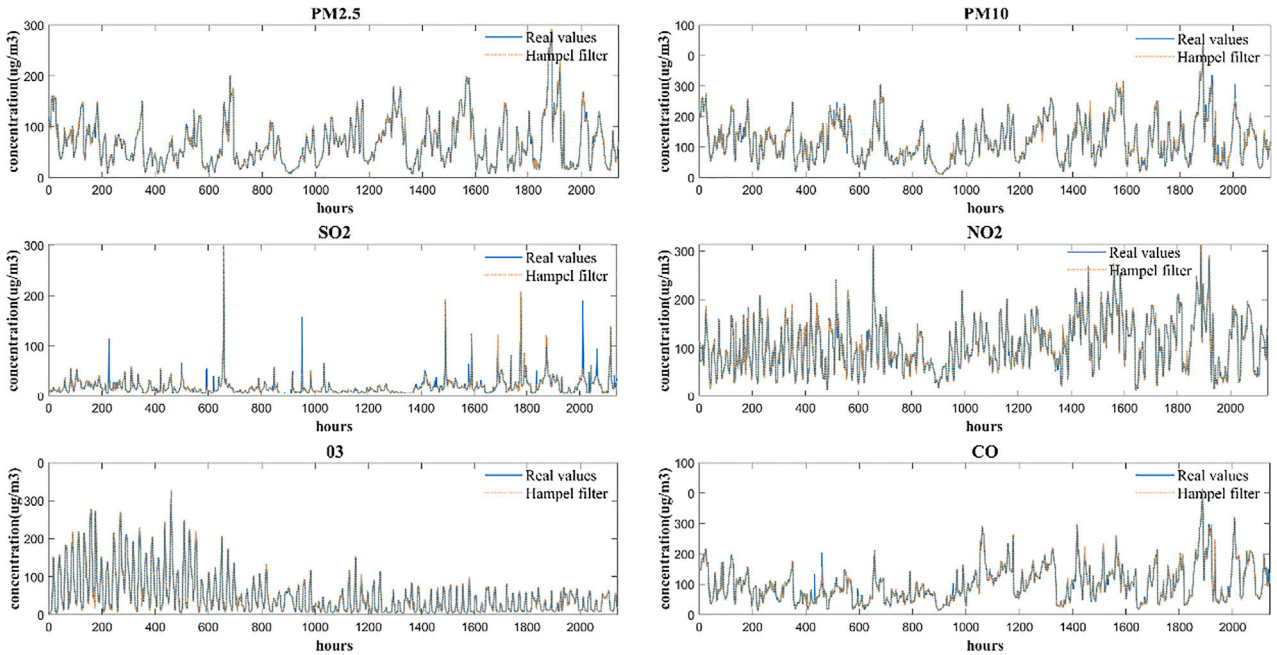


FIGURE 2 | Filtering results of the six pollutants time series.

Step 4: We chose mean squared error (RMSE), MAE, and mean absolute percentage error (MAPE) as the model evaluation criteria for the above two sets of experiments.

EMPIRICAL STUDY

Data Description

Major air pollutants in the atmosphere, including $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2 and O_3, were selected as the research objects in this study. The Ministry of Environmental Protection of the People's Republic of China (http://www.mep.gov.cn/) has provided six pollutant concentration datasets from Baoding. Sample data were collected on September 1, 2017, and November 30, 2017, in Baoding. The hourly pollution concentration data totaled 2140. These datasets were split into two categories: training and testing. The first 1814 data (approximately 85% of the total data) are training sets, and 321 data points (approximately 15%) for test.

Performance Evaluation Criteria

This study considers three assessment criteria, as in **Table 1**, to effectively evaluate the performance of the model. MAE, MSE and MAPE were chosen as error criteria to reflect the prediction performance of the forecasting models.

Mode Decomposition-Recombination Technique Process

Results of Outlier Detection

The series of original environmental pollution concentrations have obvious volatility and nonlinear characteristics and contain a few outliers. Therefore, data preprocessing is required for the original data. This section first uses a Hampel filter to process the original data. The filtering results of the six pollutant time series are shown in **Figure 2**, which shows that the filtered time series present more smooth appearance and more stable variation in local area after outliers and noise are eliminated from the original data.

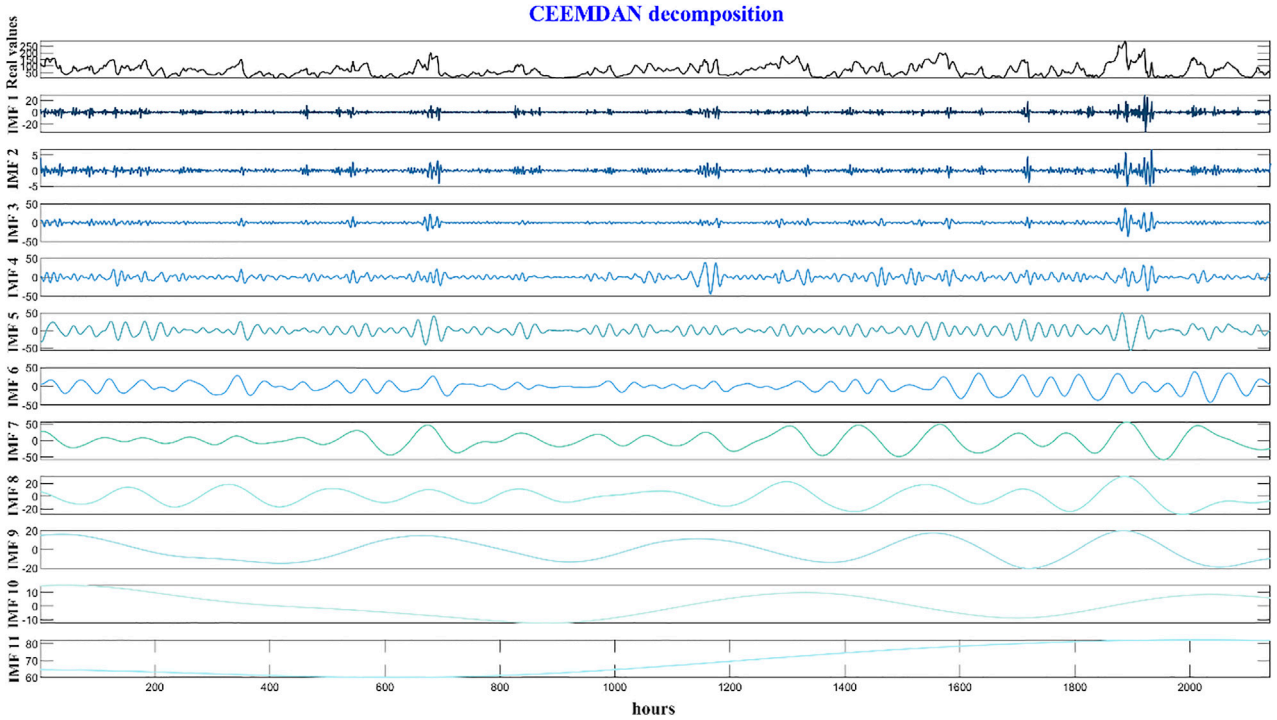


FIGURE 3 | Pollutant series decomposed by CEEMDAN.

TABLE 2 | Sample and fuzzy entropy of IMF.

		IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8	IMF9	IMF10	IMF11
$PM_{2.5}$	FE	8.047495	6.388529	1.649125	1.270372	0.959852	0.497776	0.27859	0.040618	0.023789	0.004574	9.89E-05
	SE	1.983576	0.554265	1.519343	1.491674	0.965727	0.607932	0.51636	0.488107	0.370485	0.110548	0.02104
PM_{10}	FE	1.412543	4.431981	4.210672	1.61084	0.69924	0.770135	0.408633	0.239158	0.025897	0.00488	0.000646
	SE	2.723644	1.099536	2.191735	1.993386	1.239376	0.690948	0.658875	0.400867	0.394307	0.16184	0.019505
SO_2	FE	4.267658	5.091819	1.867888	1.2594	0.592948	0.455568	0.253383	0.058592	0.012499	0.002048	0.000226
	SE	1.086146	0.337643	1.280924	0.891944	0.725795	0.608708	0.527715	0.387489	0.140906	0.042888	0.014509
NO_2	FE	4.322556	5.581931	2.524202	1.441533	0.967665	0.439743	0.194742	0.067011	0.016767	0.00236	0.00E+00
	SE	2.394989	0.813402	2.054457	1.785199	0.805091	0.674112	0.549718	0.43341	0.196665	0.075553	0.003059
O_3	FE	3.359121	4.02093	2.487006	1.975004	0.654115	0.418519	0.161993	0.037779	0.016124	0.004214	0.005488
	SE	1.717077	0.839211	2.102009	1.854934	1.134547	0.629392	0.616712	0.500818	0.197826	0.047921	0.015869
CO	FE	6.753463	0.270155	0.981246	0.746559	0.19851	0.052625	0.016649	0.002809	9.54E-05	0	0
	SE	0.013549	0	0.02167	0.080561	0.086343	0.0539	0.034768	0.010918	0.00312	0.00199	0.000891

Results of Decomposition for Six Pollutants Data

In this section, CEEMDAN is used to decompose the original amount of pollutant series into a collection of IMFs with associated frequencies and the residue component. We chose an ideal standard deviation of 0.1–0.5 and a total of 200 ensemble members. The original pollutant series decomposed using CEEMDAN is shown in **Figure 3**.

Calculate Sample and Fuzzy Entropy

Figure 3 clearly shows that the original data are decomposed by CEEMDAN to obtain different frequency components. From IMF1 to IMF11, the higher frequency of the IMF components indicates that each component contains more information and complexity. Therefore, we calculated the sample entropy and fuzzy entropy of each IMF separately to evaluate the complexity characteristics of different IMF time series. **Table 2** shows that the frequency of the sequence from IMF1 to IMF11 gradually decreases, and the calculated sample entropy and fuzzy entropy also gradually decrease, which indicates that the complexity of the IMFs decreases.

Results of K-Means Cluster

Based on the entropy value results for each IMF component obtained from **Table 2**, a cluster analysis was implemented with K-means method. The clustering centers and groupings of each pollutant were obtained, as shown in **Table 3**. From the results in

TABLE 3 | Clustering centers and groupings of each pollutant by K-means.

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
$PM_{2.5}$	Clustering Center	(0.7288,0.7868)	(6.3885,0.5542)	(8.04749,1.9835)	(0.0023,0.0657)	(1.4597,1.5055)	(0.1143,0.4583)
	Group	IMF5 IMF6	IMF2	IMF3	IMF10 IMF11	IMF3 IMF4	IMF7 IMF8 IMF9
PM_{10}	Clustering Center	(0.7346,0.9651)	(0.1358,0.3270)	(4.4319,1.0995)	(1.4125,2.7236)	(1.6108,1.9933)	(4.2106,2.1917)
	Group	IMF5 IMF6	IMF7 IMF8 IMF9 IMF10 IMF11	IMF2	IMF1	IMF4	IMF3
SO_2	Clustering Center	(0.0183,0.1464)	(5.0918,0.3376)	(1.2594,0.8919)	(0.4339,0.6207)	(1.8678,1.2809)	(4.2676,1.0861)
	Group	IMF8 IMF9 IMF10 IMF11	IMF2	IMF4	IMF5 IMF6 IMF7	IMF3	IMF1
NO_2	Clustering Center	(0.0561,0.2516)	(5.5819,0.8134)	(4.3225,2.3949)	(0.7037,0.7396)	(1.4415,1.7851)	(2.5242,2.0544)
	Group	IMF7 IMF8 IMF9 IMF10 IMF11	IMF2	IMF1	IMF5 IMF6	IMF4	IMF3
O_3	Clustering Center	(0.6541,1.1345)	(2.2310,1.9784)	(4.0209,0.8392)	(0.0086,0.0872)	(0.2060,0.5823)	(3.3591,1.7170)
	Group	IMF5	IMF3 IMF4	IMF2	IMF9 IMF10 IMF11	IMF6 IMF7 IMF8	IMF1
CO	Clustering Center	(0.01180,0.0176)	(6.7419,0.0134)	(0.7940,0.0808)	(0.1979,0.0863)	(0.2685,4.389e-07)	(1.0444,0.0219)
	Group	IMF6 IMF7 IMF8 IMF9 IMF10 IMF11	IMF1	IMF4	IMF5	IMF2	IMF3

TABLE 4 | Prediction performance of data preprocessing methods.

Pollutants	Criteria	LSTM	Hampel*	CEEMDAN*	SFE-K-Means*
$PM_{2.5}$	*MSE*	178.5243	187.8801	41.9019	38.4436
	MAE	8.3156	8.2933	3.4539	3.4379
	MAPE	14.59%	14.95%	7.29%	6.86%
PM_{10}	*MSE*	421.1828	521.2269	78.2811	75.5773
	MAE	13.8984	15.1586	5.1386	4.9915
	MAPE	12.52%	13.51%	4.66%	4.93%
SO_2	*MSE*	133.8757	72.4359	32.8127	14.8968
	MAE	5.6349	3.9291	2.9422	1.8789
	MAPE	23.86%	18.63%	12.32%	7.91%
NO_2	*MSE*	69.0062	64.0526	51.8403	29.8058
	MAE	5.7719	5.5731	4.6785	2.8757
	MAPE	14.65%	13.35%	12.26%	7.52%
O_3	*MSE*	30.1279	34.0037	8.1717	3.4115
	MAE	3.9207	4.2903	2.3501	1.3998
	MAPE	16.23%	18.12%	13.18%	6.89%
CO	*MSE*	0.0404	0.0582	0.0091	0.0121
	MAE	0.1349	0.1559	0.0568	0.0568
	MAPE	11.28%	12.61%	5.71%	6.22%

Table 3, we found that the 11 IMFs were clustered and reintegrated into 6 clusters. Each cluster is composed of IMF components with similar characteristics. The IMF components of each cluster are added together to form the final feature extraction datasets.

Comparison of Forecasting Results

Comparison of Data Preprocessing Methods

In this experiment, the concentrations of six pollutants in Baoding were predicted and analyzed. This experiment compares the performance of three preprocessing models—Hampel-LSTM, CEEMDAN-LSTM, and our proposed model. Additionally, the evaluation criteria of *MSE*, *MAE*, and *MAPE* were used to measure the prediction performance of the models and the results are presented in **Table 4**. Boldly marked values are used to indicate the best values of the model in different evaluation metrics. Further discussion of the experimental results is as follows.

For the different data processing methods of the LSTM-based hybrid models, **Table 4** shows that Hampel, CEEMDAN, and SFE-K-Means integrated with the same LSTM have obvious differences in prediction accuracy. However, compared with a single LSTM prediction model, the three hybrid models with signal processing tools—Hampel-LSTM(Hampel*), CEEMDAN-LSTM(CEEMDAN*), and SFE-K-Means-LSTM (SFE-K-Means*)—have better prediction performance. Therefore, it is safe to conclude that the use of mixed-preprecessing can significantly improve the data quality for the later hybrid model to

TABLE 5 | Comparison of benchmark methods.

Pollutants	Criteria	BPNN	ENN	ELM	LSTM	Proposed Model
$PM_{2.5}$	*MSE*	279.2252	326.0792	223.4873	178.5243	38.4436
	MAE	9.551	10.9026	8.4289	8.3156	3.4379
	MAPE	14.45%	18.10%	13.90%	14.59%	6.86%
PM_{10}	*MSE*	489.6128	8.93E+02	446.8126	421.1828	75.5773
	MAE	13.7485	18.0084	13.4037	13.8984	4.9915
	MAPE	11.74%	14.84%	11.74%	12.52%	4.93%
SO_2	*MSE*	142.8222	235.1045	171.601	133.8757	14.8968
	MAE	5.9537	7.8745	6.2234	5.6349	1.8789
	MAPE	25.87%	33.46%	25.49%	23.86%	7.91%
NO_2	*MSE*	75.7468	94.0673	71.512	69.0062	29.8058
	MAE	5.9489	6.5661	5.4114	5.7719	2.8757
	MAPE	14.14%	16.13%	13.76%	14.65%	7.52%
O_3	*MSE*	33.7319	44.2665	37.1021	30.1279	3.4115
	MAE	4.145	5.1152	4.3318	3.9207	1.3998
	MAPE	17.55%	25.35%	19.30%	16.23%	6.89%
CO	*MSE*	0.0404	0.0672	0.0535	0.0404	0.0121
	MAE	0.1234	0.1635	0.1358	0.1349	0.0568
	MAPE	10.16%	12.14%	10.71%	11.28%	6.22%

obtain better prediction results. Subsequently, three hybrid prediction models based on different signal processing tools—Hampel*, CEEMDAN*, and SFE-K-Means*—were compared, and SFE-K-Means was found to have the highest prediction accuracy. For example, as for $PM_{2.5}$, the *MAPE* values of LSTM, Hampel*, CEEMDAN*, and SFE-K-Means* were 14.59, 14.95, 7.29, and 6.86%, respectively. Thus, LSTM integrated with SFE-K-Means outperforms the other data preprocessing models.

Comparison of Benchmark Methods

This experiment compares the performances of four single benchmark prediction models, including BPNN, ENN, and ELM. The models' prediction performance was assessed using the MSE, MAE, and MAPE evaluation criteria; the results are presented in **Table 5**. The best results in the numerous evaluation metrics are emphasized by bold font. The results of the experiments are summarized below.

Table 5 clearly shows that LSTM seems to have more substantial predictive power than BPNN, ENN, and ELM. In the six pollutant concentration predictions, LSTM was superior to the other comparative models for all evaluation indexes. For example, the MAPE values for $PM_{2.5}$ via BPNN, ENN, ELM, LSTM and proposed model were 14.45, 18.10,13.90, 14.59 and 6.86%, respectively. The proposed model, which integrates SFE-K-Means with LSTM, results in the smallest MAE, MSE, and MAPE values, which says it should outperform the other benchmark methods to compare with. Notably, as a novel data preprocessing approaches, SFE-K-Means is critical for enhancing the forecast accuracy for environmental pollutant concentration.

CONCLUSION

The practical analysis and forecasting of pollutant concentrations are critical for environmental management and public health. Owing to the fluctuation and complexity of the pollutant data series, a novel mode decomposition-recombination technique is proposed to capture valuable information and characteristics. Six pollutant concentration series collected from Baoding were used as test cases to conduct the empirical study. Two experiments were implemented to compare the performances of the data preprocessing and forecasting methods, respectively. The evaluation criteria of MAE, MSE and MAPE were used to examine the prediction performance of the models. Based on the results of hourly pollutant concentration forecasting, some vital conclusions were drawn as follows. First, compared with Hampel*, CEEMDAN*, and SFE-K-Means*, the proposed SFE-K-Means* was found to have the highest prediction accuracy. Shown in **Table 4** as for PM2.5, the MAPE values of LSTM, Hampel*, CEEMDAN*, and SFE-K-Means* were 14.59, 14.95, 7.29, and 6.86%, respectively. These errors explain that LSTM integrated with SFE-K-Means outperformed the other data preprocessing models. Second, compared with BPNN, ENN, and ELM, the proposed model, which integrates SFE-K-Means and LSTM, obtains lower values of MAE, MSE, and MAPE. This indicates that the proposed model can obtain the best forecasting performance among the compared models. Notably, the novel data preprocessing methods (SFE-K-Means) play an essential role in improving the prediction accuracy of environmental pollutant concentration.

In summary, the Hybrid model can change the traditional passive response of air quality management and provide strong technical support for urban air pollution early warning decisions, scientific air quality management, and regional joint prevention. Further, it can improve the level of air pollution control for air environment risk prevention.

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

CN: Writing and Editing. ZN: Methodology. ZQ: Conceptualization and Writing-Reviewing. LW: Formal analysis. YL: Data curation and Visualization.

FUNDING

This work was supported by the National Social Science Foundation of China (Grant No. 17BTQ056) and the National Natural Science Foundation of China (Grant No. 72004086).

REFERENCES

Aksoy, A., Ertürk, Y. E., Erdoğan, S., Eyduran, E., and Tariq, M. M. (2018). Estimation of Honey Production in Beekeeping Enterprises from Eastern Part of Turkey through Some Data Mining Algorithms. *Pjz* 50, 2199–2207. doi:10.17582/journal.pjz/2018.50.6.2199.2207

Allen, D. P. (2009). A Frequency Domain Hampel Filter for Blind Rejection of Sinusoidal Interference from Electromyograms. *J. Neurosci. Methods* 177, 303–310. doi:10.1016/j.jneumeth.2008.10.019

Bai, Y., Li, Y., Wang, X., Xie, J., and Li, C. (2016). Air Pollutants Concentrations Forecasting Using Back Propagation Neural Network Based on Wavelet Decomposition with Meteorological Conditions. *Atmos. Pollut. Res.* 7, 557–566. doi:10.1016/j.apr.2016.01.004

Chen, L., and Pai, T.-Y. (2015). Comparisons of GM (1,1), and BPNN for Predicting Hourly Particulate Matter in Dali Area of Taichung City, Taiwan. *Atmos. Pollut. Res.* 6, 572–580. doi:10.5094/APR.2015.064

Chen, W., Zhuang, J., Yu, W., and Wang, Z. (2009). Measuring Complexity Using FuzzyEn, ApEn, and SampEn. *Med. Eng. Phys.* 31, 61–68. doi:10.1016/j.medengphy.2008.04.005

Elbayoumi, M., Ramli, N. A., and Fitri Md Yusof, N. F. (2015). Development and Comparison of Regression Models and Feedforward Backpropagation Neural Network Models to Predict Seasonal Indoor PM2.5-10 and PM2.5 Concentrations in Naturally Ventilated Schools. *Atmos. Pollut. Res.* 6, 1013–1023. doi:10.1016/j.apr.2015.09.001

Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., et al. (2018). Predicting Monthly High-Resolution PM2.5 Concentrations with Random Forest Model in the North China Plain. *Environ. Pollut.* 242, 675–683. doi:10.1016/j.envpol.2018.07.016

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., et al. (1998). The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis. *Proc. R. Soc. Lond. A* 454, 903–995. doi:10.1098/rspa.1998.0193

Li, C., and Zhu, Z. (2018). Research and Application of a Novel Hybrid Air Quality Early-Warning System: A Case Study in China. *Sci. Total Environ.* 626, 1421–1438. doi:10.1016/j.scitotenv.2018.01.195

Li, H., Wang, J., Li, R., and Lu, H. (2019). Novel Analysis-Forecast System Based on Multi-Objective Optimization for Air Quality Index. *J. Clean. Prod.* 208, 1365–1383. doi:10.1016/j.jclepro.2018.10.129

Liu, C.-W., and Kang, S.-C. (2014). "A Video-Enabled Dynamic Site Planner," in Computing in Civil and Building Engineering - Proceedings of the 2014 International Conference on Computing in Civil and Building Engineering, Orlando, Florida, June 23-25, 2014, 1562–1569. doi:10.1061/9780784413616.194

Liu, H., and Chen, C. (2020). Prediction of Outdoor PM2.5 Concentrations Based on a Three-Stage Hybrid Neural Network Model. *Atmos. Pollut. Res.* 11, 469–481. doi:10.1016/j.apr.2019.11.019

Liu, L., Breitner, S., Pan, X., Franck, U., Leitte, A. M., Wiedensohler, A., et al. (2011). Associations between Air Temperature and Cardio-Respiratory Mortality in the Urban Area of Beijing, China: A Time-Series Analysis. *Environ. Health* 10. doi:10.1186/1476-069X-10-51

Liu, Y., Zhang, S., Chen, X., and Wang, J. (2018). Artificial Combined Model Based on Hybrid Nonlinear Neural Network Models and Statistics Linear Models-Research and Application for Wind Speed Forecasting. *Sustainability* 10, 4601. doi:10.3390/su10124601

Macqueen, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations," in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, California, Berkeley, June 21-July 18, 1965, 281–297.

Pandey, A., Brauer, M., Cropper, M. L., Balakrishnan, K., Mathur, P., Dey, S., et al. (2021). Health and Economic Impact of Air Pollution in the States of India: the Global Burden of Disease Study 2019. *Lancet Planet. Health* 5, e25–e38. doi:10.1016/S2542-5196(20)30298-9

Qi, Y., Li, Q., Karimian, H., and Liu, D. (2019). A Hybrid Model for Spatiotemporal Forecasting of PM2.5 Based on Graph Convolutional Neural Network and Long Short-Term Memory. *Sci. Total Environ.* 664, 1–10. doi:10.1016/j.scitotenv.2019.01.333

Riches, N. O., Gouripeddi, R., Payan-Medina, A., and Facelli, J. C. (2022). K-Means Cluster Analysis of Cooperative Effects of CO, NO2, O3, PM2.5, PM10, and SO2 on Incidence of Type 2 Diabetes Mellitus in the US. *Environ. Res.* 212, 113259. doi:10.1016/j.envres.2022.113259

Richman, J. S., Moorman, J. R., Randall, J., and Physi, M. (2000). Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy. *Am. J. Physiology-Heart Circulatory Physiology* 278, H2039–H2049. doi:10.1152/ajpheart.2000.278.6.H2039

Shang, Z., Deng, T., He, J., and Duan, X. (2019). A Novel Model for Hourly PM2.5 Concentration Prediction Based on CART and EELM. *Sci. Total Environ.* 651, 3043–3052. doi:10.1016/j.scitotenv.2018.10.193

Tang, J. W., Lai, F. Y. L., Nymadawa, P., Deng, Y.-M., Ratnamohan, M., Petric, M., et al. (2010). Comparison of the Incidence of Influenza in Relation to Climate Factors during 2000-2007 in Five Countries. *J. Med. Virol.* 82, 1958–1965. doi:10.1002/jmv.21892

Tessum, C. W., Hill, J. D., and Marshall, J. D. (2015). Twelve-month, 12 Km Resolution North American WRF-Chem v3.4 Air Quality Simulation: Performance Evaluation. *Geosci. Model Dev.* 8, 957–973. doi:10.5194/gmd-8-957-2015

Wang, D., Wei, S., Luo, H., Yue, C., and Grunder, O. (2017a). A Novel Hybrid Model for Air Quality Index Forecasting Based on Two-phase Decomposition Technique and Modified Extreme Learning Machine. *Sci. Total Environ.* 580, 719–733. doi:10.1016/j.scitotenv.2016.12.018

Wang, P., Zhang, H., Qin, Z., and Zhang, G. (2017b). A Novel Hybrid-Garch Model Based on ARIMA and SVM for PM 2.5 Concentrations Forecasting. *Atmos. Pollut. Res.* 8, 850–860. doi:10.1016/j.apr.2017.01.003

Wu, L., and Zhao, H. (2019). Using FGM(1,1) Model to Predict the Number of the Lightly Polluted Day in Jing-Jin-Ji Region of China. *Atmos. Pollut. Res.* 10, 552–555. doi:10.1016/j.apr.2018.10.004

Wu, Z., and Huang, N. E. (2009). Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method. *Adv. Adapt. Data Anal.* 01, 1–41. doi:10.1142/S1793536909000047

Xu, Y., Yang, W., and Wang, J. (2017). Air Quality Early-Warning System for Cities in China. *Atmos. Environ.* 148, 239–257. doi:10.1016/j.atmosenv.2016.10.046

Yahya, K., Zhang, Y., and Vukovich, J. M. (2014). Real-time Air Quality Forecasting over the Southeastern United States Using WRF/Chem-MADRID: Multiple-Year Assessment and Sensitivity Studies. *Atmos. Environ.* 92, 318–338. doi:10.1016/j.atmosenv.2014.04.024

Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., and Li, F. (2021). Multi-hour and Multi-Site Air Quality Index Forecasting in Beijing Using CNN, LSTM, CNN-LSTM, and Spatiotemporal Clustering. *Expert Syst. Appl.* 169, 114513. doi:10.1016/j.eswa.2020.114513

Yan, X., and Enhua, X. (2020). "ARIMA and Multiple Regression Additive Models for PM2.5 Based on Linear Interpolation," in Proceedings - 2020 International Conference on Big Data and Artificial Intelligence and Software Engineering, ICBASE, Bangkok, Thailand, Oct. 30 2020 to Nov. 1 2020, 266–269. doi:10.1109/ICBASE51474.2020.00062

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., et al. (2019). Evaluation of Random Forest Regression and Multiple Linear Regression for Predicting Indoor Fine Particulate Matter Concentrations in

a Highly Polluted City. *Environ. Pollut.* 245, 746–753. doi:10.1016/j.envpol.2018.11.034

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control* 8, 338–353. doi:10.1016/S0019-9958(65)90241-X

Zhang, L., Lin, J., Qiu, R., Hu, X., Zhang, H., Chen, Q., et al. (2018). Trend Analysis and Forecast of PM2.5 in Fuzhou, China Using the ARIMA Model. *Ecol. Indic.* 95, 702–710. doi:10.1016/j.ecolind.2018.08.032

Zhang, S., Liu, Y., Wang, J., and Wang, C. (2019). Research on Combined Model Based on Multi-Objective Optimization and Application in Wind Speed Forecast. *Appl. Sci.* 9, 423. doi:10.3390/app9030423

Zhou, Q., Jiang, H., Wang, J., and Zhou, J. (2014). A Hybrid Model for PM 2.5 Forecasting Based on Ensemble Empirical Mode Decomposition and a General Regression Neural Network. *Sci. Total Environ.* 496, 264–274. doi:10.1016/j.scitotenv.2014.07.051

Zhu, S., Lian, X., Wei, L., Che, J., Shen, X., Yang, L., et al. (2018). PM2.5 Forecasting Using SVR with PSOGSA Algorithm Based on CEEMD, GRNN and GCA Considering Meteorological Factors. *Atmos. Environ.* 183, 20–32. doi:10.1016/j.atmosenv.2018.04.004

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.

Copyright © 2022 Niu, Niu, Qu, Wei and Li. 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 Environmental Science

ORIGINAL RESEARCH
published: 11 July 2022
doi: 10.3389/fenvs.2022.904939

Utilizing a Two-Dimensional Data-Driven Convolutional Neural Network for Long-Term Prediction of Dissolved Oxygen Content

Dashe Li [1,2,3]* *and Xuan Zhang* [1,2,3]*

[1]School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China, [2]Key Laboratory of Intelligent Information Processing, Shandong Technology and Business University, Yantai, China, [3]Co-innovation Center of Shandong Colleges and Universities, Future Intelligent Computing, Shandong Technology and Business University, Yantai, China

It is significant to establish a precise dissolved oxygen (DO) model to obtain clear knowledge ablout the prospective changing conditions of the aquatic environment of marine ranches and to ensure the healthy growth of fisheries. However Do in marine ranches is affected by many factors. DO trends have complex nonlinear characteristics. Therefore, the accurate prediction of DO is challenging. On this basis, a two-dimensional data-driven convolutional neural network model (2DD-CNN) is proposed. In order to reduce the influence of missing values on experimental results, a novel sequence score matching-filling (SSMF) algorithm is first presented based on similar historical series matching to provide missing values. This paper extends the DO expression dimension and constructs a method that can convert a DO sequence into two-dimensional images and is also convenient for the 2D convolution kernel to further extract various pieces of information. In addition, a self-attention mechanism is applied to construct a CNN to capture the interdependent features of time series. Finally, DO samples from multiple marine ranches are validated and compared with those predicted by other models. The experimental results show that the proposed model is a suitable and effective method for predicting DO in multiple marine ranches. The MSE MAE, RMSE and MAPE of the 2DD-CNN prediction results are reduced by 51.63, 30.06, 32.53, and 30.75% on average, respectively, compared with those of other models, and the R2 is 2.68% higher on average than those of the other models. It is clear that the proposed 2DD-CNN model achieves a high forecast accuracy and exhibits good generalizability.

Keywords: convolutional neural network, self-attention mechanism, dissolved oxygen, marine ranch, prediction

OPEN ACCESS

Edited by:
Yiliao Song,
University of Technology Sydney, Australia

Reviewed by:
Salim Heddam,
University of Skikda, Algeria
Rana Muhammad Adnan Ikram,
Hohai University, China

***Correspondence:**
Dashe Li
201513101@sdtbu.edu.cn
Xuan Zhang
2020410010@sdtbu.edu.cn

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Environmental Science

Received: *26 March 2022*
Accepted: *17 June 2022*
Published: *11 July 2022*

Citation:
Li D and Zhang X (2022) Utilizing a Two-Dimensional Data-Driven Convolutional Neural Network for Long-Term Prediction of Dissolved Oxygen Content. Front. Environ. Sci. 10:904939. doi: 10.3389/fenvs.2022.904939

1 INTRODUCTION

DO content of water quality, is necessary for all kinds of aquatic organisms. And changes in DO can reflect changes in the water quality of an aquaculture (Ni et al., 2019). Most fish stop feeding when the oxygen level is lower than 2 mg/L. Large numbers of fish die when the oxygen level is less than 1 mg/L. A low DO content is also a warning sign of eutrophication (Takahashi et al., 2021). To ensure the sound development of fisheries, the accurate prediction and control of DO are necessary tasks in the management of marine ranch fisheries.

Accurate water quality prediction has been challenging due to the complex effects of physical, chemical, biological, hydrometeorological and human-related processes. Some scholars have used traditional machine learning models to predict water quality. Tiyasha et al. (2021) used four types of prediction models, including a random forest (RF), to predict the DO content in the Klang River, Malaysia. Traditional machine learning techniques were applied by Valera et al. (2020) to reconstruct and predict nearshore DO concentrations in small coastal bays. Ahmed and Lin (2021) used a forest of quantile regression models to predict the DO levels in three rivers. Traditional machine learning prediction models can produce effective predictions for small sample sets with relatively simple relationships, but they fail to meet the prediction accuracy requirements for nonlinear, vaguely uncertain water quality features. In light of these problems with traditional machine learning models, the parameter optimization of traditional machine learning models is greatly influenced by human subjective factors. Some models using a meta-learning algorithm for local fine searching and pheromone dynamic updating have emerged. Liu et al. (2014) used an improved particle swarm optimization algorithm and least squares support vector regression to predict the DO content in a crab culture. Heddam and Kisi (2017) proposed an optimally pruned extreme learning machine (OP-ELM), which was newly applied to predict DO concentration with and without water quality variables as predictors. The above literature shows that this hybrid machine learning model can improve the prediction accuracy for DO and overcome the defects of traditional methods. However, its complex modeling methods and steps are still prone to falling into local minima during optimization, so the existing DO prediction models are not intelligent and must still be further improved.

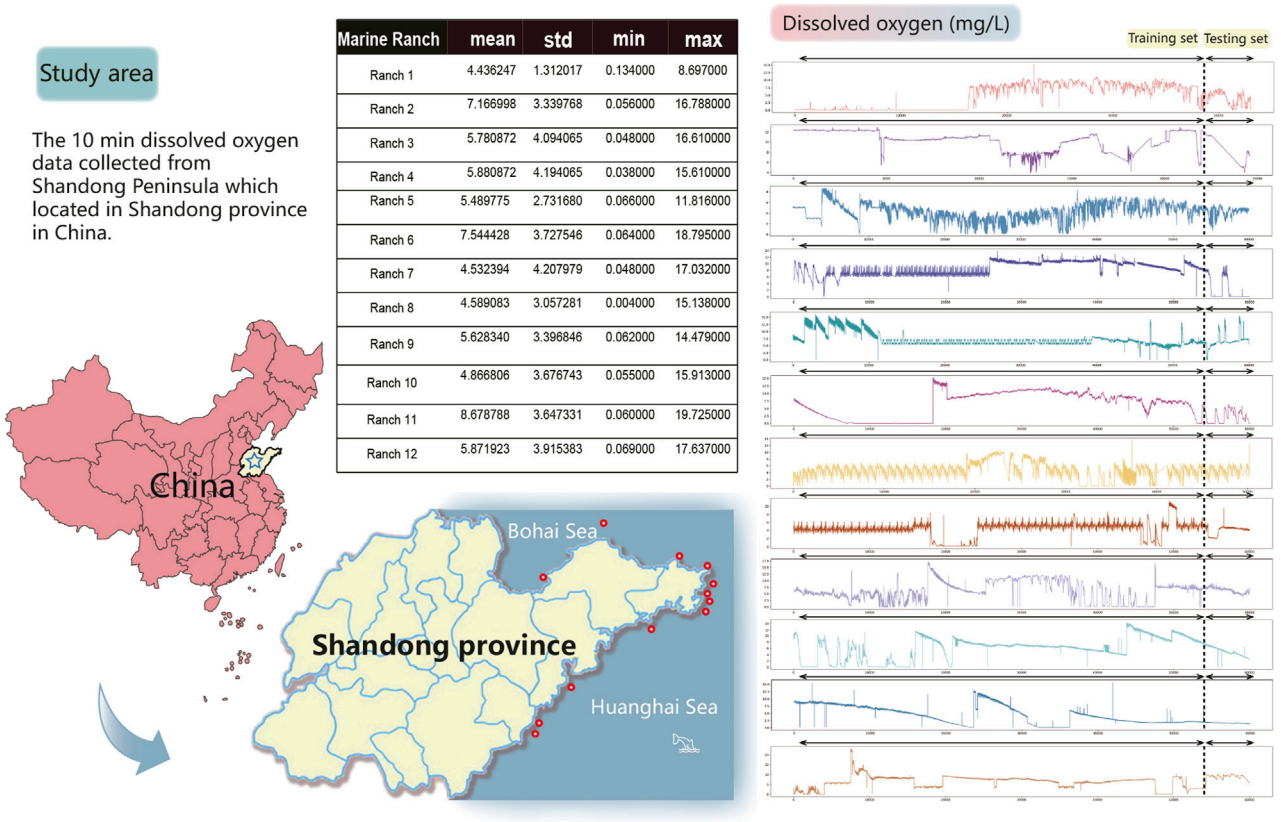
In recent years, many scholars have attempted to predict water quality using neural networks. Compared to traditional predictive models, neural network models have a high self-learning ability and excellent generalizability, allowing them to solve complex nonlinear approximation problems. These methods yield good simulation and prediction effects for trends in the water environment. Zhang et al. (2019) proposed a novel model based on multilayer artificial neural networks (MANNs) and mutual information (MI) to predict the trends of DO. proposed a new clustering-based softplus class-specific extreme learning machine to predict DO changes in time series. Rozario and Devarajan (2021) used a fuzzy C-means clustering method to construct a radial basis function neural network to predict changes in DO. Wu et al. (2018) presented a new model for DO content prediction based on a sliding window, particle swarm optimization, and error backpropagation.

The above models of water quality prediction are based on shallow networks. However, because of the small number of shallow network neurons used, the feature extraction ability of these models is not strong. And the data of some complex functions cannot be used in learning and training. Therefore, some scholars have improved the prediction accuracy of traditional models by developing deep neural network models. Zhi et al. (2021) applied long short-term memory (LSTM) to predict DO levels in several rivers. Cao et al. (2021) proposed a gradient-boosted regression tree algorithm based on an attention gate recurrent unit to predict DO levels in three dimensions. Yaqub et al. (2020) propose a long short-term memory (LSTM)-based neural network and developed to predict the ammonium, total nitrogen, and total phosphorus. Zhu et al. (2021) proposes a DO prediction model incorporating deep learning algorithms of ResNets, BiLSTM, and Attention. The LSTM mentioned above is a recurrent neural structure commonly used in sequence modeling. Compared with the traditional recurrent neural network (RNN), LSTM can alleviate gradient disappearance or explosion problems. However, due to the relatively complex internal structure, the training efficiency is much lower than that of the traditional RNN with the same computational resources, and the training is more difficult overall.

A convolutional neural network (CNN) (Kim, 2017) is a type of feedforward deep neural network containing a convolutional layer, which is composed of five structures: a convolutional layer, a pooling layer, a fully connected layer and a softmax layer. Due to its characteristics of local computation, sparse connection and weight sharing, among the available neural networks, CNNs can effectively reduce network complexity and are robust and fault tolerant. Additionally, CNNs are easy to train and optimize and have been successfully applied in many scientific fields, including computer vision (Hu et al., 2018; Luo et al., 2018), image classification (Sun et al., 2020; Pei et al., 2021), speech recognition (Haque et al., 2020; Song, 2020), natural language processing (Xiao et al., 2020; Yu et al., 2020) and others. Due to the advantages of CNNs in capturing features, they have been increasingly applied in hydrology. Khosravi et al. (2020) used CNN algorithm to develop a flood susceptibility map for Iran.Chen et al. (2020) designed an improved CNN model to establish a CNN calibration approach for the quantitative determination of water pollution with near-infrared data. Barzegar et al. (2020), Barzegar et al. (2021) improved the accuracy of forecasts achieved by a hybrid CNN LSTM deep learning (DL) model. Baek et al. (2020) used a combined CNN-LSTM model for water level and water quality prediction. Yan et al. (2021) predicted water quality using a one-dimensional residual CNN. However, most of the above studies used combined CNN models, which extracted deep features with a CNN and then used another model for prediction. These combined the modeling methods are cumbersome, and they generally adopt a one-dimensional CNN for data feature extraction; however, this approach cannot capture all the relevant spatiotemporal information.

In this paper, an improved 2DD-CNN DO prediction model is proposed. A method for converting a one-dimensional time series into a two-dimensional image is proposed. With this approach, the time dependence of the data is preserved, and the spatial characteristics are obtained. Then, we improve the two-dimensional CNN to perform regression fitting and increase the precision of the prediction model by adding an attention module. The established model strengthens the connections with DO. The main contributions of this article are as follows.

(1)A model-based data-driven two–dimensional model-based CNN is constructed, which can effectively improve the prediction accuracy of water quality parameters.



Marine Ranch	mean	std	min	max
Ranch 1	4.436247	1.312017	0.134000	8.697000
Ranch 2	7.166998	3.339768	0.056000	16.788000
Ranch 3	5.780872	4.094065	0.048000	16.610000
Ranch 4	5.880872	4.194065	0.038000	15.610000
Ranch 5	5.489775	2.731680	0.066000	11.816000
Ranch 6	7.544428	3.727546	0.064000	18.795000
Ranch 7	4.532394	4.207979	0.048000	17.032000
Ranch 8	4.589083	3.057281	0.004000	15.138000
Ranch 9	5.628340	3.396846	0.062000	14.479000
Ranch 10	4.866806	3.676743	0.055000	15.913000
Ranch 11	8.678788	3.647331	0.060000	19.725000
Ranch 12	5.871923	3.915383	0.069000	17.637000

FIGURE 1 | Research information of the original data.

(2)To solve the problem of discontinuity of feature information caused by partial missing values in the data set, a novel sequence score matching-filling (SSMF) algorithm is proposed. The historical feature sequence is used as the reference object for missing values, rendering the filled-in data more reliable.

(3)To solve the problem of discontinuity of feature information caused by partial missing values in the data set, a novel sequence score matching-filling (SSMF) algorithm is proposed. The historical feature sequence is used as the reference object for missing values, rendering the filled-in data more reliable.

(4)To resolve the problem of local perception of the convolution kernel, an attention module is added to the model to construct a CNN to capture the interdependent features of the time series.

The rest of this article is organized as follows. **Section 2** describes the study area and data sources considered in this paper and the proposed study method. **Section 3** describes the steps in establishing the 2DD-CNN prediction model. **Section 4** analyzes the model prediction performance, compares the 2DD-CNN model with other models and assesses DO data from multiple ranches. **Section 5** summarizes the study and the existing modeling problems.

2 MATERIALS AND METHODS

2.1 Study Area and Data Source

Shandong Province, China, is rich in marine resources, with a coastline length of approximately 2078 kilomiles. The national marine ranch demonstration area accounts for 40% of China, ranking first among the demonstration areas in China. This study included 12 marine ranches along the coastal waters of Shandong Peninsula (35°05′N ~ 37°50′N, 119°16′E ~ 122°42′E). The Shandong Peninsula extends into the Bohai Sea and the Yellow Sea, opposite the Liaodong Peninsula. Twelve marine ranches are spread along the coastline of Shandong Peninsula at depths of less than 656 feet. Therefore, the distribution of DO is different from that in the open ocean and is greatly affected by climate and land characteristics. The DO level changes constantly throughout the year, and the changes are complicated. The characteristics of DO and the locations of marine ranches are shown in **Figure 1**.

In the marine ranch environment, DO data are collected for 10 min, with 144 consecutive samples per day. Notably, 55,000 samples are obtained for each ranch between 2019 and 2021, including 50,000 samples that formed the training set and 5,000 that formed the test set. The same rolling prediction mechanism is used for the training and test sets.

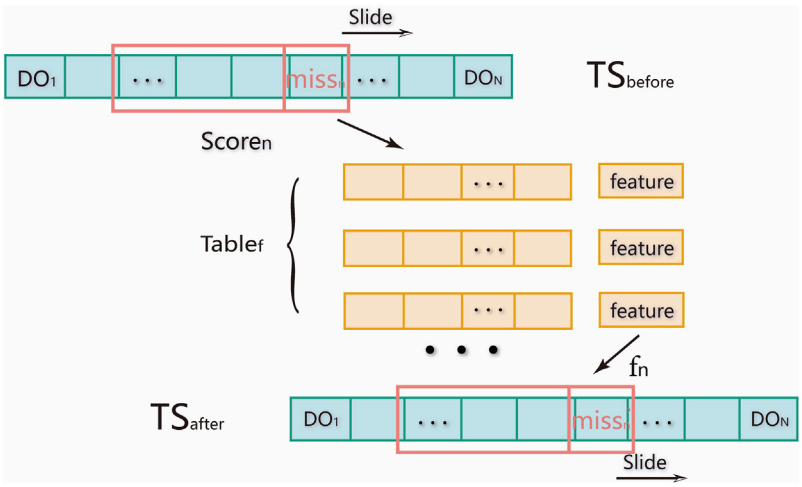


FIGURE 2 | SSMF method for filling missing values.

2.2 Data Processing Strategy

Due to the interference from sensor equipment, the environment and human factors, the collected time series contained some missing values and outliers. Poor-quality datasets containing large numbers of missing values and outliers will result in low-quality forecasting results. Data preprocessing can improve the quality of the data, thus improving the accuracy and performance of the subsequent learning process of the model (Niu and Wang, 2019).

In this paper, we identified the outliers first; then, we removed the outliers and treated them as missing values. Finally, we filled missing values based on the sequence score matching-filling (SSMF) method proposed, as shown in **Figure 2**. The SSMF approach divides the data into several sets of sequences, determines the score of each set of sequences according to the defined rules, and finds the sequence most similar to the selected sequence based on a score comparison approach, which is regarded as sequence score matching. The next feature in the matching sequence is used to fill the corresponding missing value in the selected sequence. This method is based on featurization. Historical data are used for matching, and the time series trend of historical data is used to estimate the value at the next moment to fill in the data gap. Thus, the problem of discontinuous feature information in training datasets can be avoided.

The following symbols are defined for the SSMF process:

- $TS_{before} = \{ts_1, ts_2, ts_3, \cdots ts_{N-1}, ts_N\}$: Original dataset containing missing values, where N is the sequence length;
- $TS_{after} = \{ts_1', ts_2', ts_3', \cdots ts_{N-1}', ts_N'\}$: Dataset after filling the missing values;
- L: Length of feature sequence;
- num_{Lable_m} : The number of occurrences of each category in the overall dataset, where M is the total number of categories $Lable_m = round\ (ts_n, 2)$, and $m \leq M$;
- $F\ [f_{n-S}, f_{n-S+1}, \cdots f_{n-2}, f_{n-1}]$: Features dictionary, where f is feature sequence and $n \leq N$;
- $Table_f\ [score_n]$: The feature query dictionary;

The procedure for filling missing values is as follows.

Step 1: Determine the length of feature sequence. For a time series of length N, the shortest length L of the feature sequence can be obtained according to **Eq. 1**. To minimize the number of calculations, the number of methods used per L data permutations should be greater than or equal to the total number of time series. Here, L is the length of the feature sequence before the missing values are filled. Let $miss_n$ be a missing value and n be the missing value's position in a sequence. In this paper, missing values are postprocessed from L consecutive values.

$$L = argmin_L \left(A_L^L \geq N + 1 - L\right) \tag{1}$$

Step 2: Calculates the probabilities for each category $P(label_m)$. First, each value is classified based on $Label_m = Label\ (Ts_n)$,Where, the Label function is used to divide the number of categories. Each value is regarded as a category, and two decimal places are retained. The number of occurrences of each category in the overall dataset is counted as. Then, the probability of occurrence of each category is obtained according to **Eq. 2**.

$$P(label_m) = \frac{num_{label_m}}{N}\ (m = 1, 2, \cdots M) \tag{2}$$

Step 3: Calculate the feature score of each feature sequence and establish the feature query dictionary$Table_f$. Consecutive values are regarded as a set of feature sequences denoted as $Seq = f_{n-L}, f_{n-L+1}, \cdots f_{n-2}, f_{n-1}, f_n$. The L+1 value in the feature sequence is regarded as the feature label of the sequence, e.g., $F\ [f_{n-L}, f_{n-L+1}, \cdots f_{n-2}, f_{n-1}, f_n] = f_n$. If the first L features of the feature sequence are known, the corresponding feature scores can be approximated with **Eq. 3**. All the feature scores for the original sequence are calculated, and the feature dictionary is then established as $Table_f\ [score_n] = f_{n-L}, f_{n-L+1}, \cdots f_{n-2}, f_{n-1}, f_n$, where $n \in (L + 1, N)$. **Eq. 3** is used to determine the final trend of this set of feature data.

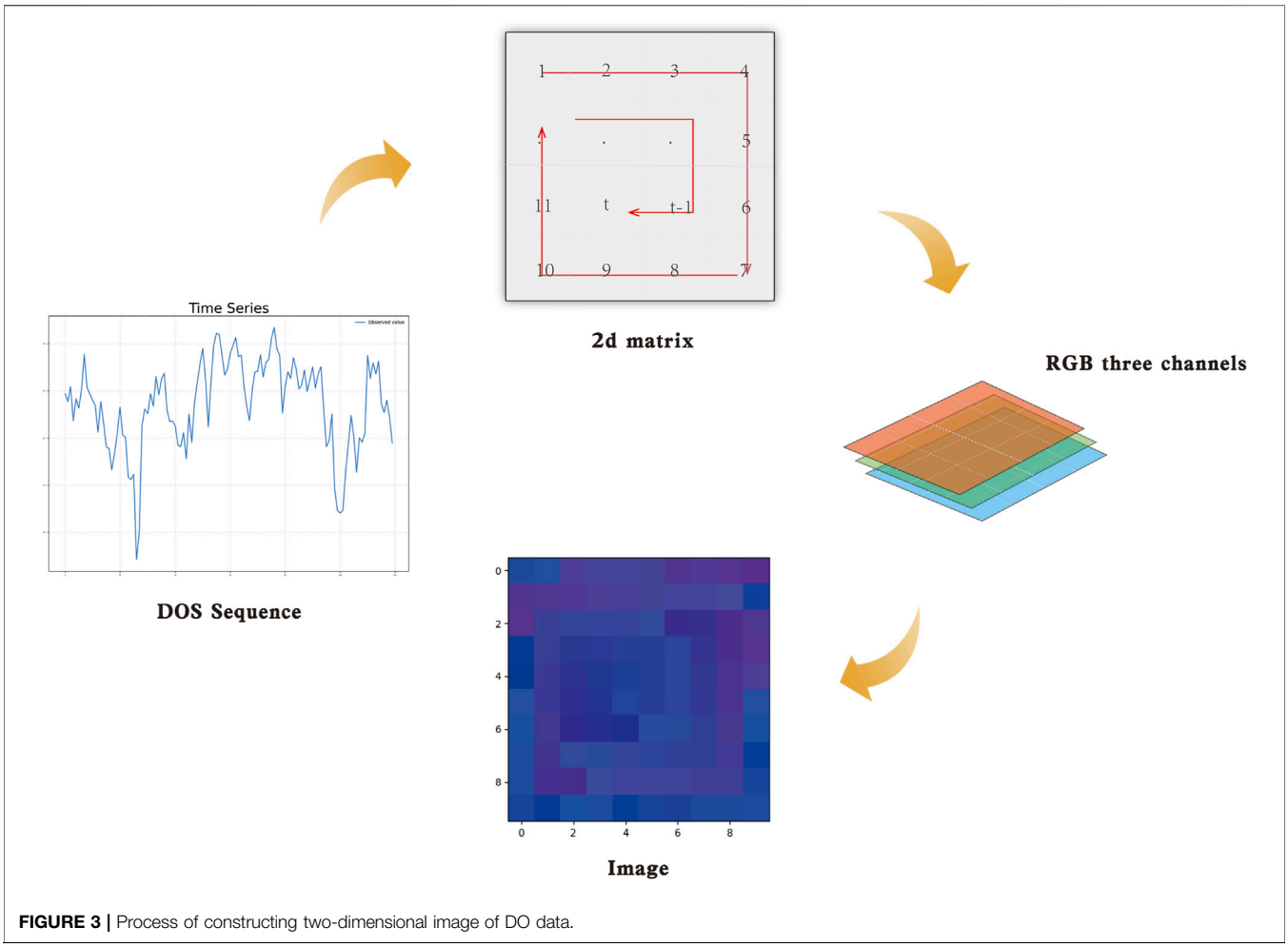


FIGURE 3 | Process of constructing two-dimensional image of DO data.

$$score_n = w\prod_{i=1}^{L} e^{(i)*P\left(Label\left(f_{n-S-1+i}\right)\right)},\ w = \begin{cases} 1, f_{n-1} - f_{n-2} \geq 0 \\ -1, f_{n-1} - f_{n-2} < 0 \end{cases} \tag{3}$$

Step 4: A sliding window is used to traverse the original sequence, and the dictionary $Table_f$ is queried to fill the missing values. SSMF fills the missing values in a data sequence. The sliding window is used to traverse the original sequence TS_{before}, and the corresponding feature score is calculated according to the L-1 values before the missing value, which are obtained from querying $Table_f$. Next, the set of features with the closest score is obtained. The consistency among the distribution characteristics of L values and missing values is assessed to find the feature that yields the highest matching score in the feature dictionary. This process is regarded as sequence feature matching, and $miss_n = F\ [Table_f\ [score_n]] = f_n$. This equation returns the processed sequence TS_{after}.

2.3 Two-Dimensional Graph of DO Data

To realize the transformation of a DO sequence from temporal dependence to spatial dependence, we must reduce the amount of redundant information in the data transformation process. Notably, here, we transform one-dimensional data convertting into two-dimensional images to match the input of the 2DD-CNN, which is used for feature extraction (Ashourloo et al., 2020). In this paper, a method of converting DO data into two-dimensional images is developed, and this approach can effectively learn the characteristics and structures of time series.

The process of constructing a two-dimensional image of DO data in this paper is shown in **Figure 3**. First, the internal rotation matrix is used to arrange the one-dimensional time series and to transform the one-dimensional time series into a two-dimensional matrix D. D is obtained according to **Eq. 4**, where k is the number of columns in the matrix.

$$D = \begin{bmatrix} D_{(1)} & D_{(2)} & . & D_{(k)} \\ D_{(4k-4)} & D_{(4k-3)} & . & D_{(k+1)} \\ . & . & . & . \\ . & D_{(k*k)} & D_{(k\times k-1)} & . \\ . & . & . & . \\ D_{(3k-2)} & . & . & D_{(2k-1)} \end{bmatrix} \tag{4}$$

Secondly, the value at each position in the two-dimensional matrix is extended to RGB three-channel form. Specifically, the

two-dimensional matrix is transformed into an RGB three-channel image. In the image, each pixel represents a value. According to the color in the image, the overall distribution of pixel values can be intuitively assessed. All values can be uniformly expressed with different colors. The data are stored without loss in three RGB channels. RGB is the color of red, green and blue channels. The R, G, and B channels have 256 levels. Red, green and blue are abbreviated as R,G,B respectively. The brightness of the R, G, and B channels ranges from 0 to 255. Based on these characteristics, this paper divides all the data n into three sets include n_1, n_2 and n_3, $n = n_1 + n_2 + n_3$, where $n_1 \in (0, 99)$ represents R, $n_2 \in (0.00, 0.99)$ represents G,$n_3 \in (0.0000, 0.0099)$ represents B. To maximize the visual weight of the first channel, the values in the first set are uniformly filled in the interval of $(0, 255)$. The value output by the first channel is R_n:

$$R_n = \lceil \frac{255}{max(n_1)} \rceil \times n_1 \tag{5}$$

The value output by the second channel is G_n:

$$G_n = n_2 \times 100 \tag{6}$$

The value output by the third channel is B_n:

$$B_n = n_3 \times 10000 \tag{7}$$

Finally, the PIL image processing library is used, and the resulting image is stored in png lossless format.

2.4 Configurations of the Designed CNN

The CNN constructed in this paper consists of three parts, namely, a convolutional layer, a pooling layer and a fully connected layer (Jiang et al., 2020; Yang et al., 2021). A linear weighted filter with a local receptive field, namely, a convolution layer, is alternately applied with the pooling layer to sample the extracted features. The fully connected layer distributes the data according to a nonlinear function. The calculation process involving these layers is as follows.

(1) The convolution layer provides local calculations and sparse connections and applies weight sharing. The convolution operation process is represented by **Eq. 8**, which W_i represents the original matrix and y_i represents the convolutional kernel. F is the convolution operation and retention factor of the characteristic matrix. The feature map first performs the F convolution operation and then adds b_i. The convolutional kernel is slid according to the padding threshold until the entire feature graph has been obtained.

$$Z_{i+1} = F(W_i, y_i) + b_i \tag{8}$$

(2) The pooling layer does not contain parameters; it performs feature selection based on the matrix. The pooling layer retains important features to reduce the number of subsequent operations and avoid overfitting. Common pooling layer operations include maximum pooling and average pooling.

(3) The fully connected layer reshapes the output tensor of the pooling layer into a one-dimensional vector and then maps it to a sample label pool.

The goals of the convolution layer and pooling layer in the CNN are to extract features and reduce the number of computations. One-dimensional vectors are predicted with the fully connected layer. The overall architecture of the CNN built in this paper is shown in **Figure 4**. This architecture contains two convolutional layers, two pooling layers and two fully connected layers. The input image is transformed into an input feature matrix. This matrix is then passed through the convolution and pooling layers and then transformed into one-dimensional data before being passed to the fully connected hidden layer and fully connected output layer. Finally, the prediction result is output. In this paper, the CNN architecture shown in **Figure 4** is constructed. No upper limit is set for the input window, and the lower limit of the input window is a 4×4 matrix.

The first convolutional layer adopts a 32-layer convolution core of 3×3. The second pooling layer adopts a maximum pooling core of 2×2. The third convolutional layer uses 64 layers of size 3×3. The fourth pooling layer also adopts 2×2 maximum pooling. The final layer of the feature graph is flattened to connect the fully connected hidden layer to a one-dimensional vector. The final result is output with the fully connected output layer. The activation function used in the middle layer of the proposed model is a ReLU function. Unlike the traditional classification model, the proposed model does not use an activation function in the final layer, which is used to directly output the final results.

3 DISSOLVED OXYGEN PREDICTION MODEL

In this paper, a 2DD-CNN is proposed to predict the DO level in marine ranches. The model prediction process is divided into four steps: data preprocessing, constructing the two-dimensional graph of DO data, applying an self-attention module and implementing the 2DD-CNN prediction framework. The process of 2DD-CNN model prediction is shown in **Figure 5**.

Step 1: Data Preprocessing. The collected DO sequence contains some missing values and outliers. First, the σ principle is used to identify the outliers. This principle can be used to identify low-probability events outside the standard normal distributed interval $(u - 3\sigma, u + 3\sigma)$. Such values should be removed and regarded as missing values. Second, the SSMF algorithm proposed previously is used to fill the missing values. If we find that sequence *Seq* has a similar score to the feature sequence *Seq*' for the target missing value $miss_n$. Then the next value in the similar sequence is used to fill the missing position $miss_n$ in the target sequence.

Step 2: Encoding Time Series to Images. This step converts a DO sequence into an image. First, the DO sequence is transformed into a two-dimensional matrix by internal rotation. Then, the two-dimensional matrix is mapped to RGB channels. Thus, the transformation of the DO sequence from temporal dependence to spatial dependence is realized.

Step 3: Self-attention Module. To solve the problem, the prediction effect is limited by the local perception of the CNN convolution kernel, the global receptive field is added. In this step, an self-attention mechanism (Wang et al., 2021) is established,

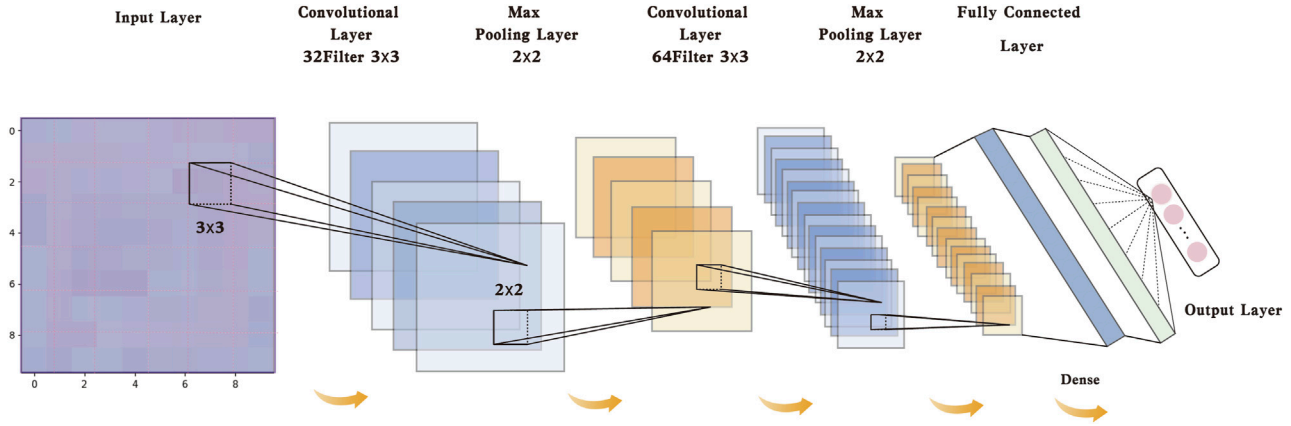


FIGURE 4 | Convolutional neural network structure.

and the self-attention module is built before the CNN model (Vaswani et al., 2017; Wang et al., 2018) to mine the influence weight of the information at each position in the matrix based on the available prediction results; then, a new weighting matrix is constructed. The correlation between values is calculated by matrix multiplication. Then, these correlation scores are combined to obtain a weighting matrix. The specific steps are as follows.

Firstly, three 1 × 1 convolution kernels are defined as W_q, W_k, W_v. These kernels are established with the original image to obtain three feature maps, expressed as $Q(x)$, $K(x)$and$V(x)$. As shown in **Eq. 9**, these three feature maps have a triplet structure.

$$Q(x) = W_q x, K(x) = W_k x, V(x) = W_v x \tag{9}$$

Secondly, obtaining the attention map $Score'_{i,j}$ based on the matrix dotted product of $K(x)$ and $Q(x)$, as shown in **Eq. 10**, is obtained with the softmax function. Finally, as expressed in **Eq. 11**, the self-attention feature map obtained through the dot product of $Score'_{i,j}$, and $V(x)$ is $Attention(x)$. $Attention(x)$ is used as the input of the 2DD-CNN. At this time, our model captures the characteristics of the input matrix considering the corresponding weights.

$$Score'_{i,j} = \frac{exp\left(Score_{i,j}\right)}{\sum_{i=1}^{N} exp\left(Score_{i,j}\right)}, \ where \ Score_{i,j} = Q(x_i)^T V\left(x_j\right) \tag{10}$$

$$Attention(x) = Score'_{i,j} V\left(x_j\right) \tag{11}$$

The partial perception of the CNN convolutional kernels results in each kernel only calculating area information. As the neural network layers deepen, the convolutional kernel region information is limited to only one area. Thus, the regions outside of the convolutional kernel area are not considered, and the effect of prediction is limited. Thus, adding the self-attention module to the model is a good way to solve this problem.

Step 4: 2DD-CNN Prediction Network Framework. First, the number and order of convolutional and pooling layers are determined, and the detailed structure is shown in **Figure 4**. Second, the ratio of training data to verification data is set as 16:1, and the trained back-propagation 2DD-CNN is used. The initial values of the weight and bias parameters in the input layer and output layer of 2DD-CNN are set. Then, the input dataset size and output dataset size are determined according to the 2DD-CNN features. Finally, the model is optimized by the root mean square prop (RMSProp) algorithm. Based on repeated experimental analyses, the prediction effect based on the 2DD-CNN is the best when the model parameters are learningrate = 0.02, batchsize = 15 and epochs = 100. The trained model is used in the prediction of DO levels.

4 RESULTS AND DISCUSSION

To verify the model proposed in this paper, the 2DD-CNN is used to predict the oxygen sequences for 12 ranches. A comparison experiment with other algorithms and a generalization experiment involving multiple ranches are performed. The *Python* 3.8 language is used in the experiments, and the hardware included an Inter(R)Core(TM) i5-8265 CPU at3.30 GHz and 8 GB memory.

4.1 Performance Criteria

To verify the excellent performance of the 2DD-CNN model and analyze the errors between the predicted and observed values of the model, this paper applies five measurement indices: the mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and coefficient of determination (R^2). The corresponding mathematical expressions are given in **Equations 12–14**, **Equation 15** and **Equation 16**, where $\hat{o}_i$ is the predicted values and o_i is the observed values. The MSE is generally used to detect the deviation between the predicted values and the observed values of the model. It calculates the sum of squares of the distance between the predicted values and the observed values. The quadratic form is convenient for derivation, so it is often used as a loss function in linear. The MAE is the absolute value of the difference between the observed values and the

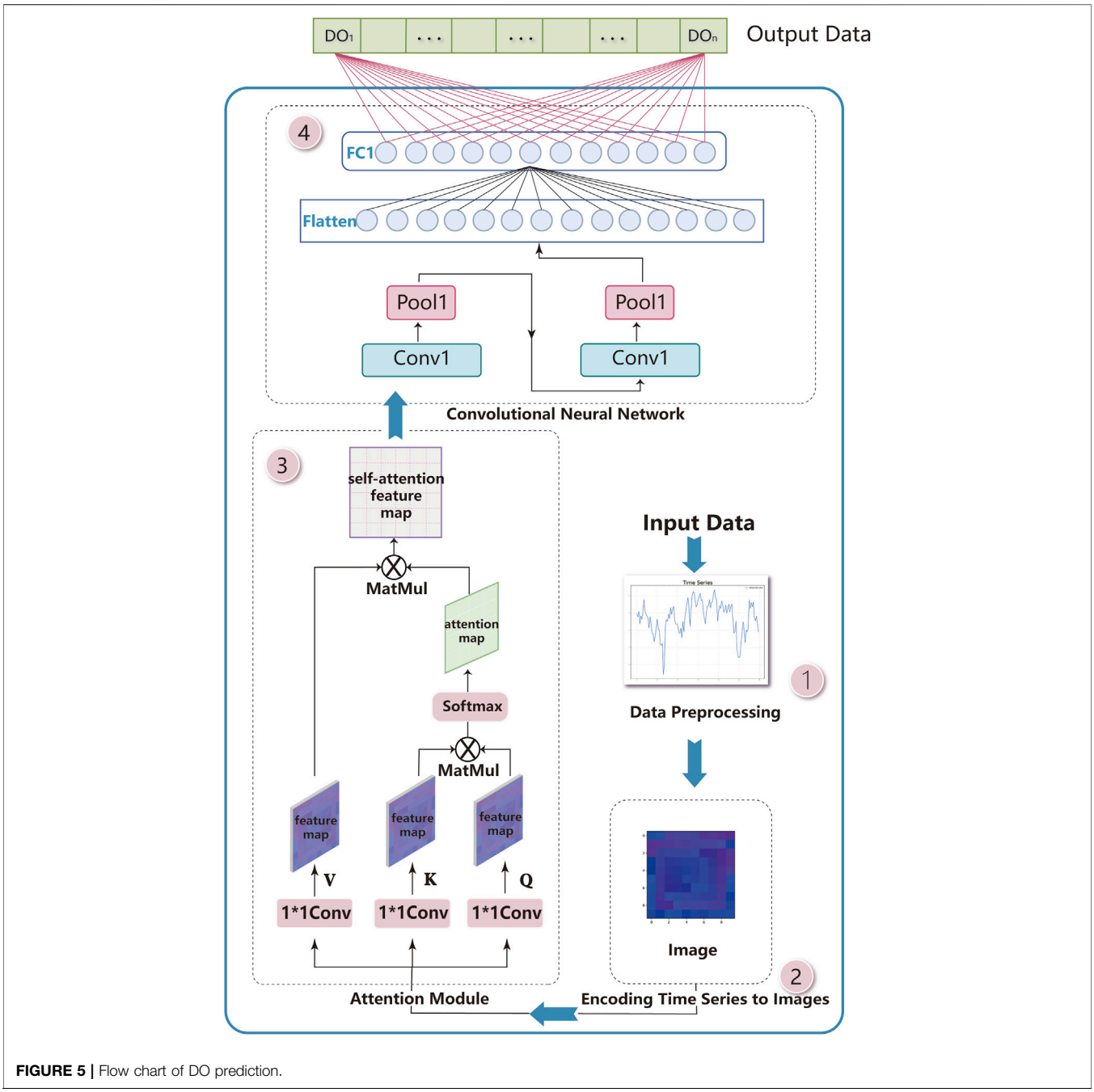


FIGURE 5 | Flow chart of DO prediction.

predicted values. It is less affected by outliers (outliers separated from the sample distribution) and can better reflect the actual situation of the predicted values error. The RMSE adds the square root sign on the basis of the MSE, which is more intuitive in comparison. The range of MAPE is [0, + ∞), MAPE is expressed as a percentage, for which 0% indicates perfect model, while a value greater than 100% indicates an inferior model. Further, MAPE is easier to explain. R^2 is indicator used to evaluate the quality of fitting. The lower the MSE, MAE, RMSE and MAPE values are, the smaller that the prediction error is. Additionally, the higher that the R^2 value is, the better the fit is between the predicted and observed values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{o}_i - o_i)^2 \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{o}_i - o_i| \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{o}_i - o_i)^2}, \tag{14}$$

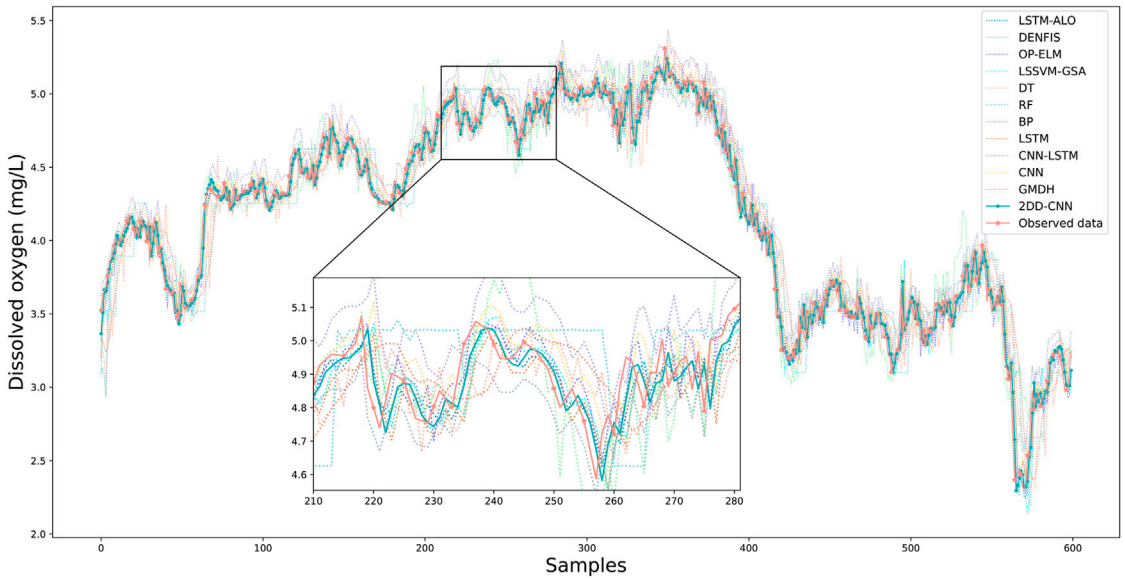


FIGURE 6 | The DO content prediction results.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{o}_i - o_i}{o_i}\right|, \tag{15}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{o}_i - o_i)^2}{\sum_{i=1}^{n}(\bar{o}_i - o_i)^2} \tag{16}$$

4.2 Single Ranch Prediction Evaluation With Other Prediction Models

In this case, 55,000 DO values from the Luhaifeng marine ranch in Qingdao are used as an example, Data from other Marine ranches were analyzed in the same way as the focus of the study. The data are divided into a training set and test set at a 10:1 ratio. Under the same conditions, the 2DD-CNN is compared with other models. The comparison model includes a CNN, an LSTM model, CNN-LSTM, a back Propagation Neuron Network (BP) (Zhang and Lou, 2021), a decision tree (DT) (Anmala and Turuganti, 2021), a RF (Karijadi and Chou, 2022), dynamic evolving neural fuzzy inference system (DENFIS) (Adnan et al., 2019), a group method of data handling (GMDH) neural networks (Adnan et al., 2020) a hybrid model based on long short-term memory neural network and ant lion optimizer (LSTM-ALO) (Yuan et al., 2018), a hybrid model based on an optimally pruned extreme learning machine (OP-ELM) and a hybrid model based on the least squares support vector machine and gravitational search algorithm (LSSVM–GSA) (Zeng et al., 2021).

Figure 6 shows the prediction results based on 600 observations from the test set and various prediction models. The 2DD-CNN model and other models exhibit good performance in predicting trends. For comparison, we enlarge part of the plot of 71 data points showing the 2DD-CNN predictions. Notably, among all the predictions, these values are closest to the observed values. Specifically, the lag of the neural network prediction model is obvious. The prediction effects of the LSTM model and the LSTM's hybrid model are second to that of the 2DD-CNN. The results in **Figure 6** show that the prediction effect of the 2DD-CNN is generally superior to that of the other models.

TABLE 1 | Comparison of evaluation indexes of model prediction error.

Model	*MSE*	*MAE*	*RMSE*	*MAPE*	R^2
2DD-CNN	**0.007595**	**0.065866**	**0.087150**	**1.653,131**	**0.983,616**
CNN	0.015655	0.095538	0.125,120	2.359,237	0.966,227
CNN-LSTM	0.028474	0.138,715	0.168,743	3.268,304	0.942,419
LSTM	0.014010	0.089261	0.118,362	2.241,713	0.967,440
BP	0.008812	0.080576	0.093872	2.039655	0.974,442
GMDH	0.042199	0.150,444	0.205,425	3.866,152	0.906,206
RF	0.043860	0.155,353	0.209,427	4.014477	0.908,870
DT	0.036191	0.140,532	0.190,240	3.637,882	0.922,261
LSTM-ALO	0.018542	0.082875	0.136,168	2.168,491	0.975,121
DENFIS	0.016854	0.087585	0.129,822	2.214,271	0.963,552
OP-ELM	0.018616	0.082947	0.136,439	2.172,834	0.975,097
LSSVM-GSA	0.015000	0.090607	0.122,474	2.323,732	0.967,818

The bold values are the result of the 2DD-CNN model.

The experimental results of all models are further verified by comparing the corresponding evaluation indics MSE, MAE, RMSE, MAPE, and R^2. The results are shown in **Table 1**. The predicted values of the 2DD-CNN display the lowest MSE, MAE, MAPE and RMSE and the highest R^2, corresponding to the lowest prediction error. Compared with LSTM, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 45.8, 26.21, 26.4 and 26.3%, respectively. Compared with BP, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 13.8, 18.3, 7.2 and 18.95%, respectively. Compared with RF, the 2DD-CNN reduces the MSE, MAE,

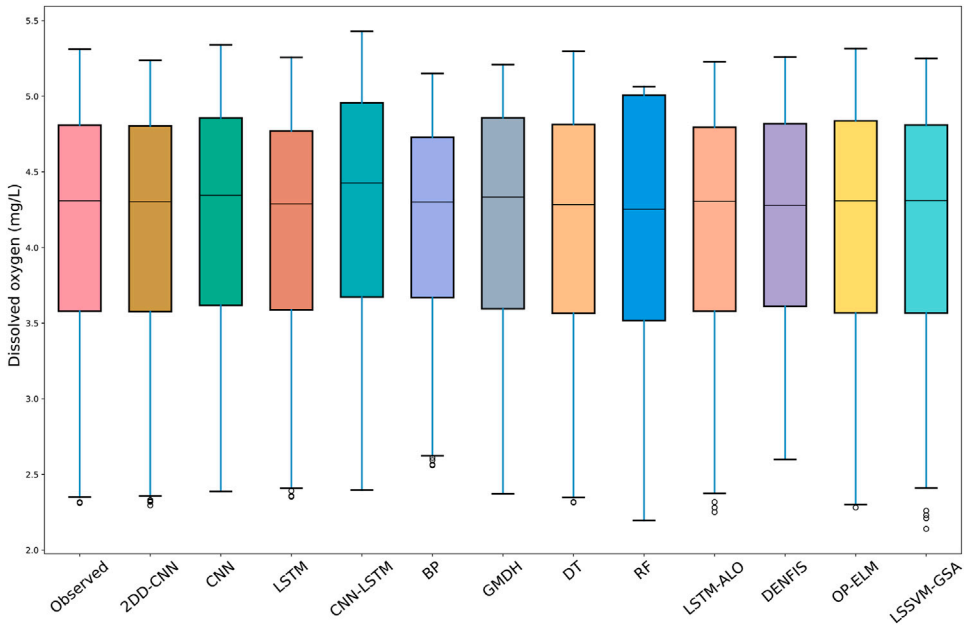


FIGURE 7 | Comparison of predicted and observed values of DO for different models.

RMSE and MAPE of the predictions by 87.7, 57.6, 58.4 and 58.8%, respectively. Compared with DT, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 79.0, 53.13, 54.19 and 54.56%, respectively. Compared with DENFIS, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 54.94, 24.8, 32.87 and 25.34%, respectively. Compared with GMDH, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 82.00, 56.22, 57.58 and 57.24%, respectively. Compared with the traditional CNN prediction model, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 51.49, 31.06, 30.35 and 29.9%, respectively. Compared with CNN-LSTM, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 73.33, 52.52, 48.35 and 49.42%, respectively. Compared with LSTM-ALO, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 59.04, 20.52, 36.00 and 23.77%, respectively. Compared with OP-ELM, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 59.20, 20.59, 31.13 and 23.92%, respectively. Compared with LSSVM-GSA, the 2DD-CNN reduces the MSE, MAE, RMSE and MAPE of the predictions by 49.37, 27.31, 28.84 and 1.63%, respectively. On average, the MSE of predictions obtained with the 2DD-CNN is 51.63% lower than that obtained with other models, the MAE is 30.06% lower, the RMSE is 32.53% lower, the MAPE is 30.75% lower and the R^2 is 2.68% higher. From the comparison of the results, the 2DD-CNN model performs significantly better than the other models in predicting DO levels. Additionally, the existing CNN prediction models performed worse than the studied LSTM and BP models. Nevertheless, the prediction performance of the improved 2DD-CNN model is better than that of all of the other models. In summary, improving the CNN to establish the 2DD-CNN model proposed in this paper yields a significant improvement in the accuracy of DO prediction.

Figure 7 shows a box plot of the predicted and observed DO values for eleven models in individual marine ranches. The 2DD-CNN predictions are similar to the observed values overall. The other model results differ from the observed values based on the upper quartile, mean, maximum and minimum values. The upper limits of the predicted values of the LSTM model and DT model are similar to the upper limit of the observed values. However, the values obtained with the 2DD-CNN are most similar to the observed values based on the mean value, upper and lower quartiles and lower limit. The BP model, RF model and DENFIS model results largely differed from the observed values based on the upper and lower quartiles and the upper and lower limits. The range of predicted values of the BP model is smaller than the range of observed values, with predictions concentrated near the mean value; this result indicates that the prediction of maximum and minimum values by the BP model is not accurate. The prediction range of the RF model exceeds that of the observed values, and the prediction of extreme values is inaccurate. The prediction range of RF model exceeds the observed values, and the prediction of extreme values is also inaccurate. The DENPFIS model is not very accurate in predicting the results at lower values. Compared with the observed values, the values predicted by the CNN model and CNN-LSTM model moved upward overall. The data of the predicted values and observed value of the hybrid model had great similarities on the whole, but some outliers appeared in the model prediction, which might have been due to model overfitting. In conclusion, the prediction accuracy of the BP model, RF model and DT model is not good. Other CNN

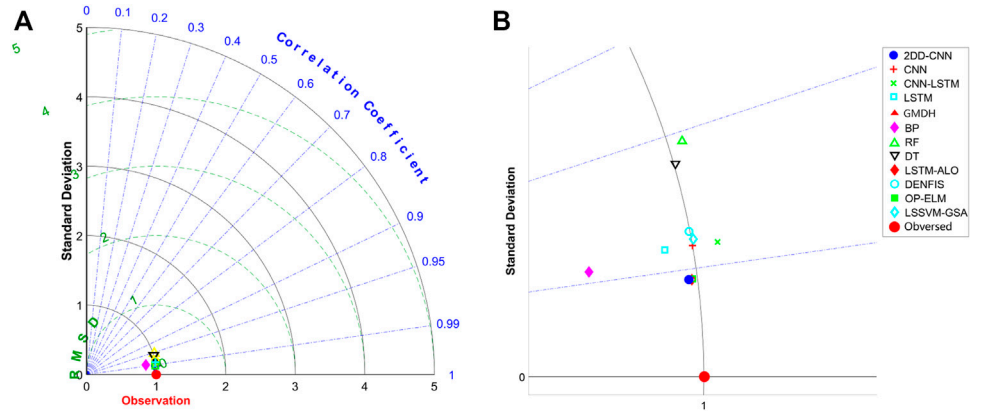


FIGURE 8 | Comparison of predicted and observed values of DO for different models. **(A)** is a whole Taylor diagram, and **(B)** is a partially enlarged Taylor diagram.

models, the CNN-LSTM model and the LSTM-ALO model have certain deviations in prediction, and the models must be adjusted. The DENFIS, GMDH, OP-ELM, and LSSVM-GSA models have poor prediction effects on some outliers and edge values. The data distributions of the 2DD-CNN-predicted values and observed values are very close.

Figure 8 shows the Taylor diagram of the performance of ten models. The scattered dots in the figure represent the model, the radiating lines represent the correlation coefficients, the horizontal and vertical axes represent the standard deviations, and the semicircular dotted lines represent the RMSE. **Figure 8A** shows the whole part of the Taylor diagram, and 8 B shows the enlarged display of eleven model indicators. For the prediction of DO, the discrete correlation coefficient, standard deviation and RMSE proposed in this paper are 0.991,817, 0.988,898 and 0.1280, respectively, and the prediction result is the best. Based on the observed values, the closer that a value is to the red dot representing the observed value in the Taylor diagram, the better that the prediction performance is. In summary, the prediction of DO proposed in this paper is the best.

To further analyze the prediction ability of the models, the prediction results for 144 samples each day with different models are presented in the form of line charts and density correlation graphs. **Figure 9** shows the values predicted by each model compared to the observed values. In **Figure 9**, the data predicted by the model proposed in this paper are closest to the observed values, and the prediction effect of peaks and valleys is the best among all models. We analyze the correlation between the predicted and observed values based on the density correlation plot in **Figure 9**. In this figure, the closer that a scatter point is to the light-colored dotted line, the closer that the predicted values is to the corresponding observed value. The line fit based on the scatter point is the dark dotted line.

This outcome can be clearly observed in the figure. The graph in **Figure 9A** shows that the predicted values of 2DD-CNN is almost the same as the observed DO, indicating good performance. The graph on the right of **Figure 9A** shows that all points are located near the straight line, and the linear regression line of these points almost covers the straight line. This outcome shows that 2DD-CNN can predict DO data with high accuracy. **Figure 9B** shows the prediction results of the CNN. Compared with the curves in **Figures 9A,B**, the predicted values of the CNN are more difference from the observed values than are predicted values of 2DD-CNN. Compared with the right figure of **Figures 9A,B**, the fitted line is farther from the y = *x* line, and the point dispersion is greater. This outcome indicates that the prediction effect of the CNN model is inferior to that of 2DD-CNN. The model in this paper is an improvement on the CNN model. Compared with the CNN model, the self-attention module is added to the model in this paper, and two-dimensional convolution is adopted. The results show that the improved model improves the prediction accuracy of DO. In the line chart in **Figure 9C**, the values predicted by the CNN and CNN-LSTM models exhibit obvious variations in positions compared with the positions of the observed values. These differences are clear in the density correlation diagram in **Figure 9D**, which shows the predicted value of the BP model. As a shallow neural network, the BP has the characteristics of a simple structure. However, due to limited neurons and shallow networks, the accuracy is not as good as the predicted value of the deep neural network in the fitting experiment of complex DO trends. In **Figure 9E**, the prediction accuracy of the LSTM model and LSTM's hybrid are second only to that of the 2DD-CNN. As an RNN model, LSTM can meet most accuracy requirements, but the training efficiency is not high due to its relatively complex internal structure. There is some deviation between the predicted and observed values in the line chart and density correlation diagram. The values predicted by the BP, DT and RF models in **Figure 9F**, **Figure 9G** and **Figure 9H**, respectively, deviate from the observed values, and the prediction of peaks and valleys is not accurate. Additionally, the RF can only predict the general trend of the DO. Two traditional machine learning models, DT and

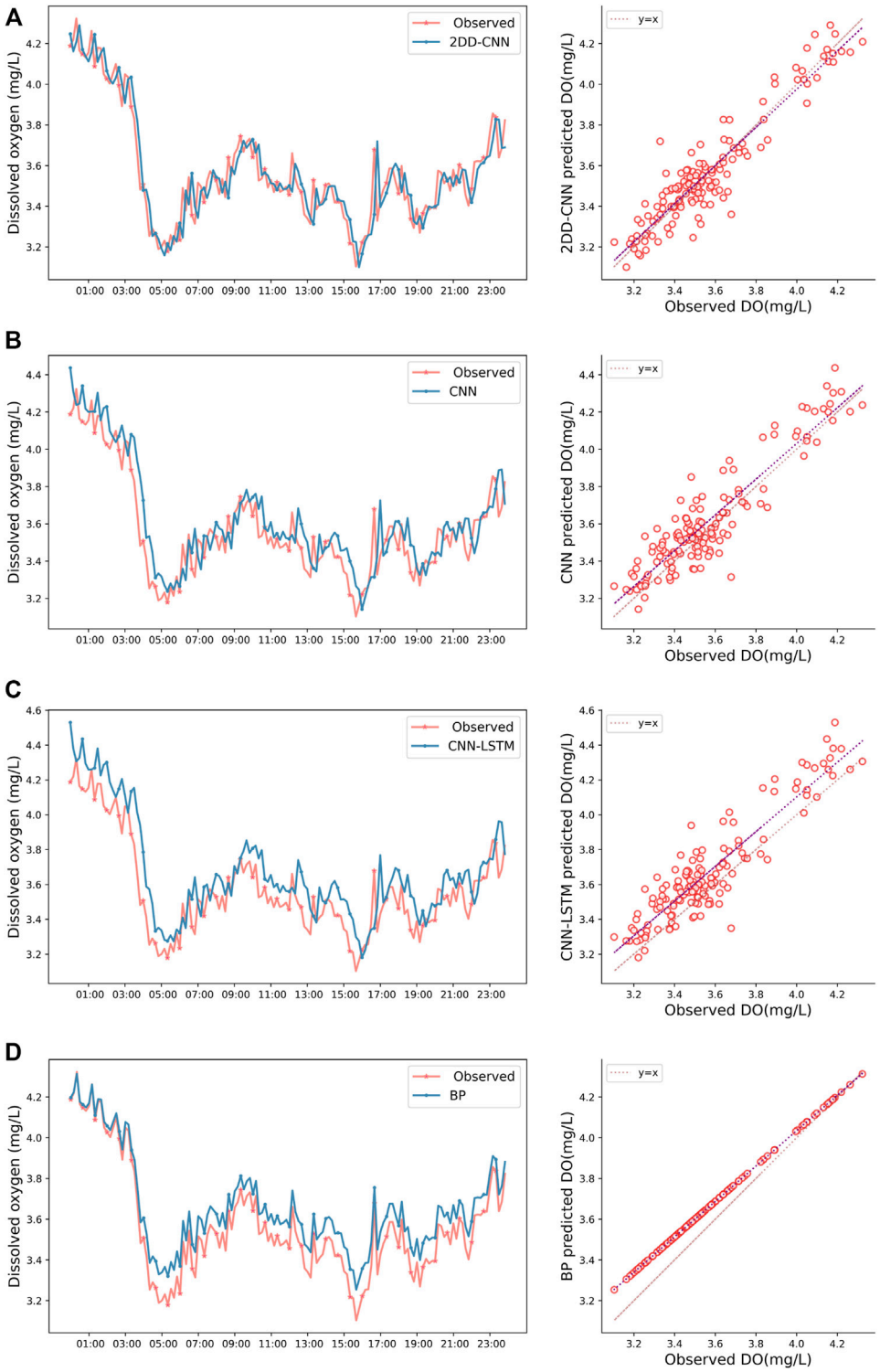


FIGURE 9 | Comparison of predicted and observed values of DO for multiple models. **(A)** is the result of 2DD-CNN. **(B)** is the result of CNN. **(C)** is the result of CNN-LSTM. **(D)** is the result of BP. **(E)** is the result of LSTM. **(F)** is the result of DT. **(G)** is the result of RF. **(H)** is the result of DENFIS. **(I)** is the result of LSTM-ALO. **(J)** is the result of OP-ELM. **(K)** is the result of LSSVM-GSA. **(L)** is the result of GMDH.

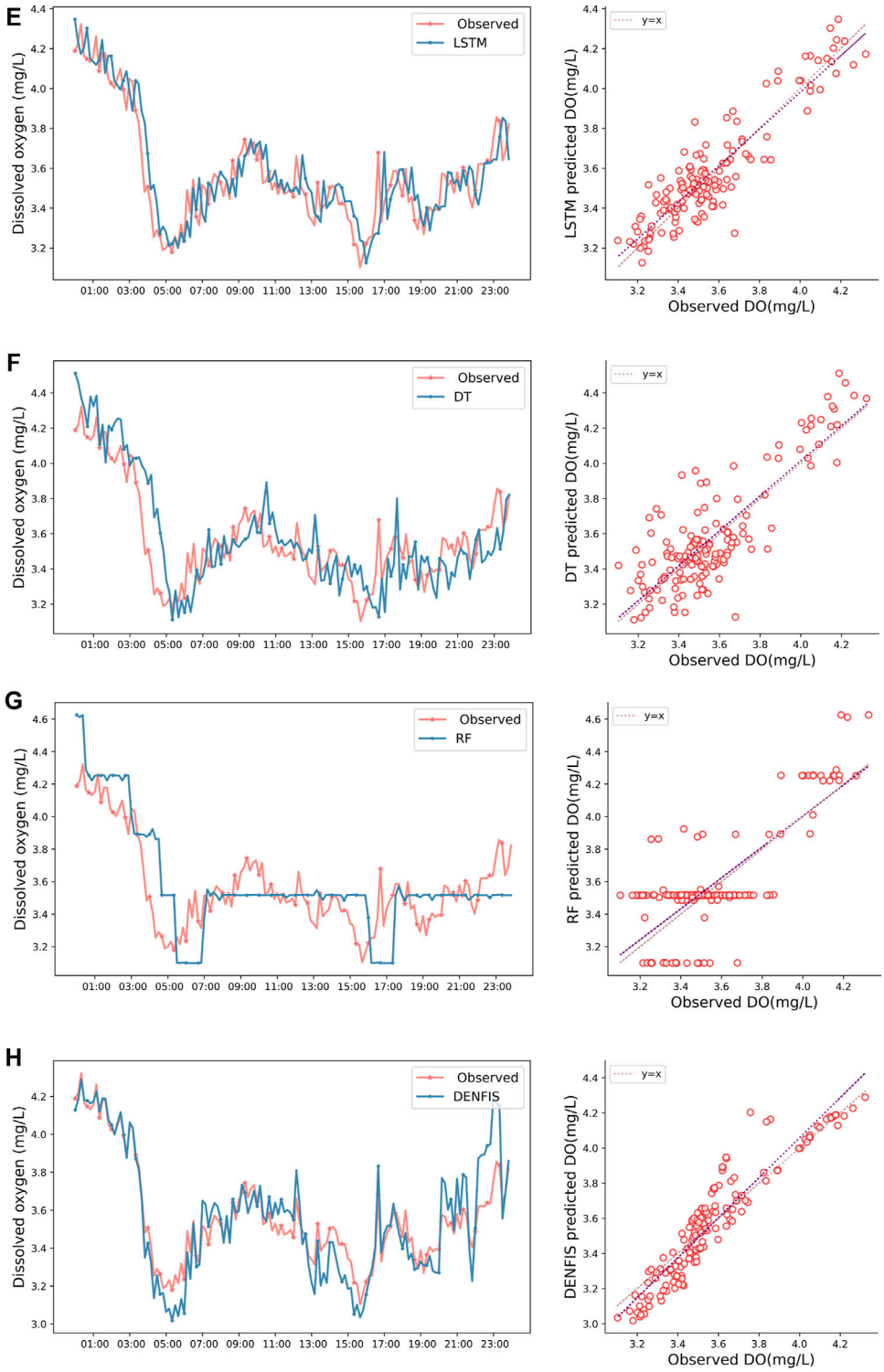


FIGURE 9 | (Continued).

RF, perform poorly in fitting complex nonlinear DO data. **Figure 9H** shows the predicted value of the DENFIS model. DENFIS performs poorly in predicting the peak value, showing a situation of amplifying the peak value and underestimating the value. In the density correlation diagram, it can also be observed that the predicted values

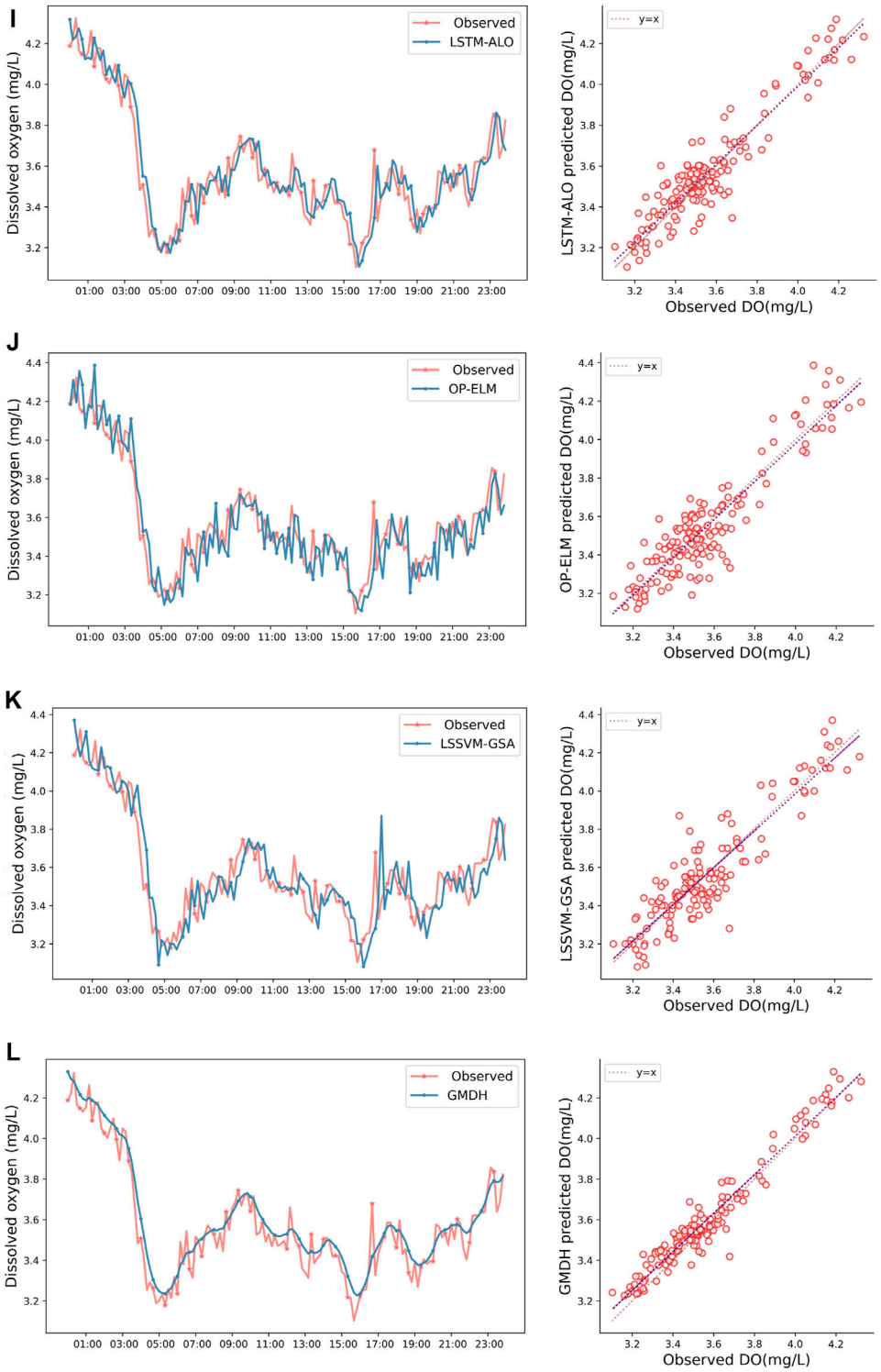


FIGURE 9 | (Continued).

of the DENFIS model deviate greatly from the observed values compared with the medium-high value segment and the low-value segment. As a mathematical fuzzy inference model, DENFIS is not as effective as a deep learning model in predicting DO. **Figures 9I,J,K** show the prediction effects of the LSTM-ALO, OP-ELM and LSSVM-GSA models,

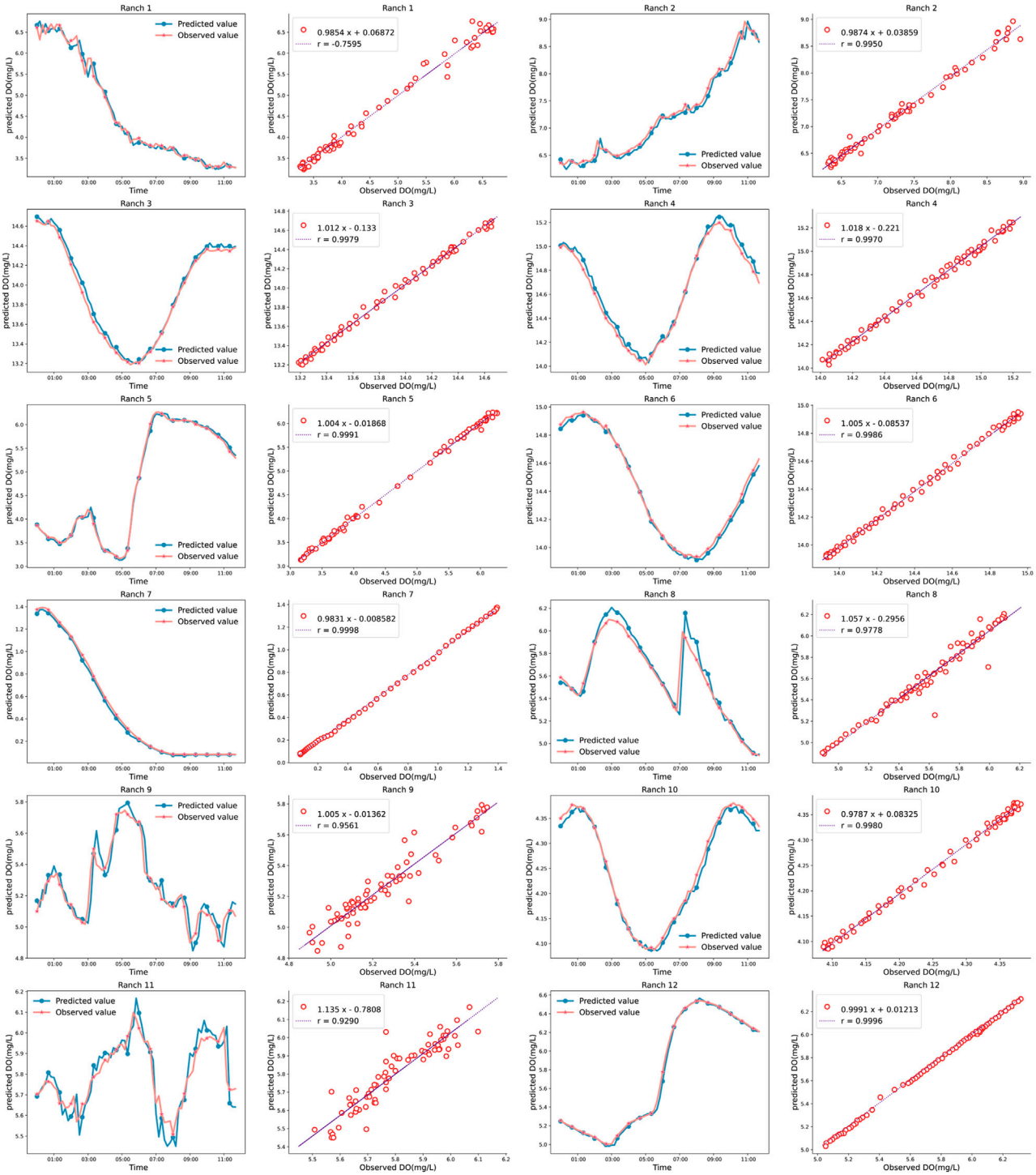


FIGURE 10 | Application of the model to data from different ranches.

respectively, in comparison with the hybrid model. The parameters of the deep learning LSTM model, shallow ELM neural network, GMDH and LSSVM-GSA machine learning model were optimized using a metaheuristic algorithm. The prediction results showed good performance with slight errors. Although the model parameters were optimized to a large extent, the structure of the model remained unchanged. The prediction accuracy is still limited by the model itself, which is not as accurate as the model proposed in this paper. The operation process of the hybrid model is complicated.

TABLE 2 | Prediction error evaluation indexes of 12 ranch models.

Ranch Number	Marine ranch	*MSE*	*MAE*	*RMSE*	*MAPE*	R^2
Ranch 1	Qingdao Luhaifeng National	0.010691	0.075311	0.103,396	1.854,217	0.983,372
Ranch 2	Xixiakou Group National	0.009136	0.074883	0.095581	1.026838	0.985,410
Ranch 3	Rongcheng Hongtai Fishing	0.001645	0.033601	0.040560	0.240,704	0.994,178
Ranch 4	Ryongcheng Broussonetia Ranch	0.002021	0.037824	0.044960	0.270,987	0.992,899
Ranch 5	Changdao Xiangyu Reef Casting	0.002888	0.032824	0.053741	0.770,817	0.990,028
Ranch 6	Weihai Yutai Fishing	0.000824	0.022561	0.028698	0.171,321	0.998,891
Ranch 7	Rongcheng Swan Lake Fishing	0.000063	0.004945	0.007967	4.348,656	0.998,488
Ranch 8	Rizhao Aquatic Group Reef Casting	0.039713	0.055784	0.199,281	1.567,903	0.971,123
Ranch 9	Rongcheng Yandunjiao	0.003290	0.034221	0.057354	1.182,211	0.998,119
Ranch 10	Rongcheng Chengshan Hongyuan	0.000086	0.007644	0.009270	0.185,958	0.997,184
Ranch 11	Rizhao Wanbao Fishing	0.005856	0.042857	0.076525	0.708,804	0.976,465
Ranch12	Shandong Oriental Ocean National	0.000426	0.013398	0.020630	0.314,491	0.999,757

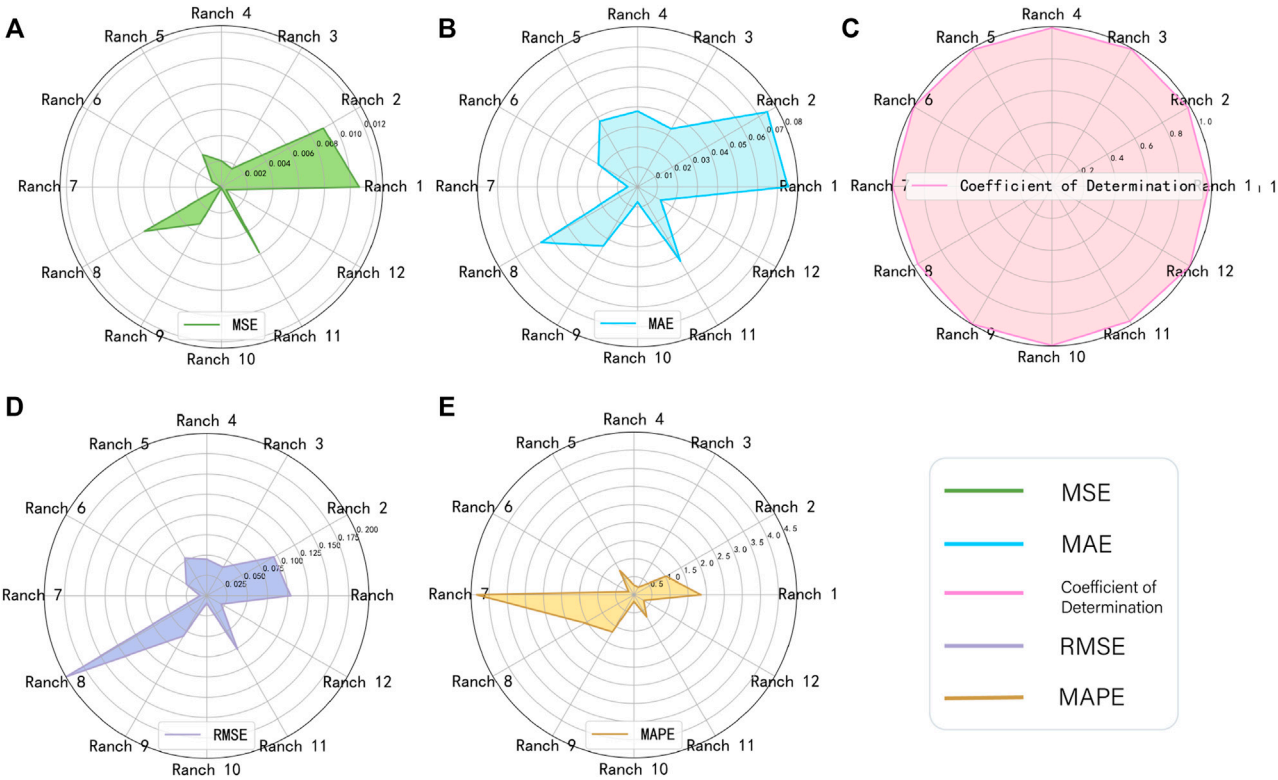


FIGURE 11 | Radar diagrams of the prediction error evaluation indexes for 12 ranches based on the model proposed in this paper. **(A)** is the value of MSE. **(B)** is the value of MAE. **(C)** is the value of RMSE. **(D)** is the value of MAPE. **(E)** is the value of R^2.

The above experimental analysis indicates that 2DD-CNN not only outperforms the other models in predicting DO but also performs better in predicting peaks and valleys, displaying the best overall fit to observations.

4.3 Prediction Accuracy Evaluation Based on Data From Multiple Ranches

DO data from 12 marine ranches are used as samples to verify and evaluate whether 2DD-CNN could be applied to analyze DO data from different marine ranches with large environmental differences. In this paper, the DO level is predicted for 12 h in 12 ranches in the research area. The lowest value reached approximately 0.2 mg/L, and the highest value reached approximately 15 mg/L. All the data selected in this paper are sufficiently representative.

The predicted results are shown in **Figure 10**, and comparisons of predicted and observed values for each ranch are shown through line charts and density correlation plots. The 2DD-CNN displays good performance for all the considered

ranches, and the blue predictions and red observations exhibit high overlap. The positive correlation reaches a maximum when the value is 1 in the density correlation diagram. Generally, when r exceeds 0.5, a strong correlation exists. The correlation between the predicted and observed values is more than 0.7 for all ranches with the 2DD-CNN model.

The correlation between the predicted and observed values for ranches 2, 3, 4, 5, 6, 7, 10 and 12 reached greater than 0.99. The fitting lines of the scatter plots in the density correlation map are very close to y = x, with a small inclination and small intercept. The line chart and density correlation diagram for each ranch show that 2DD-CNN exhibits good performance in predicting peak and valley values, and the agreement between the fitting line and points in the density correlation diagram is high. The model best predicts the DO values for ranches 3, 4, 6, 7, 10 and 12, but the results are not as good for ranches 1, 2, 8, 9 and 11. Notably, the DO data for ranches 3, 4, 6, 7, 10, and 12 are relatively smooth, but in other cases, the data contain a small amount of noise. Although the predicted DO in cases with noise is not as good as that in other cases, the prediction results still display high accuracy; thus, even if DO data contain a small amount of noise, the model can still achieve accurate predictions.

The MSE, MAE, MAPE, RMSE and R^2 are used to measure the accuracy of the 2DD-CNN in DO prediction for multiple marine ranches, as shown in **Table 2** and **Figure 11**. The MSE, MAE, MAPE, RMSE for all ranches are very low, and R^2 is greater than 0.97. **Figure 11** illustrates a three-part radar diagram. **Figure 11A** shows the MSE of the predicted and observed values for 12 ranches. The MSEs vary for different ranches but are all below 0.02. **Figure 11B** shows the MAE of the predicted and observed values for 12 ranches. In **Figure 11B**, the MAEs of multiple ranches displayed in the radar chart are similar to those in **Figure 11A**, with values below 0.08. **Figure 11D** shows the RMSE of the predicted and observed values for 12 ranches with values below 0.2. **Figure 11E** shows the MAPE of the predicted and observed values for 12 ranches with values below 2. **Figure 11C** shows the R^2 results based on the predicted and observed values for the 12 ranches. In **Figure 11C**, an almost circular shape is observed because all values are close to 1. In summary, by analyzing and evaluating the predicted DO values for 12 marine ranches, we find that the 2DD-CNN can effectively forecast DO data in different intervals and in cases with different influencing factors. Thus, the model displays strong generalization ability.

4.4 Discussion

In this part, we will discuss the research results of 2DD-cnn from the aspects of missing value filling, time series transformation into image work and convolution neural network prediction model, so as to further discuss the effectiveness and progressiveness of the proposed dissolved oxygen prediction method. Each section is discussed below.

4.4.1 The Superiority of the Newly Developed SSMF Algorithm

Common data filling methods generally include providing KNN(Qi et al., 2021) data, interpolation, means, medians, etc., and returning the predicted values for the model. The former algorithm is simple, and the characteristics represented by the filled-in data are too singular. Although the predicted value of the latter filling model can accurately match the changes in the time series, the modeling process of the algorithm is too complex. In this paper, we propose a new SSMF algorithm to provide the missing values. This method uses the sequence before the missing value to match the historical data, defines the historical sequence feature as the score formula, considers the historical feature sequence as the decision reference object of the missing value, and takes the final entries of the most similar historical data as the missing value, thereby rendering the provided data more reliable. Usually, we observe the sequence according to whether there are similar fragments in past periods of time. The correlation between the past time series and the current time series and the subsequences related to the past will be used for decision-making regarding the current time point. Similar to this method, the logic is also the method of providing the predicted value of the model, such as using a machine learning model like RF (Deng et al., 2019) to pretest and provide the value after learning the historical data, but the method proposed in this paper is simpler and does not require modeling. Our algorithm can capture the sequence features, and the amount of computation is only O(n). The accuracy of the provided data is guaranteed, while the calculation is simple.

4.4.2 The Superiority of the Two-Dimensional Graph of the DO Data Strategy

In order to obtain more accurate prediction, we should use as many features embedded in time series as possible in the prediction model. Recent studies have shown that by converting one-dimensional time series data into two-dimensional images in some way, more features embedded in the original time series can be retained. Therefore, a new framework is explored to visualize time series, so as to learn the features and structures of time series with the help of the success of deep learning in the field of computer vision. However, at present, most of the research on time series focuses on the classification of time series. For example, A time series classification method based on CNN and recursive graph is proposed (Hatami et al., 2018). In this method, firstly, the recursive graph is used to convert the time series into two-dimensional texture images. Yang et al. (2019) proposed a framework for sensor classification using multivariate time series sensor data as input, which encodes multivariate time series data into two-dimensional color images. However, we propose a novel framework for encoding time series as two-dimensional images to predict DO, this method preserves the time series information in the form of matrix arrangement. At the same time, in addition to the adjacent data, the periodic data or interval data also contain some rules and characteristics. We think that the closer the data is, the more meaningful it is. Therefore, this cyclotron arrangement method transforms the time information into spatial information to a certain extent. The method of transforming sequences into two-dimensional pictures

proposed in this paper aims to mine various forms of information, and is also convenient for two-dimensional convolution kernel to further extract a variety of information. We use the method of converting dissolved oxygen data into two-dimensional pictures and inputting them into convolutional neural network, which rarely appears in the study of dissolved oxygen prediction. Moreover, according to the experiment in **Section 4.2**, the prediction accuracy of our proposed method is higher than that of other models.

4.4.3 The Effectiveness of Self-Attention in the CNN Module

The long-term dependence of the capture sequence plays an important role in deep learning prediction models. However, the convolution operation has a significant weakness in that it only operates on a local neighbourhood, thus missing global information (Bello et al., 2019). With the deepening of the network, there has always been the problem of local calculation, limiting the performance of the model. To resolve this problem, this paper uses a self-attention mechanism to improve the CNN. The self-attention mechanism is a variant of the attention mechanism, decreasing the dependence on external information. A self-attention mechanism is used to mine the influence weight of information at each position in the input matrix on the prediction results, which can accurately capture the internal correlation of data or features and better assist the optimization process of CNN models (Jia et al., 2021). Its application in DO sequence prediction is mainly through calculating the interaction between DO sequences, to solve the problem of long-distance dependence. The self-attention mechanism is a variant of the attention mechanism, decreasing the dependence on external information, and it is better at capturing the internal correlation of data or features. At this time, the CNN prediction model is more focused on capturing the characteristics of the input matrix. Through the learning of feedforward neural networks, we can better consider the context information of time series.

5 CONCLUSION

Because the aquatic environment of marine pastures is affected by various factors, the change in DO is complex and nonlinear. To improve the prediction accuracy for DO, the change trend in it can be accurately predicted. In this paper, an improved 2DD-CNN DO prediction model is proposed. In the pretreatment stage, an SSMF method is proposed to provide missing values, and a new method is used to convert the time series of water quality parameters into pictures and input them into a two-dimensional CNN. At the same time, the two-dimensional CNN model is improved, and a convolutional self-attention module is added to the network to resolve the long-distance dependence problem by calculating the interaction between DO sequences. The model proposed in this paper achieves good improvement in prediction accuracy. The 2DD-CNN model has a very good prediction effect and exhibits good generalizability for the prediction error, fitting degree, peak valley value and data segments with large and gentle fluctuations. This model is applicable not only to the prediction of one water quality parameter but also to the prediction of other water qualities. The prediction of water quality parameters plays an important role in marine ranch management by providing quantitative information for the solution of emerging environmental problems and the decision-making of sustainable management.

Although 2DD-CNN has achieved good results in predicting DO, there remain many aspects that can be improved. First, DO data preprocessing has a significant impact on the accuracy of data modeling and is an important method to improve the accuracy of DO prediction. In the method that the SSMF used to provide missing values in this paper, parameter optimization is greatly affected by human subjective factors and cannot ensure the optimization of set parameters. Therefore, optimizing SSMF parameters will be the focus of the next improvement. Second, the research in this paper only involves the prediction of one-dimensional DO, but due to the interaction of water quality parameters of marine pastures, DO is affected by many water quality parameters. To further capture the variation characteristics of DO, predicting DO according to multidimensional water quality parameters is an important research direction. In addition, the method of transforming time into images designed in this paper could store more feature data, so further research work could be performed in the direction of feature expression in the future to better mine the internal relationships of data and to improve the prediction accuracy.

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 authors.

AUTHOR CONTRIBUTIONS

DL designed the study and revised the manuscript. XZ performed the study and wrote the manuscript. DL and XZ revised the manuscript. All authors approved the publication of the final manuscript.

FUNDING

This work is supported in part by the National Natural Science Foundation of China under Grants 61772319, 61472227, and 61602277; the Yantai Science and Technology Innovation Development Project (2022YTJC06003022).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenvs.2022.904939/full#supplementary-material

REFERENCES

Adnan, R. M., Liang, Z., El-Shafie, A., Zounemat-Kermani, M., and Kisi, O. (2019). Prediction of Suspended Sediment Load Using Data-Driven Models. *Water* 11, 2060. doi:10.3390/w11102060

Adnan, R. M., Liang, Z., Parmar, K. S., Soni, K., and Kisi, O. (2020). Modeling Monthly Streamflow in Mountainous Basin by Mars, Gmdh-Nn and Denfis Using Hydroclimatic Data. *Neural Comput. Applic.* 33, 2853–2871. doi:10.1007/s00521-020-05164-3

Ahmed, M. H., and Lin, L.-S. (2021). Dissolved Oxygen Concentration Predictions for Running Waters with Different Land Use Land Cover Using a Quantile Regression Forest Machine Learning Technique. *J. Hydrol.* 597, 126213. doi:10.1016/j.jhydrol.2021.126213

Anmala, J., and Turuganti, V. (2021). Comparison of the Performance of Decision Tree (DT) Algorithms and Extreme Learning Machine (ELM) Model in the Prediction of Water Quality of the Upper Green River Watershed. *Water Environ. Res.* 93, 2360–2373. doi:10.1002/wer.1642

Ashourloo, D., Shahrabi, H. S., Azadbakht, M., Rad, A. M., Aghighi, H., and Radiom, S. (2020). A Novel Method for Automatic Potato Mapping Using Time Series of Sentinel-2 Images. *Comput. Electron Agric.* 175, 105583. doi:10.1016/j.compag.2020.105583

Baek, S.-S., Pyo, J., and Chun, J. A. (2020). Prediction of Water Level and Water Quality Using a Cnn-Lstm Combined Deep Learning Approach. *Water* 12, 3399. doi:10.3390/w12123399

Barzegar, R., Aalami, M. T., and Adamowski, J. (2020). Short-term Water Quality Variable Prediction Using a Hybrid CNN-LSTM Deep Learning Model. *Stoch. Environ. Res. Risk Assess.* 34, 415–433. doi:10.1007/s00477-020-01776-2

Barzegar, R., Aalami, M. T., and Adamowski, J. (2021). Coupling a Hybrid Cnn-Lstm Deep Learning Model with a Boundary Corrected Maximal Overlap Discrete Wavelet Transform for Multiscale Lake Water Level Forecasting. *J. Hydrol.* 598, 126196. doi:10.1016/j.jhydrol.2021.126196

Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. (2019). "Attention Augmented Convolutional Networks," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 27 – November 2, 2019, Seoul, 3286–3295. doi:10.1109/iccv.2019.00338

Cao, X., Ren, N., Tian, G., Fan, Y., and Duan, Q. (2021). A Three-Dimensional Prediction Method of Dissolved Oxygen in Pond Culture Based on Attention-Gru-Gbrt. *Comput. Electron. Agric.* 181, 105955. doi:10.1016/j.compag.2020.105955

Chen, H., Chen, A., Xu, L., Xie, H., Qiao, H., Lin, Q., et al. (2020). A Deep Learning Cnn Architecture Applied in Smart Near-Infrared Analysis of Water Pollution for Agricultural Irrigation Resources. *Agric. Water Manag.* 240, 106303. doi:10.1016/j.agwat.2020.106303

Deng, W., Guo, Y., Liu, J., Li, Y., Liu, D., and Zhu, L. (2019). A Missing Power Data Filling Method Based on Improved Random Forest Algorithm. *Chin. J. Electr. Eng.* 5, 33–39. doi:10.23919/CJEE.2019.000025

Haque, M. A., Verma, A., Alex, J. S. R., and Venkatesan, N. (2020). "Experimental Evaluation of Cnn Architecture for Speech Recognition," in First International Conference on Sustainable Technologies for Computational Intelligence, August 22–26, 2020, Shenzhen (Springer), 507–514. doi:10.1007/978-981-15-0029-9_40

Hatami, N., Gavet, Y., and Debayle, J. (2018). "Classification of Time-Series Images Using Deep Convolutional Neural Networks," in Tenth International Conference on Machine Vision (ICMV 2017), November 13–15, 2017, Vienna, Austria (International Society for Optics and Photonics), 106960Y. doi:10.1117/12.2309486

Heddam, S., and Kisi, O. (2017). Extreme Learning Machines: a New Approach for Modeling Dissolved Oxygen (Do) Concentration with and without Water Quality Variables as Predictors. *Environ. Sci. Pollut. Res.* 24, 16702–16724. doi:10.1007/s11356-017-9283-z

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18–22, 2018, Salt Lake City, USA, 7132–7141. doi:10.1109/cvpr.2018.00745

Jia, M., Gao, Y., Li, S., Yue, J., and Ye, M. (2021). An Explicit Self-Attention-Based Multimodality Cnn In-Loop Filter for Versatile Video Coding. *Multimed. Tools Appl.*, 1–15. doi:10.1007/s11042-021-11214-2

Jiang, H., Zhang, C., Qiao, Y., Zhang, Z., Zhang, W., and Song, C. (2020). Cnn Feature Based Graph Convolutional Network for Weed and Crop Recognition in Smart Farming. *Comput. Electron. Agric.* 174, 105450. doi:10.1016/j.compag.2020.105450

Karijadi, I., and Chou, S.-Y. (2022). A Hybrid Rf-Lstm Based on Ceemdan for Improving the Accuracy of Building Energy Consumption Prediction. *Energy Build.* 259, 111908. doi:10.1016/j.enbuild.2022.111908

Khosravi, K., Panahi, M., Golkarian, A., Keesstra, S. D., Saco, P. M., Bui, D. T., et al. (2020). Convolutional Neural Network Approach for Spatial Prediction of Flood Hazard at National Scale of iran. *J. Hydrol.* 591, 125552. doi:10.1016/j.jhydrol.2020.125552

Kim, P. (2017). "Convolutional Neural Network," in *MATLAB Deep Learning* (California and La Jolla, CA: Springer), 121–147. doi:10.1007/978-1-4842-2845-6_6

Liu, S., Xu, L., Jiang, Y., Li, D., Chen, Y., and Li, Z. (2014). A Hybrid WA-CPSO-LSSVR Model for Dissolved Oxygen Content Prediction in Crab Culture. *Eng. Appl. Artif. Intell.* 29, 114–124. doi:10.1016/j.engappai.2013.09.019

Luo, H., Xiong, C., Fang, W., Love, P. E. D., Zhang, B., and Ouyang, X. (2018). Convolutional Neural Networks: Computer Vision-Based Workforce Activity Assessment in Construction. *Automation Constr.* 94, 282–289. doi:10.1016/j.autcon.2018.06.007

Ni, W., Li, M., Ross, A. C., and Najjar, R. G. (2019). Large Projected Decline in Dissolved Oxygen in a Eutrophic Estuary Due to Climate Change. *J. Geophys. Res. Oceans* 124, 8271–8289. doi:10.1029/2019JC015274

Niu, X., and Wang, J. (2019). A Combined Model Based on Data Preprocessing Strategy and Multi-Objective Optimization Algorithm for Short-Term Wind Speed Forecasting. *Appl. Energy* 241, 519–539. doi:10.1016/j.apenergy.2019.03.097

Pei, Y., Huang, Y., Zou, Q., Zhang, X., and Wang, S. (2021). Effects of Image Degradation and Degradation Removal to Cnn-Based Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 1239–1253. doi:10.1109/tpami.2019.2950923

Qi, X., Guo, H., and Wang, W. (2021). A Reliable Knn Filling Approach for Incomplete Interval-Valued Data. *Eng. Appl. Artif. Intell.* 100, 104175. doi:10.1016/j.engappai.2021.104175

Rozario, A. P. R., and Devarajan, N. (2021). Monitoring the Quality of Water in Shrimp Ponds and Forecasting of Dissolved Oxygen Using Fuzzy C Means Clustering Based Radial Basis Function Neural Networks. *J. Ambient. Intell. Hum. Comput.* 12, 4855–4862. doi:10.1007/s12652-020-01900-8

Song, Z. (2020). English Speech Recognition Based on Deep Learning with Multiple Features. *Computing* 102, 663–682. doi:10.1007/s00607-019-00753-0

Sun, Y., Xue, B., Zhang, M., Yen, G. G., and Lv, J. (2020). Automatically Designing Cnn Architectures Using the Genetic Algorithm for Image Classification. *IEEE Trans. Cybern.* 50, 3840–3854. doi:10.1109/TCYB.2020.2983860

Takahashi, S., Hori, R. S., Yamakita, S., Aita, Y., Takemura, A., Ikehara, M., et al. (2021). Progressive Development of Ocean Anoxia in the End-Permian Pelagic Panthalassa. *Glob. Planet. Change* 207, 103650. doi:10.1016/j.gloplacha.2021.103650

Tiyasha, T., Tung, T. M., Bhagat, S. K., Tan, M. L., Jawad, A. H., Mohtar, W. H. M. W., et al. (2021). Functionalization of Remote Sensing and On-Site Data for Simulating Surface Water Dissolved Oxygen: Development of Hybrid Tree-Based Artificial Intelligence Models. *Mar. Pollut. Bull.* 170, 112639. doi:10.1016/j.marpolbul.2021.112639

Valera, M., Walter, R. K., Bailey, B. A., and Castillo, J. E. (2020). Machine Learning Based Predictions of Dissolved Oxygen in a Small Coastal Embayment. *J. Mar. Sci. Eng.* 8, 1007. doi:10.3390/jmse8121007

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention Is All You Need," in Advances in Neural Information Processing Systems, California and La Jolla, CA: Neural Information Processing Systems, 5998–6008.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-local Neural Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18–22, 2018, Salt Lake City, USA, 7794–7803. doi:10.1109/cvpr.2018.00813

Wang, Y., Zhang, Z., Feng, L., Ma, Y., and Du, Q. (2021). A New Attention-Based Cnn Approach for Crop Mapping Using Time Series Sentinel-2 Images. *Comput. Electron. Agric.* 184, 106090. doi:10.1016/j.compag.2021.106090

Wu, J., Li, Z., Zhu, L., Li, G., Niu, B., and Peng, F. (2018). Optimized Bp Neural Network for Dissolved Oxygen Prediction. *IFAC-PapersOnLine* 51, 596–601. doi:10.1016/j.ifacol.2018.08.132

Xiao, X., Zhang, D., Hu, G., Jiang, Y., and Xia, S. (2020). CNN-MHSA: A Convolutional Neural Network and Multi-Head Self-Attention Combined Approach for Detecting Phishing Websites. *Neural Netw.* 125, 303–312. doi:10.1016/j.neunet.2020.02.013

Yan, J., Liu, J., Yu, Y., and Xu, H. (2021). Water Quality Prediction in the Luan River Based on 1-drcnn and Bigru Hybrid Neural Network Model. *Water* 13, 1273. doi:10.3390/w13091273

Yang, C.-L., Chen, Z.-X., and Yang, C.-Y. (2019). Sensor Classification Using Convolutional Neural Network by Encoding Multivariate Time Series as Two-Dimensional Colored Images. *Sensors* 20, 168. doi:10.3390/s20010168

Yang, X., Zheng, C., Zou, C., Gan, H., Li, S., Huang, S., et al. (2021). A Cnn-Based Posture Change Detection for Lactating Sow in Untrimmed Depth Videos. *Comput. Electron. Agric.* 185, 106139. doi:10.1016/j.compag.2021.106139

Yaqub, M., Asif, H., Kim, S., and Lee, W. (2020). Modeling of a Full-Scale Sewage Treatment Plant to Predict the Nutrient Removal Efficiency Using a Long Short-Term Memory (Lstm) Neural Network. *J. Water Process Eng.* 37, 101388. doi:10.1016/j.jwpe.2020.101388

Yu, S., Liu, D., Zhu, W., Zhang, Y., and Zhao, S. (2020). Attention-based Lstm, Gru and Cnn for Short Text Classification. *J. Intell. Fuzzy Syst.* 39, 333–340. doi:10.3233/JIFS-191171

Yuan, X., Chen, C., Lei, X., Yuan, Y., and Muhammad Adnan, R. (2018). Monthly Runoff Forecasting Based on Lstm-Alo Model. *Stoch. Environ. Res. Risk Assess.* 32, 2199–2212. doi:10.1007/s00477-018-1560-y

Zeng, F., Nait Amar, M., Mohammed, A. S., Motahari, M. R., and Hasanipanah, M. (2021). Improving the Performance of LSSVM Model in Predicting the Safety Factor for Circular Failure Slope through Optimization Algorithms. *Eng. Comput*, 1–12. doi:10.1007/s00366-021-01374-y

Zhang, D., and Lou, S. (2021). The Application Research of Neural Network and BP Algorithm in Stock Price Pattern Classification and Prediction. *Future Gener. Comput. Syst.* 115, 872–879. doi:10.1016/j.future.2020.10.009

Zhang, Y., Fitch, P., Vilas, M. P., and Thorburn, P. J. (2019). Applying Multi-Layer Artificial Neural Network and Mutual Information to the Prediction of Trends in Dissolved Oxygen. *Front. Environ. Sci.* 7, 46. doi:10.3389/fenvs.2019.00046

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., et al. (2021). From Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale? *Environ. Sci. Technol.* 55, 2357–2368. doi:10.1021/acs.est.0c06783

Zhu, N., Ji, X., Tan, J., Jiang, Y., and Guo, Y. (2021). Prediction of Dissolved Oxygen Concentration in Aquatic Systems Based on Transfer Learning. *Comput. Electron. Agric.* 180, 105888. doi:10.1016/j.compag.2020.105888

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.

Copyright © 2022 Li and Zhang. 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 Marine Science

ORIGINAL RESEARCH
published: 14 July 2022
doi: 10.3389/fmars.2022.844359

CGAN Based Improved ASI Retrieval Algorithm for Antarctic Sea Ice Concentration

Xingdong Wang[1], Yue Zhao[1], Shuhui Yang[1], Yuhua Wang[1] and Donghui Shangguan[2]*

[1]College of Information Science and Engineering, Henan University of Technology, Zhengzhou, China, [2]State Key Laboratory of Cryospheric Science, Northwest Institute of Eco-Environment and Resources, Chinese Academy Science, Lanzhou, China

Sea ice change is closely related to the change of global atmosphere and ocean circulation, which plays an important role in the study of global climate change. Sea ice concentration is one of the important parameters to study the temporal and spatial change of sea ice. Accurately retrieving sea ice concentration is the innovation of this paper. At present, the high-resolution microwave-detected sea ice concentration product was provided by the University of Bremen, which was derived by the Arctic Radiation and Turbulence Interaction Study (ARTSIST) Sea Ice (ASI) algorithm based on the Advanced Microwave Scanning Radiometer for Earth Observing System (AMSR-E) 89-GHz brightness temperature data. The AMSR-E/AMSR-2 89-GHz brightness temperature data has higher spatial resolution, but it is often affected by cloud and water vapor, which affects the recognition and subsequent use of ground feature. Although the weather filters can remove some errors in the edge regions of the sea water and the sea ice, the errors of the sea ice concentration in other regions cannot be removed. The generative model of Conditional Generative Adversarial Network (CGAN) increases the utilization of image feature information through skip connection, which improves the removal of the influence of cloud and water vapor. The discriminative model can retain the image feature information and realize the non-linear mapping from the image to the image. The loss function can reduce the pixel-level loss, which can remove the influence of cloud and water vapor. Therefore, this paper proposed an improved ASI algorithm based on CGAN. Firstly, the relatively stable relationship between the 89-GHz brightness temperature data which is not disturbed or less affected by the external environment and the 36-GHz brightness temperature data was determined, and the 89-GHz brightness temperature data with large interference was screened. Secondly, based on the 36-GHz brightness temperature data with high reliability, the 89-GHz brightness temperature data with large interference was corrected through CGAN. Finally, the ASI algorithm was used to retrieve sea ice concentration. Compared with sea ice concentration retrieved by the ASI algorithm, the results showed that the improved ASI algorithm based on CGAN was feasible. Compared with sea ice distribution obtained from the Landsat 8 OLI-L1T data, the improved ASI algorithm based on CGAN significantly improves the inversion accuracy of sea ice concentration. The improved ASI algorithm based on CGAN makes use of the reliable 36-GHz brightness temperature data, which greatly reduces the error caused by cloud and water vapor, and the method effectively corrects sea ice concentration of the pixels affected by the external

OPEN ACCESS

Edited by:
Yiliao Song,
University of Technology Sydney, Australia

Reviewed by:
Dazhou Zhang,
Central South University, China
Xiangbin Cui,
Polar Research Institute of China, China

***Correspondence:**
Shuhui Yang
ysh1563852@163.com

Specialty section:
This article was submitted to Global Change and the Future Ocean, a section of the journal Frontiers in Marine Science

Received: *28 December 2021*
Accepted: *13 June 2022*
Published: *14 July 2022*

Citation:
Wang X, Zhao Y, Yang S, Wang Y and Shangguan D (2022) CGAN Based Improved ASI Retrieval Algorithm for Antarctic Sea Ice Concentration. Front. Mar. Sci. 9:844359. doi: 10.3389/fmars.2022.844359

environment. Therefore, the improved ASI algorithm based on CGAN realizes high spatial resolution and significantly improves the inversion accuracy of sea ice concentration.

Keywords: Antarctic, sea ice concentration, CGAN, data correction, AMSR-2

1 INTRODUCTION

The polar region is an important indicator of global climate. With the increasing severity of global warming, sea ice is an important climate factor in the polar region, and the monitoring and the studies of sea ice have attracted more and more attention. Sea ice concentration is the most intuitive parameter to study the sea ice change. Sea ice concentration plays an important role in the monitoring and prediction of sea ice change, and it is of great significance to study global climate change. Passive microwave data is not limited by day and night, which is less affected by clouds and fog, and it has good temporal and spatial continuity. Passive microwave data have been used extensively for polar sea ice monitoring. Sea ice concentration can provide reliable basic data and scientific basis to study the polar region and global climate.

Sea ice concentration plays an important role in climate-change study and ship navigation in the polar regions. Many algorithms for retrieving sea ice concentration had been proposed in recent decades. Bootstrap algorithm mainly used the characteristics of the polarization difference between sea water and sea ice, which also uesd the high-frequency data and the low-frequency data of passive microwave radiometer to retrieve sea ice concentration (Comiso, 1986; Comiso, 1995). Based on Special Sensor Microwave/Imager (SSM/I) brightness temperature data, Cavalieri et al. proposed NASA Team (NT) algorithm, which can retrieve first-year ice concentration and multi-year ice concentration (Cavalieri et al., 1991). Cavalieri et al. proposed a method for determining sea ice parameters using dual-polarized multispectral brightness temperature data gathered by the Nimbus 7 Scanning Multichannel Microwave Radiometer (SMMR) (Cavalieri et al., 1984). Liu et al. proposed a fully constrained least squares algorithm based on NT algorithm to retrieve Antarctic sea ice concentration (Liu et al., 2015). On the basis of NT algorithm, Markus et al. added brightness temperature data at 89-GHz vertical polarization and proposed NT 2 algorithm (Markus and Cavalieri, 2000). Based on the SSM/I 85.5-GHz brightness temperature data, Lomax et al. proposed Lomax algorithm to retrieve sea ice concentration (Lomax et al., 1995). Hao improved the NT algorithm by introducing AMSR-E 6.9-GHz brightness temperature data and improved the accuracy of multi-year ice concentration (Hao and Su, 2015). Kern et al. proposed the SEA LION (SL) algorithm to retrieve sea ice concentration based on 37-GHz polarization difference (Kern, 2001; Kern and Heygster, 2001). The ASI algorithm was derived from the project "Arctic Radiation and Turbulence Interaction Study (ARTIST)" in 1998. Based on the concept of "polarization correction temperature", Svendsen et al. proposed a model for retrieving total sea ice concentration from a spaceborne dual-polarized passive microwave instrument near 90 GHz (Svendsen et al., 1987; Spencer et al., 1989). Kaleschke et al. improved the algorithm proposed by Svendsen et al., and used SSM/I 85-GHz brightness temperature data to conduct mesoscale numerical simulation of the atmospheric boundary layer at the edge of Arctic sea ice (Svendsen et al., 1987; Kaleschke et al., 2001). One advantage of the ASI algorithm is that, compared with other algorithms using 85-GHz brightness temperature data, it does not require additional input data (Kern, 2004). The ASI algorithm can directly retrieve sea ice concentration based on the 89-GHz brightness temperature data, and it has a similar result with sea ice concentration algorithms using other data channels (Kern et al., 2003). Spreen et al. applied the ASI algorithm to the AMSR-E 89-GHz brightness temperature data and obtained the inversion formula of sea ice concentration (Spreen et al., 2008). Wang proposed a multi-year ice concentration algorithm based on the different characteristics of the first-year ice, multi-year ice and sea water at 89-GHz brightness temperature data (Wang, 2009). Based on the 89-GHz brightness temperature data and the ASI algorithm, Su et al. carried out a series of experiments on interpolation algorithm and weather filter (Su et al., 2013). Zhang et al. proposed an algorithm to retrieve sea ice concentration using multichannel and dual-polarized data according to the radiation characteristics of sea ice and sea water (Zhang, 2012). Wu et al. proposed an enhanced ASI algorithm which used the 19-GHz polarization difference to modify the 91-GHz polarization difference (Wu et al., 2019).

The spatial resolution and inversion algorithm of satellite data are very important to accurately provide sea ice concentration. Although the ASI algorithm has advantages, compared with the low-frequency brightness temperature data, the 89-GHz brightness temperature data are more affected by cloud and water vapor. When the liquid water content in the cloud is high or there is a cyclone passing by, it will lead to a large error of sea ice concentration in the edge regions of sea water and sea ice. Therefore, the ASI algorithm needs weather filter processing (Spreen et al., 2008). Although some errors can be eliminated by weather filter, the errors of the sea ice concentration in some regions cannot be removed. The generative model of CGAN increases the utilization of image feature information through skip connection, which improves the removal of the influence of cloud and water vapor. The discriminative model can retain the image feature information and realize the non-linear mapping from the image to the image. The loss function can reduce the pixel-level loss, which can remove the influence of cloud and water vapor. Therefore, CGAN was used to realize image correction in this paper. Firstly, in order to obtain more accurate sea ice concentration, the 89-GHz brightness temperature data greatly affected by the external environment such as cloud and

water vapor is screened based on the relatively stable relationship between the 89-GHz brightness temperature data not disturbed or less affected by the external environment such as cloud and water vapor and the 36-GHz brightness temperature data. Secondly, this paper used the data correction method based on CGAN to correct the 89-GHz brightness temperature data greatly affected by the external environment such as cloud and water vapor. Finally, based on the correction data obtained in the second step, this paper used the ASI algorithm to retrieve Antarctic sea ice concentration. This method effectively corrected sea ice concentration of the pixels affected by the external environment and greatly reduced the error caused by cloud and water vapor. The Landsat 8 OLI-L1T data were used to verify sea ice concentration retrieved by the improved ASI algorithm based on CGAN.

2 DATASETS

The Advanced Microwave Scanning Radiometer for EOS (AMSR-E), carried on the NASA satellite Aqua, is a 12-channel, 6-frequency microwave radiometer that measures brightness temperatures at 6.925 GHz, 10.65 GHz, 18.7 GHz, 23.8 GHz, 36.5 GHz, and 89.0 GHz with vertical polarization and horizontal polarization. Spatial resolution of the individual measurements varies from 5.4 km at 89.0 GHz to 74 × 43 km at 6.9 GHz, and it is the lower-frequency channels that provided the SST measurement capability. The Advanced Microwave Scanning Radiometer-2 (AMSR2) is a multi-frequency total-power microwave radiometer with dual-polarization channels onboard the Global Change Observation Mission (GCOM) 1st-Water (GCOMW1) (Imaoka et al., 2010). The basic characteristics are almost identical to those of a predecessor sensor, AMSR-E. AMSR2 continues AMSR-E observations with several improvements. The AMSR-E/AMSR-2 data can provide a variety of parameters of land, ocean and atmosphere. Such as precipitation rate, sea surface temperature, sea ice concentration, soil humidity, wind speed and water vapor in the atmosphere. The AMSR-E/AMSR-2 89-GHz brightness temperature data and 36-GHz brightness temperature data are used in this paper. (https://seaice.uni-bremen.de/).

The Landsat 8 is the eighth satellite in the Landsat series. It was originally called Landsat Data Continuity Mission (LDCM). The Landsat 8 carries the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). The OLI includes 9 bands with a spatial resolution of 30 meters, including a 15-meter panchromatic band (Knight and Kvaran, 2014). In this paper, the Landsat 8 OLI-L1T data released by United States Geological Survey(USGS) was selected as the verification data (https://earthexplorer.usgs.gov/). In this paper, the 89-GHz brightness temperature data and the 36-GHz brightness temperature data from October 2019 to March 2020 were selected as the input of CGAN. The high-resolution optical data of Landsat 8 satellite were selected to verify sea ice concentration obtained in this paper.

3 METHODOLOGY

3.1 ASI Algorithm

The ASI algorithm used the polarization difference between the 89-GHz vertical brightness temperature and the 89-GHz horizontal brightness temperature to retrieve sea ice concentration and used the low-frequency brightness temperature data as the weather filters to remove the errors of sea ice concentration in the regions of the low sea ice concentration and the sea water (Svendsen et al., 1987; Spreen et al., 2008). The AMSR-E/AMSR-2 89-GHz brightness temperature data is significantly affected by cloud and water vapor in the atmosphere. In particular, the cyclones in the sea water regions will weaken the polarization difference of sea water, make this part of sea water close to the polarization difference of sea ice, and it may lead to this part of sea water being mistaken for sea ice. Therefore, it is very necessary to use the weather filter to remove the errors of sea ice concentration due to the external environment such as cloud and water vapor.

Up to now, all weather filters basically use low-frequency data. In 1986, Comiso used the Gradient Ratio (GR) at 36.5 GHz and 18.7 GHz to reduce the influence of cloud and water vapor. Because GR (37/19) of the sea water is greater than 0, while GR (37/19) of the sea ice is close to 0 or less than 0 (Comiso, 1986). In 1995, Comiso improved the weather filter by adding GR (23/19) in addition to the original GR (37/19), because the GR at 23 GHz and 19 GHz is more sensitive to the water vapor of the atmosphere (Comiso, 1995).

3.2 An Improved ASI Algorithm

At present, the weather filter used in the ASI algorithm only removes the misjudged sea ice in the open ocean, and it does not change sea ice concentration affected by cloud and water vapor. Therefore, in order to obtain more accurate sea ice concentration with the high spatial resolution, the process is as follows. Firstly, we screened the 89-GHz brightness temperature data greatly affected by the external environment, such as cloud and water vapor. Then we proposed the data correction method based on CGAN to correct the 89-GHz brightness temperature data greatly affected by the external environment such as cloud and water vapor, so as to replace the weather filter used in the ASI algorithm.

3.2.1 Data Screening

The external environment such as cloud and water vapor basically has no impact on the 36-GHz brightness temperature data, but has a great impact on the 89-GHz brightness temperature data. Under sunny weather, the Polarization Ratio(PR) of the 89-GHz brightness temperature data and the 36-GHz brightness temperature data is stable (Iwamoto et al., 2013). But when there is external interference such as cloud and water vapor, the PR of the 89-GHz brightness temperature data and the 36-GHz brightness temperature data will be reduced, and the degree of reduction is related to the impact of the external environment

such as cloud and water vapor. The PR value is obtained as shown in equation (1). Therefore, under sunny weather, take the PR of the 36-GHz brightness temperature data as the abscissa and the PR of the 89-GHz brightness temperature data as the ordinate, and draw the PR scatter plot, as shown in **Figure 1**. In the PR scatter plot, the abscissa is equally divided into several intervals, and the average value and standard deviation of the ordinate in each interval are calculated. Then, the best fitting curve based on the least square method is drawn by subtracting the value of twice the standard deviation from the average value in each interval, which is similar to the quadratic equation.

$$PR = \frac{TB_v - TB_H}{TB_v + TB_H} \tag{1}$$

Where TB_v and TB_H are the vertical polarization brightness temperature data and the horizontal polarization brightness temperature data respectively.

$$PR_{89} = a * \left(PR_{36}\right)^2 - b * a_0 PR_{36} + c \tag{2}$$

Where PR_{89} is the PR of the 89-GHz brightness temperature data and PR_{36} is the PR of the 36-GHz brightness temperature data. And a, b, c are constants.

3.2.2 Data Correction

CGAN can better fit complex nonlinear noise data and introduce additional condition information to guide data generation, so that CGAN has better denoising effect. The generative model and the discriminative model of CGAN can get the relationship between the data affected by the external environment and the data not affected by the external environment through confrontation training. If the generative model outputs an image with poor correction results, the network parameters are continuously updated through the feedback mechanism of the discriminative model, to guide the generative model to correct the data affected by the external environment. The core idea of CGAN model is to achieve Nash equilibrium through the game (The game function of CGAN model is shown in equation (3).) between the generative model and the discriminative model. The purpose of the generative model is to generate data that is not affected by the external environment, to improve the generation ability and reduce the discrimination ability of the discriminative model. The discriminative model judges the difference of input data through the loss function, updates the parameters of CGAN model through the feedback mechanism, and finally obtains the optimal CGAN model.

$$\begin{aligned} \min_G \max_D V(G,D) = {} & E_{x\sim pdata(x)}[log(D(x \mid y))] \\ & + E_{Z\sim p_{z(z)}}[log(1-(D(G(z \mid y))))] \end{aligned} \tag{3}$$

Where x is the data affected by the external environment, y is the additional information, z is the input random noise, G (z | y) is the data that not affected by the external environment and that output by the generation network of CGAN, D (G(z | y)) is the probability that the discriminative network judges whether the input data is false. Since the goal of the generative model of CGAN is to make the generated data close to the data that not affected by the external environment as much as possible, the loss function is set to 1 - (D(G(z | y))) to ensure that the probability of judging the output false image of the discriminative network as small as possible. The goal of the discriminative model is to improve the ability to judge the difference of input data. Therefore, the larger D (x | y) is, the better it is. At the same time, the smaller the noise impacts, the better it is. The loss function is set to D (x | y) + 1 – (D(G(z | y)). Use $mi_G^n \, {}^{max}_D V(D,G)$ to represent the process of the game.

The convolutional neural network (CNN) based on the U-Net model can integrate the characteristics of different network layers by the skip connection and improve the denoising performance, and it has strong adaptability and can effectively retain the structural information of the image. Thus, the CNN based on the U-Net model was used as the generative model in this paper, and it includes input layer, convolution layer, pooling layer, activation layer, and output layer. We set the size of the pool layer to 2 × 2, and set the size of the convolution filter to 3 × 3, and selected Rectified Linear Unit (ReLU) as the activation function.

The function of discriminative model is to distinguish two sets of relationships, that is, the relationship between the 89-GHz brightness temperature data with large interference and the 36-GHz brightness temperature data obtained by the generative model, and the relationship between the undisturbed 89-GHz brightness temperature data and the 36-GHz brightness temperature data with the high reliability. The discriminative model adopted the CNN network. Firstly, the image was input into the discriminative model, and then the batch normalization (BN) operation was performed on the input image. Secondly, the feature is extracted through convolution. Thirdly, the ReLU activation function is used for the non-linear mapping, and the final loss is calculated by cross entropy. Finally, the corrected 89-GHz brightness temperature data was obtained. So CGAN was applied to data correction as follows.

(1) Before the training, adjust the data set, such as rotation, translation, to increase the number of data set. After that, the training set and the test set are normalized.
(2) Input the training set into the generative model, and then perform continuous BN + convolution + ReLU + pooling to complete the down sampling operation.
(3) Continuous operations such as deconvolution, ReLU and dropout are performed on the feature map obtained by down sampling to complete up sampling.
(4) The output characteristic diagram of the down sampling is connected with the output characteristic diagram of the up sampling (In the same network layer, each neural network node of the current neural network layer uses the dense jump connections for the feature fusion. In different network layers, from top-layer neural network to bottom-layer neural network, the output feature maps of the down sampling and the up sampling of the next neural network layers are fused.). Then the relationship between the 89-GHz brightness temperature data and the 36-GHz brightness temperature data with the high reliability is obtained.

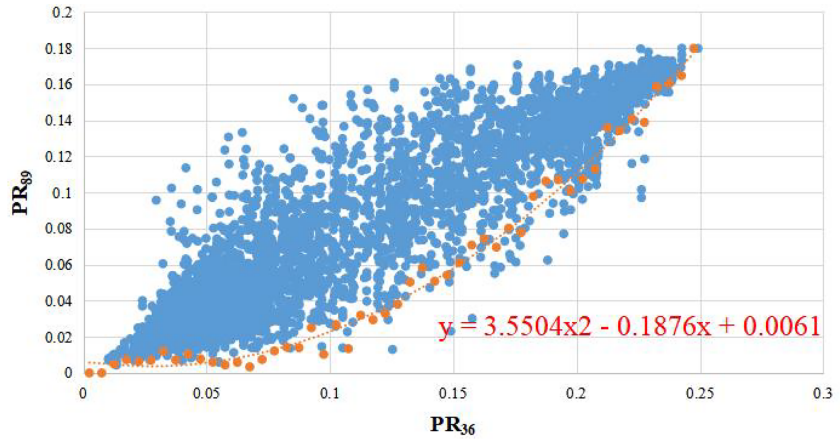


FIGURE 1 | Scatter plot of PR.

(5) The test set data (undisturbed 89-GHz brightness temperature data and 36-GHz brightness temperature data) and the relationship between the 89-GHz brightness temperature data obtained in the previous step and the 36-GHz brightness temperature data with the high reliability are input into the discriminative model.
(6) Carry out BN + convolution + ReLU + pooling to complete the down-sampling operation.
(7) The cross entropy is used to judge the results obtained by the discriminative model. If the loss function reaches the minimum value, the corrected 89-GHz brightness temperature data is output. Otherwise, return to step (2), and repeat the above steps until the loss function reaches the minimum value. The flowchart of the affected 89-GHz brightness temperature data correction based on CGAN model is shown in **Figure 2**.

4 RESULTS AND VERIFICATION

Based on the AMSR-2 89-GHz brightness temperature data, we used the ASI algorithm and the improved ASI algorithm based on CGAN to retrieve Antarctic sea ice concentration on February 1, 2021 as shown in **Figure 3**, and then further verified sea ice concentration by the Landsat 8 OLI-L1T data.

According to the Landsat 8 OLI-L1T data (resolution: 30m), we selected Landsat images for the different regions from October 2019 to March 2020 for the further verification. Based on the Landsat 8 OLI-L1T data, we used NDSI (normalized difference snow index) calculated by the near-infrared band and the short-wave near-infrared band to identify the sea ice distribution. Because this method can identifiy sea ice and sea water according to the reflectivity difference between sea ice and sea water (Perovich, 1996; Riggs et al., 1999; Hall et al., 2001; Riggs and Hall, 2015; Liu et al., 2016).

In order to further verify that the sea ice concentration retrieved by the improved ASI algorithm based on CGAN has higher accuracy than the ASI algorithm, the Landsat 8 OLI-L1T data is used to verify the results retrieved by the improved ASI algorithm based on CGAN and the ASI algorithm from. The Landsat 8 OLI-L1T data were all used under a clear sky, and the errors caused by cloud and water vapor in the processing process can be eliminated. Sea ice distribution obtained by the ASI algorithm, the improved ASI algorithm based on CGAN and the Landsat 8 OLI-L1T data shown in **Figures 5A–C** (February 1, 2020), **Figures 5D–F** (October 16, 2019), **Figures 5G–I** (November 10, 2019), and **Figures 5J–L** (March 6, 2020).

In **Figures 5A–C**, the accuracy of the sea ice distribution obtained by the improved ASI algorithm based on CGAN is

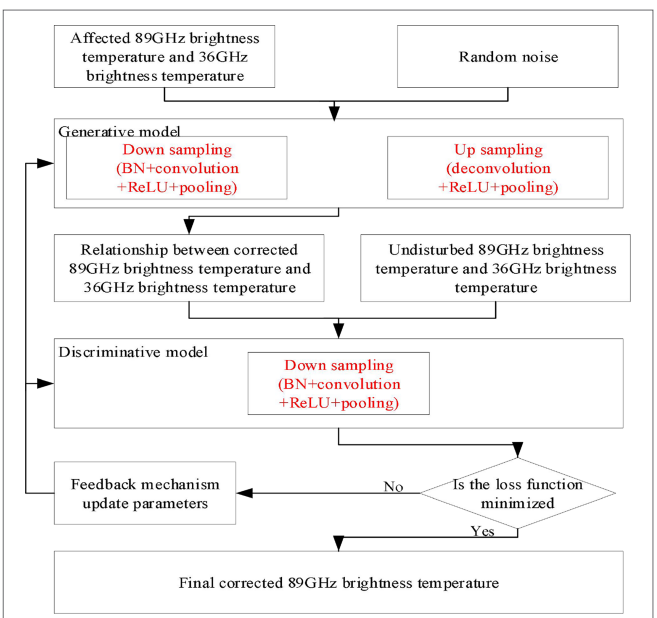


FIGURE 2 | Flowchart of the affected 89-GHz brightness temperature data correction based on CGAN model.

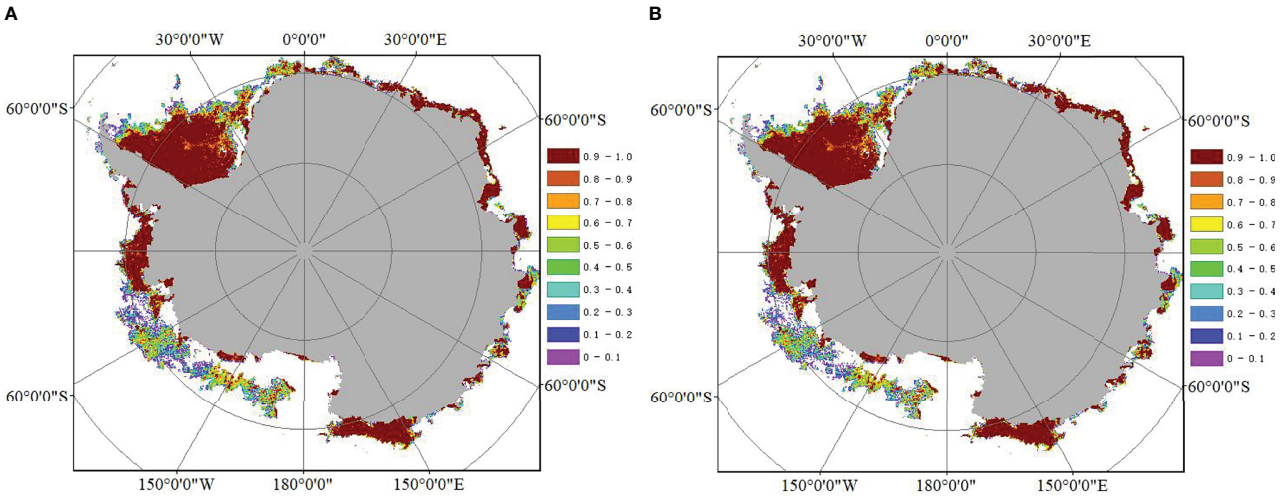


FIGURE 3 | (A) Sea ice concentration was retrieved by the ASI algorithm; **(B)** Sea ice concentration was retrieved by the improved ASI algorithm based on CGAN. (Projection: polar stereographic projection) Figure 3 showed sea ice concentration retrieved by the ASI algorithm and the improved ASI algorithm based on CGAN, respectively. By comparing the results in Figure 3, it can be seen that sea ice concentration are relatively similar. Then the extents of the multi-year ice, first-year ice and total sea ice from the U.S. National Ice Center (USNIC) data, the improved ASI algorithm based on CGAN and the ASI algorithm were comared on February 1, 2021 as shown in **Figure 4**. It can see from **Figure 4** that the sea ice extent of the improved ASI algorithm based on CGAN is between USNIC and the original ASI algorithm.

about 91%, and the accuracy of the sea ice distribution obtained by the ASI algorithm is about 83%. In **Figures 5D–F**, the accuracy of the sea ice distribution obtained by the improved ASI algorithm based on CGAN is 83%, and the accuracy of the sea ice distribution obtained by the ASI algorithm is about 33%. In **Figures 5G–I**, the accuracy of the sea ice distribution obtained by the improved ASI algorithm based on CGAN is about 83%, and the accuracy of the sea ice distribution obtained by ASI algorithm is about 50%. In **Figures 5J–L**, the accuracy of the sea ice distribution obtained by the improved ASI algorithm based on CGAN is about 78%, and the accuracy of the sea ice distribution obtained by ASI algorithm is about 89%. The selected areas in **Figure 5** are basically located in the edge areas of **Figure 3** with the low sea ice concentration or the interface between sea water and sea ice. That is to say, the accuracy of the improved ASI algorithm based on CGAN is higher than that of the ASI algorithm in the arers with the low sea ice concentration. Therefore, through the comparison of the above results, we can draw a conclusion that the improved ASI algorithm based on CGAN has higher accuracy.

5 DISCUSSION

Sea ice concentration was retrieved by the CGAN based improved ASI retrieval algorithm based on the Landsat 8 OLI-L1T data and the AMSR-E 89-GHz brightness temperature data in this paper. Compared with the sea ice distribution obtained by the ASI algorithm, the sea ice distribution obtained by the improved ASI algorithm based on CGAN was closer to the sea ice distribution obtained from the Landsat 8 OLI-L1T data,

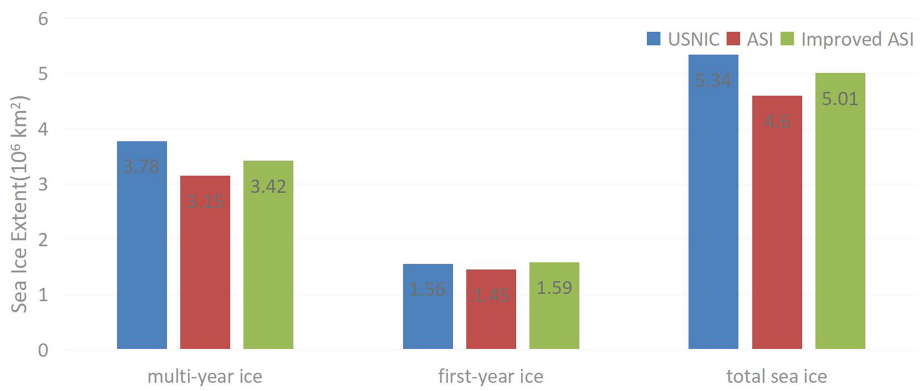


FIGURE 4 | Comparison of Antarctic sea ice extent.

so the improved ASI algorithm based on CGAN significantly improved the accuracy of sea ice concentration. The improved ASI algorithm based on CGAN made use of the reliable 36-GHz brightness temperature data, which greatly reduced the errors caused by the atmosphere, and the proposed method effectively corrected sea ice concentration of the pixels affected by the external environment.

At present, many scholars had studied the ASI algorithm for retrieving the sea ice concentration from the 89-GHz brightness temperature data. Although the ASI algorithm has advantages due to its higher spatial resolution, compared with the low-frequency brightness temperature data, the 89-GHz brightness temperature data are more affected by cloud and water vapor, which will lead to some errors in the edge regions of sea water and sea ice. Although some errors can be eliminated by the weather filter, the most of the errors in some regions cannot be removed. Based on the previous studies and the above reasons, we proposed the improved ASI algorithm based on CGAN. That is, we used CGAN to replace the weather filter in the ASI algorithm. The generative model of CGAN increases the utilization of the image feature information through the skip connection operation, which improves the removal of the influence of cloud and water vapor. The discriminative model can retain the image feature information and realize the non-linear mapping from the image to the image. The loss function can reduce the pixel-level loss, which can remove the influence of cloud and water vapor. The improved ASI algorithm based on CGAN can corrected the 89-GHz brightness temperature data affected by the external environment in the process of training. And the improved ASI algorithm based on CGAN greatly reduced the errors caused by

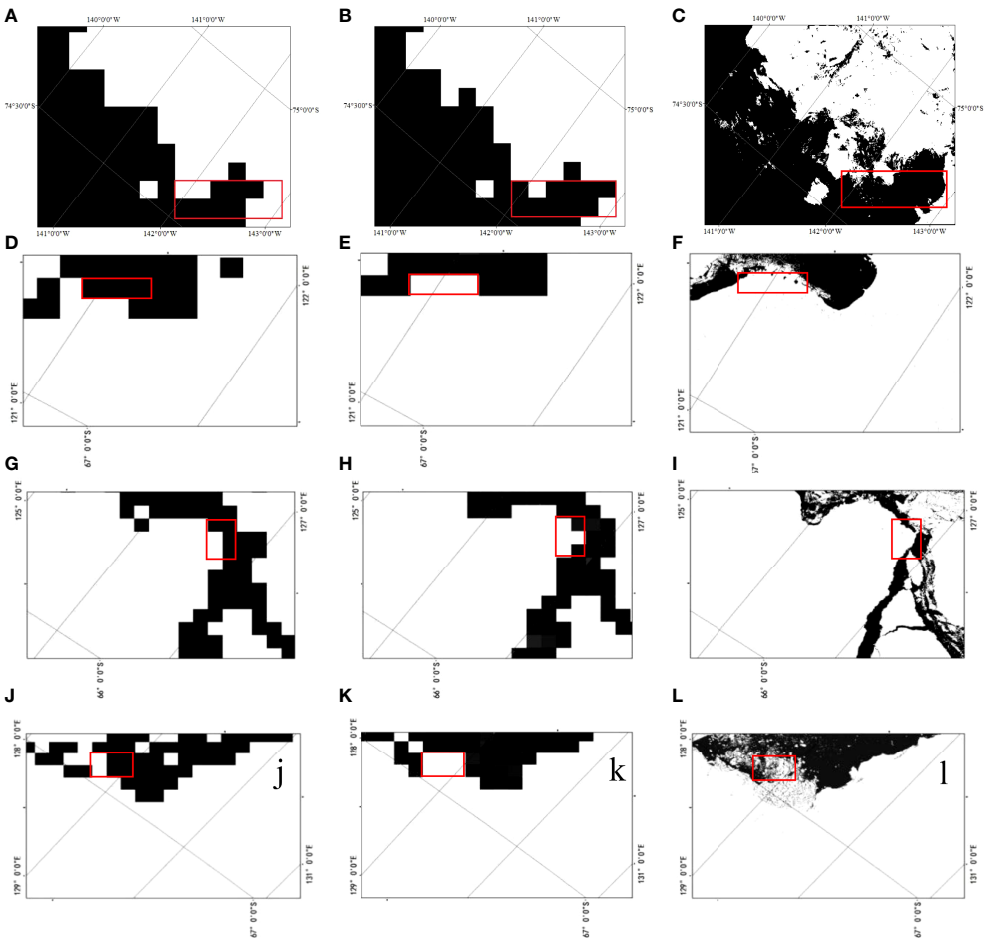


FIGURE 5 | (A) Sea ice distribution was obtained by the ASI algorithm on February 1, 2020 **(B)** Sea ice distribution was obtained by the improved ASI algorithm based on CGAN on February 1, 2020 **(C)** Sea ice distribution was obtained by the Landsat8 OLI-L1T data through the NDSI method on February 1, 2020. **(D)** Sea ice distribution was obtained by the ASI algorithm on October 16, 2019 **(E)** Sea ice distribution was obtained by the improved ASI algorithm based on CGAN on October 16, 2019 **(F)** Sea ice distribution was obtained by the Landsat8 OLI-L1T data through the NDSI method on October 16, 2019. **(G)** Sea ice distribution was obtained by the ASI algorithm on November 10, 2019 **(H)** Sea ice distribution was obtained by the improved ASI algorithm based on CGAN on November 10, 2019 **(I)** Sea ice distribution was obtained by the Landsat8 OLI-L1T data through the NDSI method on November 10, 2019. **(J)** Sea ice distribution was obtained by the ASI algorithm on March 6, 2020 **(K)** Sea ice distribution was obtained by the improved ASI algorithm based on CGAN on March 6, 2020 **(L)** Sea ice distribution was obtained by the Landsat8 OLI-L1T data through the NDSI method on March 6, 2020 (Projection: polar stereographic projection; The white area is sea ice, and the black area is sea water).

the atmosphere and significantly improved the accuracy of sea ice concentration.

However, the improved ASI algorithm based on CGAN has some limitations. Firstly, in the data screening stage, the relatively stable relationship between the 89-GHz brightness temperature data which is not disturbed or less affected by the external environment and the 36-GHz brightness temperature data is limited by the sample points. Secondly, there are the time difference between the Landsat 8 OLI-L1T data and the AMSR-E/AMSR-2 data, resulting in a certain error in the inversion of sea ice concentration. Finally, using the Landsat 8 OLI-L1T data has obvious advantages in verifying local small regions, but the Landsat 8 OLI-L1T data is not very suitable for large-scale and long-time series sea ice detection. Therefore, we will strive for breakthroughs in the following two aspects in future studies. Firstly, collect more representative sample points in order to get a more accurate screening model. Secondly, find more appropriate verification data (such as on-site data) to verify the results of sea ice concentration retrieved by the improved ASI algorithm based on CGAN.

6 CONCLUSIONS

In this study, the data correction method based on CGAN was used to correct the 89-GHz brightness temperature data affected by the external environment by the relatively stable relationship between the 89-GHz brightness temperature data which is not disturbed or less affected by the external environment and the 36-GHz brightness temperature data. This method effectively corrected sea ice concentration of the pixels affected by the external environment, which greatly reduced the errors caused by the atmosphere. The sea ice concentration was verified by the Landsat 8 OLI-L1T data. Firstly, the study determined the relatively stable relationship between the 36-GHz brightness temperature data and the 89-GHz brightness temperature data that were not disturbed or less affected by external environment, and we screened out the 89-GHz brightness temperature data with the large interference. Then, the data correction method based on CGAN corrected the 89-GHz brightness temperature data which was greatly affected by the external environment such as cloud and water vapor. Finally, the ASI algorithm was used to retrieve Antarctic sea ice concentration. Sea ice concentration obtained by the improved ASI algorithm based on CGAN was compared with sea ice concentration obtained by the ASI algorithm. The results showed that sea ice concentration retrieved by the improved ASI algorithm based on CGAN was close to that obtained by the ASI algorithm. We used sea ice concentration obtained from the Landsat 8 OLI-L1T data using the NDSI method to further verify the sea ice concentration retrieved by the improved ASI algorithm based on CGAN. The improved ASI algorithm based on CGAN significantly changed the sea ice concentration of the pixels affected by the external environment, so as to reduce the impact of cloud and water vapor on high-frequency data. Compared with sea ice concentration obtained by the ASI algorithm, sea ice concentration retrieved by the improved ASI algorithm based on CGAN had higher accuracy. The sea ice distribution obtained by CGAN does not need to design features in advance. For different data products, CGAN has the strong robustness and the migration ability.

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

Data curation, XW and YZ. Project administration, XW, DS, and YW. Resources, XW, YW, and SY. Software, SY. Supervision, XW and DS. Validation, XW and YZ. Writing—original draft, XW and YZ. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was supported by the Key Projects of Science and Technology Research of Henan Province (Grant No. 202102310334), the Key Scientific Research Project in Colleges and Universities of Henan Province of China (Grant No. 22A420004), the Cultivation Programme for Young Backbone Teachers in Henan University of Technology (Grant No. 21420070).

ACKNOWLEDGMENTS

We would like to acknowledge the work and efforts of the author who contributed to the manuscript.

REFERENCES

Cavalieri, D. J., Crawford, J. P., Drinkwater, M. R., Eppler, D. T., Farmer, L. D., Jentz, R. R., et al. (1991). Aircraft Active and Passive Microwave Validation of Sea Ice Concentration From the Defense Meteorological Satellite Program Special Sensor Microwave Imager. *J. Geophysic. Res.* 96 (C12), 21998–22008. doi: 10.1029/91JC02335

Cavalieri, D. J., Gloersen, P. and Campbell, W. J. (1984). Determination of Sea Ice Parameters With the NIMBUS 7 SMMR. *J. Geophysical. Res. Atmospheres.* 89 (D4), 5355–5369. doi: 10.1029/JD089iD04p05355

Comiso, J. C. (1986). Characteristics of Arctic Winter Sea Ice From Satellite Multispectral Microwave Observations. *J. Geophysical. Res.* 91 (C1), 975–994. doi: 10.1029/JC091iC01p00975

Comiso, J. C. (1995). "SSM/I Sea Ice Concentrations Using the Bootstrap Algorithm," in *NASA Reference Publication 1380* (Washington, D. C: National Aeronautics and Space Administration), ISBN: ISBN: 0-87590-263-4.

Hall, D. K., Riggs, G. A. and Salomonson, V. V. (2001) *Algorithm Theoretical Basis Document (ATBD) for the MODIS Snow and Sea Icemapping Algorithms[EB/OL]*. Available at: https://eospso.gsfc.nasa.gov/sites/default/files/atbd/atbd_mod10.pdf.

Hao, G. and Su, J. (2015). A Study of Multiyear Ice Concentration Retrieval Algorithms Using AMSR-E Data. *Acta Oceanologica. Simian.* 9, 102–109. doi: 10.1007/s13131-015-0656-1

Imaoka, K., Kachi, M., Fujii, H., Murakami, H., Hori, M., Ono, A., et al. (2010). Global Change Observation Mission (GCOM) for Monitoring Carbon, Water Cycles, and Climate Change. *Proc. IEEE* 98 (5), 717–734. doi: 10.1109/JPROC.2009.2036869

Iwamoto, K., Ohshima, K. I., Tamura, T., Nihashi, S. (2013). Estimation of Thin Ice Thickness From AMSR-E/AMSR-2 Data in the Chukchi Sea. *Int. J. Remote Sens.* 34 (2), 468–489. doi: 10.1080/01431161.2012.712229

Kaleschke, L., Heygster, G., Lüpkes, C., Vihma, T., Bochert, A., Hartmann, J. et al. (2001). SSM/I Sea Ice Remote Sensing for Mesoscale Ocean-Atmosphere Interaction Analysis. *Can. J. Remote Sens.* 27 (5), 526–537. doi: 10.1080/07038992.2001.10854892

Kern, S. (2001). *A New Algorithm to Retrieve the Sea Ice Concentration Using Weather-Corrected 85 GHz SSM/I Measurements* (Bremen: Dept.Physics Elect.Eng.Univ. Bremen), 1–5.

Kern, S. (2004). A New Method for Medium-Resolution Sea Ice Analysis Using Weather-Influence Corrected Special Sensor Microwave/Imager 85 GHz Data. *Int. J. Remote Sens.* 25 (21), 4555–4582. doi: 10.1080/01431160410001698898

Kern, S. and Heygster, G. (2001). Sea-Ice Concentration Retrieval in the Antarctic Based on the SSM/I 85.5 GHz Polarization. *Ann. Glaciology.* 33 (1), 109–114. doi: 10.3189/172756401781818905

Kern, S., Kaleschke, L. and Clausi, D. A. (2003). A Comparison of Two 85-GHz SSM/I Ice Concentration Algorithms With AVHRR and ERS-2 SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* 41 (10), 2294–2306. doi: 10.1109/TGRS.2003.817181

Knight, E. J. and Kvaran, G. (2014). Landsat-8 Operational Land Imager Design, Characterization and Performance. *Remote Sens.* 6 (11), 10286–10305. doi: 10.3390/rs61110286

Liu, Y. H., Key, J. and Mahoney, R. (2016). Sea and Freshwater Ice Concentration From VIIRS on Suomi NPP and the Future JPSS Satellites. *Remote Sens.* 8 (6), 523. doi: 10.3390/rs8060523

Liu, T., Liu, Y., Huang, X. and Wang, Z. (2015). Fully Constrained Least Squares for Antarctic Sea Ice Concentration Estimation Utilizing Passive Microwave Data. *Geoence. Remote Sens. Letters. IEEE* 12 (11), 2291–2295. doi: 10.1109/LGRS.2015.2471849

Lomax, A. S., lubin, D. and Whriner, R. H. (1995). The Potential for Interpreting Total and Multiyear Ice Concentrations in SSM/I 85. 5ghz Imagery. *RenoteSensing. Environ.* 54 (1), 13–26. doi: 10.1016/0034-4257(95)00082-c

Markus, T. and Cavalieri, D. J. (2000). An Enhancement of the NASA Team Sea Ice Algorithm. *IEE. Trans. Ceoscience. Renote. Sens.* 38 (3), 1387–1398. doi: 10.1109/36.843033

Perovich, D. K. (1996). *The Optical Properties of Sea Ice, CRREL Monogr. No. 96-1. U.S. Army Cold Regions Research and Engi-Neering Laboratory*. 25.

Riggs, G. A. and Hall, D. K. (2015) *MODIS Sea Ice Products User Guide to Collection 6[EB/OL]*. Available at: https://landweb.modaps.eos-dis.nasa.gov/QA_WWW/forPage/user_guide/MODISC6SeaIce-productsUserguide.pdf.

Riggs, G. A., Hall, D. K. and Ackerman, S. A. (1999). Sea Ice Extent and Classification Mapping With the Moderate Resolution Imaging Spectro-Radiometer Airborne Simulator. *Remote Sens. Environ.* 68 (2), 152–163. doi: 10.1016/S0034-4257(98)00107-2

Spencer, R. W., Goodman, H. M. and Hood, R. E. (1989). Precipitation Retrieval Over Land and Ocean With SSM/I: Identification and Characteristics of the Scattering Signal. *J. Atmospheric. Oceanic. Technol.* 6, 254–273. doi: 10.1175/1520-0426(1989)006<0254:PROLAO>2.0.CO;2

Spreen, G., Kaleschke, L. and Heygster, G. (2008). Sea Ice Remote Sensing Using AMSR-E 89-GHz Channels. *J. Geophysical. Res. Atmospheres.* 113 (113), C02S03. doi: 10.1029/2005JC003384

Su, J., Hao, G. H., Ye, X. X. and Wang, W. B. (2013). The Experiment and Validation of Sea Ice Concentration AMSR-E Retrieval Algorithm in Polar Region. *J. Remote Sens.* 17 (3), 495–513. doi: CNKI:SUN:YGXB.0.2013-03-005

Svendsen, E., Mauler, C. and Grenfell, T. C. (1987). A Model for Retrieving Total Sea Ice Concentration From a Spaceborne Dual-Polarization Passive Microwave Instrument Operating Near 90ghz. *Int. J. Remote Sens.* 8 (10), 1479–1487. doi: 10.1080/01431168708954790

Wang, H. (2009). *Multiyear Ice Retrieval Using Passive Microwave Remote Sensing Radiometer AMSR-E 89ghz Data (in Chinese)* (Qingdao: Ocean University of China).

Wu, Z., Wang, X. and Wang, X. (2019). An Improved ARTSIST Sea Ice Algorithm Based on 19 GHz Modified 91 GHz. *Acta Oceanologica. Sin.* 38 (10), 93–99. doi: 10.1007/s13131-019-1482-7

Zhang, S. (2012). *Sea Ice Concentration Algorithm and Study on the Physical Process About Sea Ice and Melt-Pond Change in Central Arctic (in Chinese)* (Qingdao: Ocean University of China).

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.

Copyright © 2022 Wang, Zhao, Yang, Wang and Shangguan. 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 Earth Science

TYPE Original Research
PUBLISHED 04 August 2022
DOI 10.3389/feart.2022.918123

Check for updates

OPEN ACCESS

EDITED BY
Yan Hao,
Shandong Normal University, China

REVIEWED BY
Qingsong He,
Huazhong University of Science and Technology, China
Lingxin Chen,
Yantai Institute of Coastal Zone Research (CAS), China
Mingxing Hu,
Southeast University, China

*CORRESPONDENCE
Jing Luo,
luojing@mail.ccnu.edu.cn

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing,
a section of the journal Frontiers in Earth Science

RECEIVED 12 April 2022
ACCEPTED 06 July 2022
PUBLISHED 04 August 2022

CITATION
Jiang L, Chen J, Zhang C, Tian Y, Wu G and Luo J (2022), Study on spatial evolution and mechanism of county economic differences in Hubei province from 2005 to 2020.
Front. Earth Sci. 10:918123.
doi: 10.3389/feart.2022.918123

COPYRIGHT
© 2022 Jiang, Chen, Zhang, Tian, Wu and Luo. 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.

Study on spatial evolution and mechanism of county economic differences in Hubei province from 2005 to 2020

Liang Jiang[1], Jie Chen[2], Chunyan Zhang[2], Ye Tian[3], Guoxi Wu[1] and Jing Luo[2]*

[1]College of Urban and Environmental Sciences, Xuchang University, Xuchang, China, [2]College of Urban and Environmental Sciences, Central China Normal University, Wuhan, China, [3]Institute for Advanced Studies in Finance and Economics, Hubei University of Economics, Wuhan, China

The spatial distribution pattern of the economic development among counties is an important external representation of a balanced and sustainable regional development in China. With the rapid development of globalization and localization, spatial pattern of economic growth is increasingly obvious. The mechanisms of regional economic growth in China are also gradually gaining attention. However, there is still a lack of research at the province and county levels. As a result, based on the per capita GDP of each county in Hubei province from 2005 to 2020 as the research index, the spatial autocorrelation and the spatial variation function are used to analyze the spatial pattern evolution and the county economy mechanism in Hubei province. The results show that 1) there is a remarkable phenomenon of county-level economic spatial agglomeration in Hubei province. The urban area of Wuhan and its surrounding counties are high–high (HH-type) county agglomeration areas. The low–low (LL-type) counties are mainly distributed in the western parts of Hubei province and scattered in the northeastern and southern parts of Hubei province; 2) the county economy of Hubei province presents a spatial distribution pattern of "high in the east and low in the west." The hot areas of the county economy are primarily located in the urban area of Wuhan and its surrounding areas. In the process of development, the hot spot areas tend to shift to Yichang, Jingmen, and Xiangyang. The cold spot areas are located on the edges of the western, northeastern, and southeastern areas of Hubei province; 3) the spatial continuity and self-organization of the county economic development are strengthened. The structural differentiation trend caused by spatial autocorrelation is also strengthened. The county economy is relatively balanced from the southeast to the northwest, and the spatial difference in economic development in other directions is increasing; and 4) the spatial evolution of county economic development in Hubei province is the result of the comprehensive effects of historical and cultural background, economic development, traffic location, and policy system, and the A-shaped point-axis structure is a reliable spatial structure for regional development in Hubei province.

KEYWORDS

county economic, spatial pattern evolution, spatial variogram, kriging interpolation, Hubei province

Introduction

The unbalanced distribution of economic activities and economic phenomena is a persistent problem for both developed and developing nations (Candelaria et al., 2009), which exhibits spatial and temporal patterns in geography. The spatial process and pattern make a difference to broad social, economic, and political processes (Massey, 1985). The objective existence of regional economic differences is determined by the local development path, economic structure, spatial linkages, and local resource endowments among different regions. China's economic prosperity over the past 40 years has made its crowded cities and metropolises rich, but rest of the countries, especially the counties with a predominantly agricultural and rural economy, has remained relatively poor (He et al., 2019), more intensifying inequalities are mainly demonstrated across prefectures and counties. Different regional economic spatial patterns are formed because of this objective existence (Jin et al., 2007; Li and Li, 2018). Hence, a need to focus on "enhancing the county-level economy" was proposed in 2002 and was emphasized in the "National territorial planning framework (2016–2030)," approved by Premier Li, which aims to extend urban development beyond urban areas and promote the development of county economy. As the most basic administrative unit and industrial undertaking carrier of China's economic development, the county is the basis of the construction of the regional economy, instead of the economic zone, province, city, *etc.* This specific pattern could be regarded as an important reference factor to promote a balanced and sustainable regional development and formulate the macro-control policy of the government (Ward, 2016). However, the ways in which regional economic differences is spatially formulated and connected at the county scale remains unknown. The research on the county spatial differences and its economic development mechanism is of great significance for understanding the regional economic pattern and its evolution and has become a major strategic problem that China's economic macro-control is committed to solving (Xu et al., 2005; Jin and Lu, 2009; Zhao and Dong, 2012).

The literature on regional economic differences is popular with terms related to geography, such as region, spatiality, locality, district, and neighborhood effect. Scholars have found that regional economic difference is sensitive to geographical clustering and agglomeration, as corroborated in many empirical analyses at regional, provincial, and city levels (Liao and Wei, 2012; Wei and Kim, 2002), and that changing trajectories and fortunes of leading or lagging regions often have a huge impact on regional inequality (Yu and Wei, 2003; Ye and Wei, 2005). Regional differences have become a burning issue in the development of the social economy and have attracted tremendous concern and scholarly attention (Li and Qiao, 2001; Xu et al., 2005). From the perspective of space, regional economic differences are spatially represented by the economic spatial structure of the region in a certain period of time. Most of the studies conducted internationally have focused on the theoretical level and have provided the basis for a classical theoretical system (Czako et al., 2014). Perroux (1950) proposed the theory of the regional growth pole in 1955, which is regarded as a great theoretical contribution to the study of regional development from the perspective of economics, and Friedman (1966) proposed the "core-periphery" structure based on the simulation of urban system formation and development to explain the evolution of regional economic spatial structure. New economic geography places increasing return to scale and agglomeration at the center of regional development (Fujita et al., 1999). Initial studies by foreign scholars have suggested that the regional differences in economic development were ultimately reflected in the problem of social polarization between rich and poor populations, so early scholars used the Gini coefficient to measure polarization (Atkinson, 1970). Esteban et al. (2007) measured the average annual Esteban–Ray polarization index of five OECD countries (Canada, Germany, Sweden, the United Kingdom, and the United States) from 1970 to 2000. Fedorov (2002) calculated and decomposed Russia's Esteban–Ray's and Kanbur–Zhang's polarization indexes of income inequality and consumption expenditure differences from 1990 to 1999. Ezcurra et al. (2006) used the Gini coefficient, Theil index, coefficient of variation, and Esteban–Ray polarization index to quantitatively analyze the overall differences and polarization level of the per capita income distribution dynamics in Europe from 1977 to 1999. Gasparini et al. (2006) found that the economic development of Latin American countries showed a highly polarized trend, while that of Caribbean countries was relatively flat (except Mexico, Nicaragua, and Peru). Zhang and Kanbur carried out a series of studies on regional economic differences and polarization, respectively, in China (Zhang and Kanbur, 2001, 2005; Kanbur and Zhang, 2005). Wan and Zhou. (2005) showed that since the 1980s, the differences in per capita net income among farmers in different regions, within regions, within provinces and cities, and even among farmers in the same township have been increasing.

At present, the research on regional economy in China mainly embodies the following aspects. ① The research object has changed from the differences between the east and the west (Lu et al., 1999; Wu D. Y., 2001; Lu and Xu, 2005), between the north and the south (Wu D. T., 2001), between urban and rural areas (Xu and Liu, 2012) to the differences between provinces

(Zhang et al., 2010; Xiong et al., 2011; Fang et al., 2013a; Tian and Zhao, 2013), and between economic zones and urban agglomerations (Ma et al., 2007; Sun et al., 2009; Hu and Zhang, 2010; Sun et al., 2013). ② The research scale has changed from the macro-level province (city, autonomous region) (Chen and Zhu, 2012) to the micro prefecture-level city and county and has begun to focus on mechanism analysis (Ke and Lu, 2011; Ke et al., 2013a; Wang and Gao, 2014). ③ Research methods have been integrated, including the traditional standard deviation (Wu and Ding, 2011), coefficient of variation (Xu et al., 2015), Gini coefficient (Cai et al., 2011), Theil coefficient (Zhong and Lu, 2010; Zhou and Zhang, 2011), and other introduced mathematical analysis methods such as GIS (Feng et al., 2015), ESDA (Fang et al., 2013b), wavelet analysis (Xu et al., 2005), and other methods to explore the spatial dependence and heterogeneity of the county economy. ④ There has been a shift from an emphasis on quantity to a focus on quality, and the factors and structure of the county have also become an important issue. Rozelle et al. (2002) have examined how rural China is successful in participating in the sequence of economic activities that will lead to modernization and emphasize the significance of the state of China's county economy. This issue has also been favored by some scholars (Ke et al., 2013b; Zhu and Wu, 2015), mainly focusing on research on land (Li and Fang, 2014), industry (Chen et al., 2016), urbanization (Liu and Yang, 2012), transportation (Liu and Zeng, 2011), ecology (Li et al., 2014), *etc.*

Regional economic difference is multi-scale in nature, and spatial agglomeration is an essential feature of geographical space. The significance of the spatial perspective has motivated a new round of methodological advances. With increasing amount of data and fast developing analytical methodologies, our understanding of regional economic differences (inequality) and spatial pattern is increasing drastically. This is particularly true when scholars started to apply spatial analytical approaches to understand the issue since the spatial distribution of economy is essentially a geographical problem (Dou et al., 2016; Ortega et al., 2018). World Development Report 2009 highlights the importance of economic geography in regional economic development (World Bank, 2009). Affected by various socioeconomic contexts, regional economic differences demonstrate many spatial features such as spatial heterogeneity (S. He, Fang, et al., 2017), spatial dependence (Anselin, 2013), spatial agglomeration (F. H. Liao and Wei, 2015), spatial mobility (Rey, 2016), spatial hierarchy, and spatial causation (Wei, 2015). Therein, spatial heterogeneity and spatial dependence are the two fundamental issues in social sciences (Goodchild, 2009). Regional economic differences are heterogeneous in both space and time. The effect of spatial dependence is also significant to understand structural instability in regional sciences. Geographical space or regions are heterogenous in nature, while a traditional inequality index measures the spread of the income distribution and fails to distinguish between convergence to the global mean and clustering around local means (Esteban and Ray, 1994). Declining overall inequality can mask rising polarization, with related concepts of geographical concentration, club convergence, and poverty trap. The development of GIS and spatial analysis has provided powerful tools to uncover the significant impact of spatial association and spatial heterogeneity on regional economic differences. Moreover, geographers also tend to pay less attention on rural economic differences, including inequality across counties and villages. Analyzing spatial association/heterogeneity and mechanisms of county-level can better explain the evolution trend of regional economy and deepen the understanding of regional inequality.

A spatial view is crucial for regional economic differences analysis, particularly in the case of the county-level China (Zhang and Dang, 2016). The relationships between regional inequality and its geospatial dimensions have been examined by a number of studies (Rey and Le Gallo, 2009; Liao and Wei, 2015). These studies investigate spatial distributional dynamics of regional economic differences, which conclude that space does matter in shaping uneven regional development. Spatial dependence, scale, and hierarchy are all significant for better understanding the complexity of China's regional economic differences. For example, Rey (2015) suggests that the role of spatial context does influence the distributional dynamics of regional inequality by comparing the case of Mexico and the United States. Wei (2015) found that mechanism of regional economic differences in China is mainly demonstrated in terms of first nature (physical geography) and second nature (agglomeration). Although these studies have investigated the county-level inequality, most of them mainly focus on an individual coastal province in China, and no formal consensus has been reached to date and a geospatial analysis on the county-level economic differences in central China is lacking. Meanwhile, the new-type urbanization is required to realize regionalization and optimize spatial patterns of regional development in China. Hence, a spatially explicit view of development dynamics in the county-level central China requires further research, how the intra-county spatial inequality has evolved over time and whether these development strategies of regions became effective remains unknown.

In the new era, the implementation of rural revitalization and the development and study of the county economy play an important role in promoting the overall planning of urban and rural areas, coordinating regions, and accelerating the development of the national economy. Hubei is one of the most developed areas in China, making it an important component of development strategies of the "Yangtze River economic zone" and "Triangle of central China." In 2020, the county population and total economic volume were

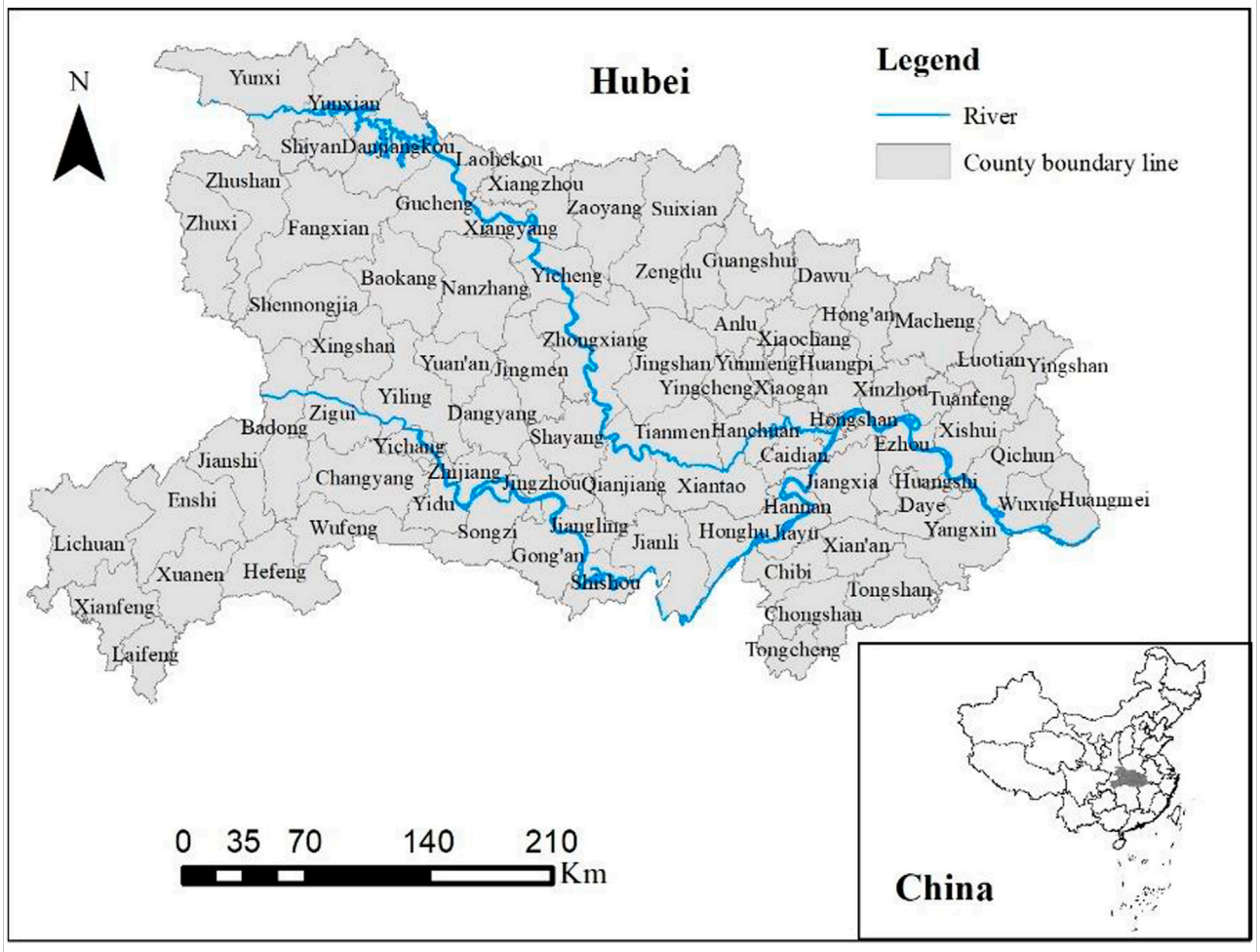


FIGURE 1
Location of the study area.

36.7536 million and 2,554.772 billion yuan, accounting for 59 and 60% of the province, respectively. With the rapid development of economy, the level of urbanization in Hubei has been continuously improved. Meanwhile, the spatial difference of county economic development in Hubei province is apparent. In the plain area along the Yangtze River in eastern Hubei with Wuhan as the core, large amounts of land resources are developed for industrial, commercial, residential, and infrastructure purposes. Conversely, the land development is restricted or forbidden in most counties of the middle and western Hubei. So there is an obvious regional gradient of economic growth from the eastern to the central and to the western Hubei, which is extremely similar to that of China. As such, Hubei province can be considered a miniature of economic development in China. In addition, Hubei province is one of the first regions to implement the reform of the financial system of the province directly governing county. This region could be regarded as a typical case for research. Moreover, the previous research works on the county economic differences in Hubei province only focused on the absolute spatial differences of the regional economy in different time sections, and there were few quantitative studies on the randomness, correlation, and structural factors affecting the spatial differences within the region, which failed to better reveal the spatial evolution process and mechanism of economic growth differences in Hubei province. Therefore, this study attempts to make contributions in the following aspects. First of all, a spatiotemporal analysis of county economy in Hubei province is conducted to examine their dynamic changes. Second, this study combines spatial heterogeneity and spatial dependence with spatial variability functions to analyze the formation mechanisms causing county-level economic differences. Finally, the results can offer novel thoughts for drafting future political strategies for regional development and rural revitalization.

2 Materials and methods

2.1 Study area and data source

Hubei province is located in the middle reaches of the Yangtze River, covering an area of 185,900 km^2, with a high terrain in the west and low terrain in the east (Figure 1). At the end of 2020, the total population was 57.75 million, with a GDP of 4,344.34 billion yuan, ranking eighth in China. The highest per capita GDP was that of Huangpi District (87,988 yuan), and the lowest was that of Hefeng County (14,583 yuan), with a difference of six times. This indicates that there is a large imbalance between county economies in Hubei province.

In this article, we chose the years 2005–2020 as the study period. During 2005 to 2020, the urbanization of Hubei province was developing rapidly, and the gap between urban and rural areas was more remarkable. For better revealing the change characteristics of county-level economy, we selected the year of 2005, 2010, 2015, and 2020 as the study time point to present their spatial–temporal changes. Then a county of Hubei province is taken as the research unit, and the per capita GDP of the county is used as the research indicator to reflect the level of economic development of the county; based on the integrity, scientificity, and availability of the data, some municipal jurisdictions are subjected to the merge application, and a total of 91 county units are obtained from 10 prefecture-level urban areas, 17 municipal districts, 24 county-level cities, and 40 counties. For the changed administrative division unit (such as the establishment of Suixian county in Zengdu District in 2009, and the transformation of Yunxian County to Yunyang District in 2014), the corresponding processing is performed in ArcGIS. The data used in this article were obtained from the China Statistical Year Book for Regional Economy, Hubei Statistical Year Book, and the China County Statistical Yearbook from 2006 to 2021.

2.2 Research method

2.2.1 Measurement method of county economic differences

1) Standard deviation and coefficient of variation: the absolute difference in per capita GDP of counties in Hubei province is calculated with the standard deviation, and the formula is $\sigma = \sqrt{[\sum_{i=1}^{n}(y_i - \bar{y})^2]/n}$. The coefficient of variation represents the relative degree of change in geographical data, which can measure the degree of differences between regions. Therefore, the coefficient of variation is used to measure the relative difference in per capita GDP of counties in Hubei province. The formula is $CV = \sqrt{[\sum_{i=1}^{n}(y_i - \bar{y})^2]/n}/\bar{y}$.
2) Gini coefficient and skewness coefficient: the Gini coefficient is the most commonly used index to measure the degree of regional imbalance. The geographical concentration of economic development is actually a kind of regional imbalance. The formula is as follows:

$$G = \frac{1}{2n^2\mu}\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j|. \quad (1)$$

The value range of the Gini coefficient is [0, 1]. The more uniform the geographical distribution of county economic development is, the smaller the Gini coefficient of location is; the higher the concentration of county economic development is, the greater the Gini coefficient of location is.

The skewness coefficient measures the asymmetric distribution of geographic data distribution and describes the skewness centered on the average value. The formula is $g = \sum_{i=1}^{n}\frac{1}{n}(\frac{y_i-\bar{y}}{\sigma})^3$. A value of the skewness coefficient greater than zero indicates that the economic differences of each county are distributed in a right state; a value of the skewness coefficient less than zero indicates that the economic differences of each county are distributed in a left state.

In the aforementioned measurement methods, y_i is the per capita GDP of the j-th county, $\bar{y}$ is the per capita GDP of the i-th county, y_j is the average per capita GDP of all counties, n is the number of counties, and μ is the average value of county y_i.

2.2.2 Exploratory spatial data analysis

The exploratory spatial data analysis is a combination of statistical principles and graphical charts to study the non-randomness or spatial autocorrelation of spatial information and reveal the spatial pattern of regional economic development, including global autocorrelation and local autocorrelation.

1) Global autocorrelation: global autocorrelation is used to describe the spatial distribution characteristics of object attributes in the whole region. It is measured by the global index $Moran'I$ and the calculation formula for this index is as follows:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)}{S^2\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}}, \quad (2)$$

where $S^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$; w_{ij} are the spatial weights; n is the total number of counties; y_i and y_j are the average per capita GDP of counties i and j, respectively; and $\bar{y}$ is the mean of the average per capita GDP of all counties. $Moran'I$ varies between -1 and 1. If $Moran'I > 0$ and passes the significance test, it indicates that there is a positive correlation in the spatial distribution of county economy. If $Moran'I < 0$ and passes the significance test, it indicates that there is a negative correlation in the spatial distribution of county economy. If $Moran'I$ is close to 0, it indicates a random pattern or absence of spatial autocorrelation.

2) Local autocorrelation: local autocorrelation is used to analyze the degree of association among each spatial element attribute and its adjacent spatial element attribute, which is measured by the index $Moran'I$ and the local G statistic. The calculation formula of $Moran'I$ is as follows:

$$\text{Local}\ \ I = \frac{\left(y_i - \bar{y}\right)}{S^2}\sum_{j=1}^{n} w_{ij}\left(y_j - \bar{y}\right). \tag{3}$$

When $Moran'I > 0$ and passes the significance test, the counties with high per capita GDP in Hubei province are adjacent to each other (this type of cluster was recorded as HH-type), or the counties with low per capita GDP are adjacent to each other (this type of cluster was recorded as LL-type); when $Moran'I < 0$ and passes the significance test, the counties with high per capita GDP are adjacent to the counties with low per capita GDP (this type of cluster was recorded as HL-type), or the counties with low per capita GDP are adjacent to the counties with high per capita GDP (this type of cluster was recorded as LH-type).

The Local G statistics can further detect the local spatial dependence and determine the location of high value or low-value elements clustering in space, and the calculation formula for this statistic is as follows:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}(d)y_j}{\sum_{j=1}^{n} y_j}. \tag{4}$$

Standardized G_i^* $Z(G_i^*) = (G_i^* - E(G_i^*))/(\sqrt{VAR}(G_i^*))$, $E(G_i^*)$ and $VAR(G_i^*)$ represent mathematical expectations and variation coefficients, and $w_{ij}(d)$ represents spatial weights. A significant positive $Z(G_i^*)$ value indicates that counties with high per capita GDP tend to gather and form hot spots, while a significant negative $Z(G_i^*)$ value indicates that counties with low per capita GDP tend to gather and form cold spots.

2.2.3 Spatial variogram

A spatial variogram is a basic means to describe the randomness and structure of regionalized variables (Liu et al., 2009). Let $Z(x)$ be the value of system attribute Z at the spatial position x, and $Z(x)$ be a regionalized random variable. h is the space separation distance of two sample points, and $Z(x_i)$ and $Z(x_i + h)$ are the observation values of regionalized variable $Z(x)$ at space position x_i and $x_i + h$, respectively; (x = 1,2,..., $N(h)$), thus, the spatial variogram can be expressed as:

$$\gamma(h) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)}[Z(x_i) - Z(x_i + h)]^2, \tag{5}$$

where $N(h)$ is the total number of sample point pairs when the separation distance is h. For different space separation distance h, the corresponding $\gamma(h)$ value can be calculated. When the $\gamma(h)$ value becomes larger, the spatial autocorrelation tends to weaken. Taking h as the abscissa and $\gamma(h)$ as the ordinate, the spatial variation function graph can be drawn to represent the spatial variation characteristics of the regionalized variable $Z(x)$ (Figure 2).

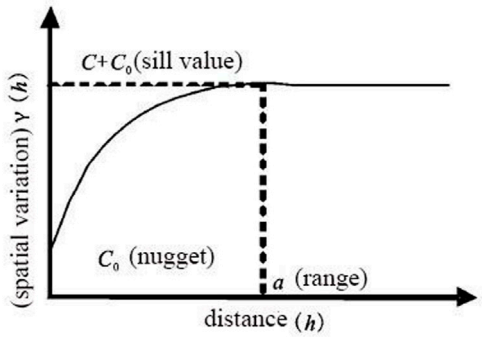


FIGURE 2
Model variogram.

With the increase in h, when the variogram $\gamma(h)$ reaches a relatively stable constant from a non-zero value, the constant is the sill value $(C + C_o)$, and C is the structural variance; when $h = 0$ and $\gamma(h) = C_0$, the value represents the nugget value, which represents the discontinuous variation in the regionalized variable when it is smaller than the observation scale. α is the range of variation, that is, the interval distance when the variogram $\gamma(h)$ reaches the sill value. After the range $\alpha \geq$ x, the spatial correlation of regionalized variables disappears. The fourth parameter of the variogram is the fractal dimension, which is used to express the characteristics of the variogram. It is determined by the relationship between the variogram $\gamma(h)$ and the interval distance h.

$$2\gamma(h) = h^{(4-2D)}. \tag{6}$$

The fractal dimension D is the slope in the linear regression equation of double logarithm. Its size represents the curvature of the variogram. It can be used as a measure of random variation. The intensity of spatial autocorrelation between different variables is compared. The closer its value is to 2, the more balanced the spatial distribution among variables is. The variogram is theoretically unknown and can be fitted by calculation. Commonly used fitting models include the linear model, exponential model, spherical model, Gaussian model, and power function model.

3 Results

3.1 Temporal characteristics of county economic differences

The change trend of county economic differences in Hubei province from 2005 to 2020 is shown in Figure 3. Among them, the

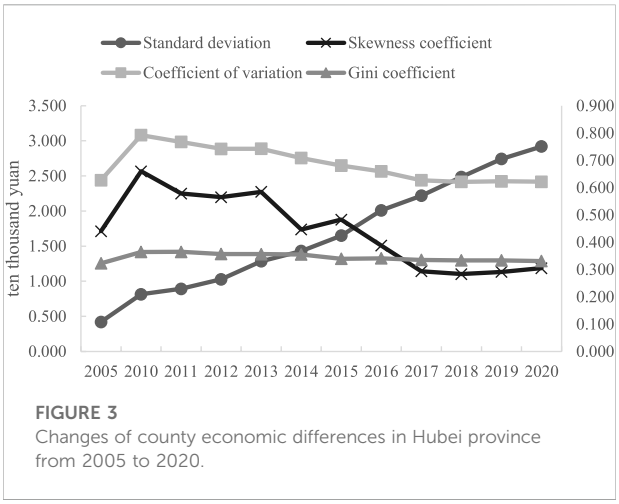


FIGURE 3
Changes of county economic differences in Hubei province from 2005 to 2020.

standard deviation continued to expand from 0.420 in 2005 to 3.075 in 2020, which reflected the absolute difference in county economic development in Hubei province have been on the rise trend. Based on the change trend of the coefficient of variation and the Gini coefficient, the relative difference in county economic development in Hubei province have been decreased generally during the study period. Meanwhile, 2010 was an important node of relative difference change. From 2005 to 2010, the Gini coefficient and coefficient of variation increased from 0.323 to 0.627 and from 0.365 to 0.793, respectively, and the relative difference in economic development among counties in Hubei province has expanded. After 2010, the relative difference in the county economy shrank steadily, and the Gini coefficient and coefficient of variation shrank to 0.329 and 0.615 in 2020. The main reason is that after 2010, with the implementation of the rise of central China, the Belt and Two Circles (the Yangtze River Economic Belt, Wuhan Metropolitan Area, and Western Hubei Eco-cultural Tourism Circle) and the construction of provincial sub-central cities, these development strategies and policies have effectively promoted the coordination of regional development in the province and continuously reduced the relative difference of county economy. Regarding the skewness coefficient, its value is greater than zero and declines in fluctuations, indicating that the economic differences among the counties in Hubei province present a right-skewed distribution, and the number of counties greater than the average GDP per capita of the province gradually increased, further indicating that the relative difference in economic development between the two countries has been shrinking.

3.2 Spatial characteristics of county economic differences

Figure 4 depicts the spatiotemporal patterns of county economy in Hubei province from 2005 to 2020. Although county economy has gained significant increase since 2000s, the spatial pattern was imbalanced and spatial agglomeration of county economy could be observed during this time period. Overall, the group of rich counties in red tended to cluster in Wuhan, Yichang, and Xiangyang, which explained the strong positive association between geographic agglomeration and economic growth documented by some scholars (Van Oort, et al., 2012). As the economic core area and growth pole of Hubei province, Wuhan was very stable; Yichang and Xiangyang were gradually taking shape as the growth poles of southwest and northwest Hubei. These three regions attracted the agglomeration of economic activities in metropolitan regions. Indicated by endogenous growth theory, the localized spillover effect would bring economic activities concentrating in neighboring rural counties because of low transport cost and easy access to innovation technologies. The group of poor counties in blue tended to cluster in mountainous areas of northwest, southwest, and northeast Hubei. Spatial agglomeration of these poor counties largely increases regional poverty because of their lack of growth poles and little mobility of capital and labor. These spatially clustered poor areas were mainly located in ecology fragile regions, such as mountainous areas, hilly areas, and restrictive/prohibited development areas.

In addition, the urban constellation theory indicates that the strategic design of sustainable regional development cannot ignore the differences between urban and rural areas, but should take the city (county) as the "star" of regional development, and connect many urban stars by the transportation axis to form a ring-shaped belt called the "City Constellation System" (Lewis, 1996). Figure 5A shows a night light map of cities (counties) in Hubei province in 2020 (http://59.175.109.173:8888/app/login.html, accessed on 21 December 2020), where the brightness of night lights can reflect the level of regional economic development and the scale of city stars. It can be seen that the brightest "first-class star" is Wuhan, and the second-brightest "second-class stars" include cities such as Xiangyang, Yichang, Ezhou, Shiyan, and Jingzhou. Among these bright cities, there were also bright lines along highways and railways. Connecting these bright cities along the bright lines constituted the "urban constellation system" of Hubei province: the ring-shaped belt connects almost all cities in Hubei, with dense cities and towns and intensive economic activities, and the central area surrounded by the ring-shaped belt is the Jianghan Plain. Therefore, with Wuhan as the vertex, an axis connects Yichang and Enshi to the southwest hinterland of Hubei, and an axis connects Xiangyang and Shiyan to the northwest hinterland of Hubei, forming an "A-shaped point-axis" spatial structure (Figure 5B). "A-shaped point-axis" covers 65 counties in Hubei province, not only connecting the three gradient regions of east, middle, and west of Hubei province but also taking into account the synchronous advancement of southwest and northwest Hubei, which is conducive to the coordinated development of Hubei province.

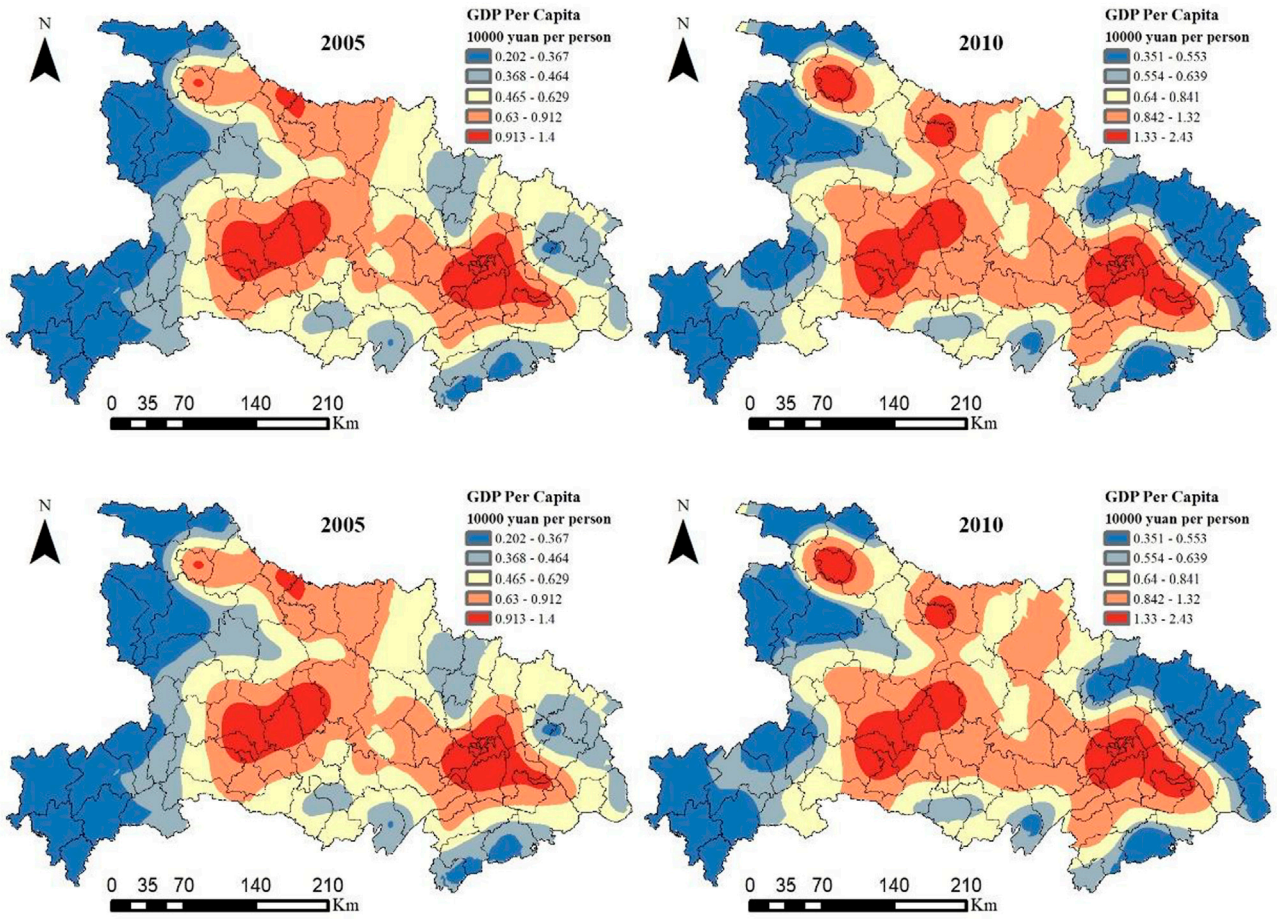


FIGURE 4
Spatial patterns of county economy in Hubei province from 2005 to 2020. (blue denotes the poor counties; gray denotes the less-developed counties; yellow denotes the well-developed counties; orange denotes the more-developed counties; and red denotes the rich counties).

3.3 Spatial connection of the county economy

3.1.1 Global spatial characteristics

The global Moran's I was computed to capture the overall tendency of geographical concentration of county economic development in Hubei (Figure 6). We can see that the global Moran's I fluctuated slightly but maintained a gradual rising trend, increased from 0.318 in 2005 to 0.444 in 2020 and all are significant at the 0.05 level. On the one hand, this result implied that the spatial distribution of county economy in Hubei is not random but showed a significant spatial agglomeration, that is, counties with high (or low) level tend to be adjacent to each other, and this agglomeration trend continues to strengthen over time. On the other hand, the increasing clustering of county economy possibly implied that the localized regional inequality across counties was decreasing. The reason is that the spatial pattern of county economic development in Hubei province is that developed areas are concentrated in eastern Hubei and relatively underdeveloped areas are concentrated in western Hubei. The reduction of the internal spatial difference between eastern and western Hubei makes the overall spatial difference of the county continuously narrow. This conclusion is that for the county-level spatial scale, it is the reduction of the spatial difference between counties (cities and districts) in the average sense, which is not contradictory to the expansion of the spatial difference on other scales. The localized spatial spillovers through rich counties pulling neighboring poor counties do make effective, resulting in the agglomeration extent of rich or well-developed counties being increased. Meanwhile, the localized clustering of poor counties tends to decrease by some poor counties getting rid of poverty. Combining coefficient of variation and Gini coefficient, they have been decreased from 2005 to 2020 despite one slight increase from 2005 to 2010. It indicated that the regional economic difference in Hubei province showed a decreasing trend at the county-level. With

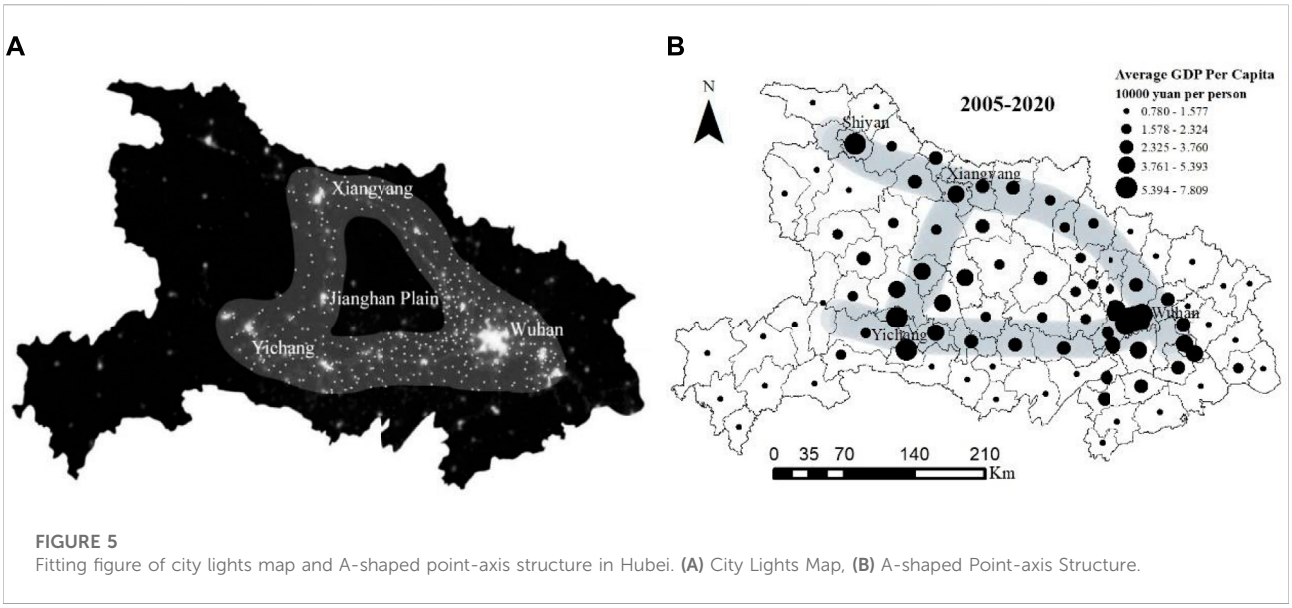


FIGURE 5
Fitting figure of city lights map and A-shaped point-axis structure in Hubei. **(A)** City Lights Map, **(B)** A-shaped Point-axis Structure.

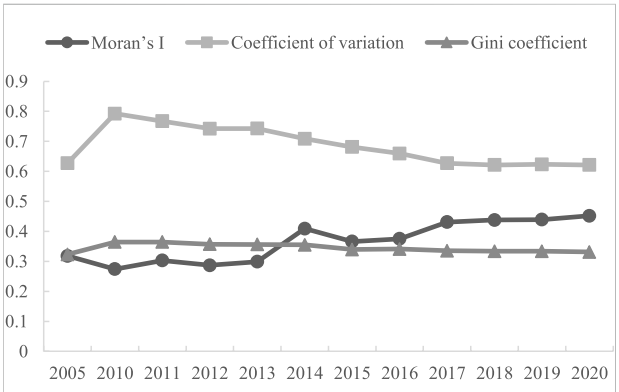


FIGURE 6
Global Moran's I of county level GDP per capita in Hubei, 2005–2020.

the economic transformation and development of Hubei and the implementation of the national targeted poverty alleviation strategy and eco-compensation mechanism, in the long run, the localized regional inequality across counties tends to be narrowing.

3.3.2 Local spatial characteristics

To further analyze the spatial agglomeration characteristic of economic development across counties, we draw their distribution maps with 2005, 2010, 2015, and 2020 as time nodes. Specifically, high–high (HH) means the counties with high per capita GDP are surrounded by other counties with high value, low–low (LL) means the counties with low per capita GDP are surrounded by other counties with low value; low–high (LH) indicates that the counties with low per capita GDP are surrounded by other counties with high value. High–low (HL) shows that the counties with high per capita GDP are surrounded by other counties with a low value.

Results in Figure 7 showed that most counties were distributed in HH and LL. This implied that county economy had a positive spatial correlation. 1) Spatial clustering patterns observed in Wuhan were dominated by HH; HH is continuously strengthened in this region and it did not undergo any apparent changes from 2005 to 2020. Wuhan, as a primary and provincial capital city, is better able to access various forms than ordinary counties in Hubei; its economic development has been more rapid, and industrial development has more diversification than those observed in the other areas. To support socioeconomic development, such cities must exert their influence to agglomerate various production factors and direct policy-making. HH in Yichang city and its surrounding counties was also strengthened except the year of 2010. Yichang showed inherent advantages based on a collection of production factors due to their status, the "Bottom-up" urbanization has strengthened linkages between local counties. In addition, infrastructure is developing rapidly to form a complete and competitive transportation network, which has benefitted regional spatial integration and coordinated development. 2) LL patterns were observed in southwest, northwest, and northeast Hubei, this is likely related to natural features in the west and east Hubei, which is characterized by mountainous terrain and large spatial distances between counties. 3) The LH pattern only fall in Yuan'n County in 2005, whereas the regions transform into Nanzhang County and Caidian District; the HL pattern was distributed in Shishou City in 2005 and Shiyan City

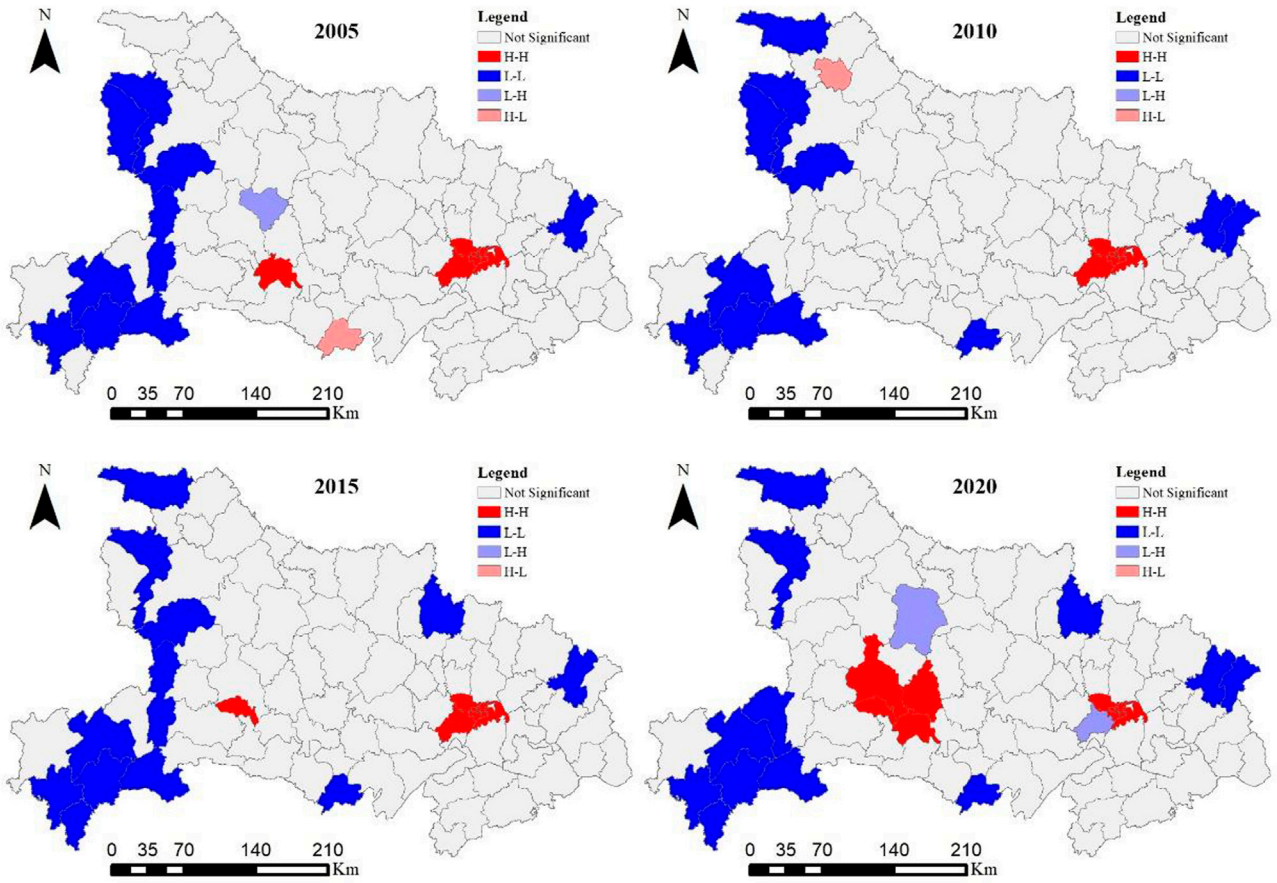


FIGURE 7
Local indicators of spatial association (LISA) of county economy in Hubei province.

in 2010, but not in other years. Shiyan City is the central city in the hinterland of northwest Hubei, with automobile and tourism as the leading industries. It has the geographical advantage of connecting Hubei, Shaanxi, and Henan provinces and the policy advantage of the water source of the mid-route south-to-north water diversion project, which exhibited HL.

3.3.3 Hot spots characteristics

This article used Getis-ord GI* in ArcGIS to identify the location of the high-value (hot spots) and low-value (cold spots) spatial clustering of counties in Hubei province. We divided the spatial agglomeration of economic development in 2005–2020 into four categories by the natural breakpoint method: hot spot area, sub-hot spot area, sub-cold spot area, and cold spot area. As illustrated in Figure 8, the spatial pattern of county economy in Hubei province is relatively stable, showing an obvious circle structure: Wuhan and its neighboring counties were hot spots, which form "hot spots, sub-hot spots, sub-cold spots, and cold spots" from outward to the edge of the province. Specifically, Wuhan and its neighboring such as Huanggang, Xiaogan, and Ezhou were always hot spots of the county economy from 2005 to 2020, and the economic development level of these counties was similar and higher than that of surrounding counties. By 2020, the number of the hot spots increased to 14. In the meantime, the hot spots moved to the southwest, adding Yichang, Zhijiang, Yuan'an, and Dangyang, which became a new core area of county economic development. The cold spots were mainly distributed in the western, northeastern, and southeastern of Hubei. From 2005 to 2020, the edge of western Hubei extended from Yunxi, Zhushan, and Zhuxi in the north to Xianfeng and Laifeng, forming a stable continuous strip covering 13 counties. In 2005, Guangshui, Anlu, Dawu, Xiaochang, Macheng, Luotian, and Yingshan in the northeastern of Hubei were cold spots. After 2005, the area presented strip-like distribution, increasing Hongan, Qichun, Wuxue, and Huangmei. In the southeastern of Hubei, Gongan, Jiangling, Shishou, Jianli, Chongyang, and Tongshan have been cold spots.

The sub-hot spots and sub-cold spots changed significantly, showing the characteristics of mutual transformation among

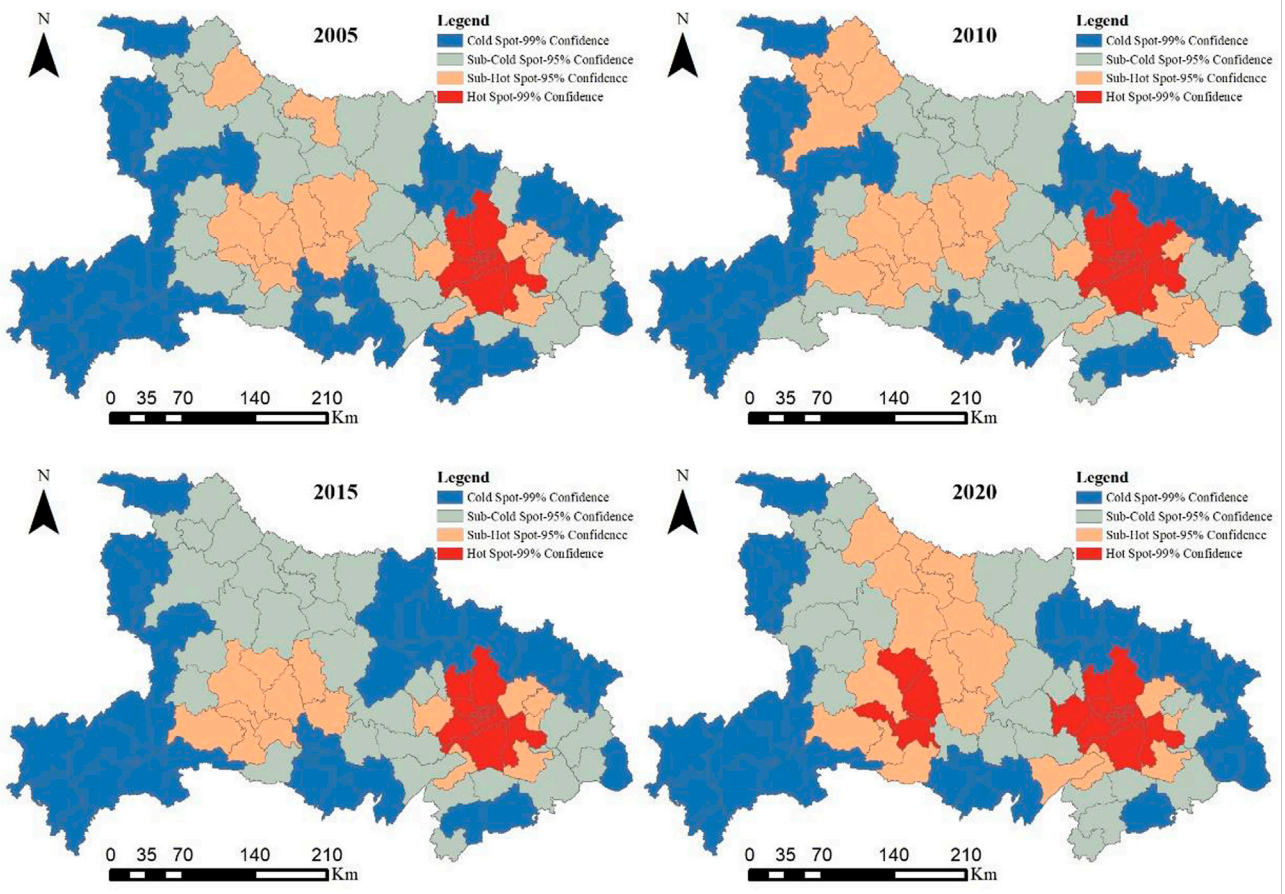


FIGURE 8
Distribution of county economy hot spots with different confidence levels from 2000 to 2016.

counties. In 2005, there were three sub-hot spots: one was the surrounding areas of Wuhan in eastern of Hubei, which was composed of six counties; the next was the central areas, which was composed of eight counties, including Yichang and Zhongxiang, to the west of the hot spots; and the third was Danjiangkou and Xiangyang in northwestern of Hubei. In 2010, Xinzhou withdrew from the sub-hot spots, adding Yangxin, and the number of central sub-hot spots increased to 10. However, the sub-hot spots in the northwest moved northward, increasing to five counties. In 2015, the sub-hot spots in the northwest disappeared, leaving only two regions in eastern and central of Hubei. In 2020, the sub-hot spots expanded again, from Danjiangkou and Laohekou to Yichu and Songzi and other 14 counties southward, covering most of the central counties. The sub-hot spots in eastern of Hubei were basically stable.

Based on the aforementioned analysis, the hot spots analysis demonstrated that the county economic development in Hubei is becoming more closely related in space, and the relationship between county economic development in neighboring areas has mutual influence and infiltration. Overall, the county economy of Hubei presents a spatial distribution pattern of "low in the west, high in the east" and "dual core structure." In terms of the change of years, the hot spots tend to move to Jingzhou–Xiangyang–Yichang urban agglomeration, demonstrating that the central and western Hubei economic circle will gradually become the growth pole of county economy, reflecting the trend of the gradual narrowing of county economic differences in Hubei province.

3.4 Mechanism of spatial pattern evolution of the county economy

"Pattern–process–mechanism" is the basic paradigm of geographical research on geographical phenomena and laws (Fu, 2014). The spatial variation function is used to investigate the internal mechanism of the spatial pattern evolution of the county economy in Hubei province. In 2005, 2010, 2015, and 2020, with the GDP per capita as the research data, the sampling step is set as 40 km (the size of the step

TABLE 1 Parameters of the variogram model on the spatial pattern of county economy in Hubei province.

Year	2005	2010	2015	2020
Range	82.6	55.59	92.2	114.2
Nugget value	0.058	0.078	0.079	0.082
Sill value	0.242	0.344	0.393	0.428
Nugget coefficient	0.239	0.226	0.201	0.191
Fitting model	Spherical	Gaussian	Spherical	Spherical
Coefficient of determination	0.731	0.589	0.697	0.811

multiplied by the number of steps is approximately 0.5 times the maximum distance between sample points), and the experimental variogram is calculated (that is, the experimental variogram is fitted from the effective spatial samples). For the sample data, the sphere model, Gaussian model, exponential model, linear model, and other models are used for fitting. Finally, the model with the highest fitting degree is selected, and the fractal dimension on different sides of each year is calculated, and then Kriging interpolation is carried out. The results are as follows:

1) It can be seen from the changes in the sill value, nugget value, and nugget coefficient that the spatial differences in county economic development in Hubei increased, the sill value increased from 0.242 in 2005 to 0.428 in 2020, and the nugget value increased from 0.058 in 2005 to 0.082 in 2020. However, the nugget coefficient showed a declining trend, from 0.239 in 2005 to 0.191 in 2020, which showed that in the spatial pattern evolution of the county economy in Hubei, the spatial differentiation of the economy caused by random factors such as data variation was decreasing, while the structural differentiation of the economic spatial pattern caused by spatial autocorrelation was increasing, and the significant spatial agglomeration exists in county economy (Table 1). Meanwhile, the parameters of the range were increasing, indicating that the scope of the spatial effect of the county economy in Hubei province was gradually expanding.
2) From the variance fitting chart, we can see that the per capita GDP of 4 years decreased first and then increased under the given step length, showing that the scope of the spatial correlation effect caused by the structural spatial gradient of county economic development in Hubei province is expanding, and the economic development tends to spread to the hot spots. According to the fitting equation of the spatial variation function selected by the least square method, the fitting effect of the spherical model and the Gauss model was better. The development of the county economy in Hubei province had different structural characteristics in different periods. The decision coefficient of model fit first decreased and then increased, indicating that the randomness of county economic development is obvious, but with the passage of time, the spatial self-organization of county economic development is strengthened, and the spatial differentiation of the economy is obvious.

The spatial differentiation of regional economy originates from the spatial heterogeneity or heterogeneity of factors affecting economic activities. Krugman. (1991) pointed out that an accidental event in history may lead to spatial self-organization behavior because this accidental event will inevitably bring about changes in certain factor endowments, and the heterogeneous space generated by such changes in factor endowments is suitable for the development of certain industries. This spatial self-organization promotes the economic growth and development of the region and become hot spots. There are structural characteristics in the development of county economy in Hubei province. Economic elements are linked together in a certain form of organization or combination and interact with each other to produce spatial correlation. This spatial correlation and self-organization structure dynamic system act on the regional economic system and promote the evolution and development of the regional economic system. With the continuous improvement of the level of social and economic development, the form of regional economic spatial organization has gradually evolved from low-level to high-level, and the pattern of regional economic spatial differentiation has shown the law of evolution and replacement from low-level equilibrium to high-level equilibrium or network. In addition, the institutional change within the country or region is an important mechanism that affects the spatial differentiation of regional economy. The marketization process in various regions of Hubei province has an asymmetric impact on the regional spatial pattern, forming a differentiated economic spatial gradient.

3) According to the fractal dimension of the spatial variation function (Table 2), the fractal dimension of all directions first increased and then decreased and increasingly deviated from the ideal value 2, and the determination coefficient first decreased and then increased, indicating that the difference in county economic development in Hubei province decreased before 2010 and gradually increased after 2010, and the spatial differentiation scale is also rising. In addition, the decrease in the degree of homogeneity in all directions showed more of the spatial difference at the macro and meso scale in Hubei province, and the spatial difference at the micro scale became less obvious. Observing the fractal dimension of each direction, aside from the fractal dimension of the southeast to northwest direction increased the fractal dimension of other directions decreased, which showed that the county economic development of Hubei province was relatively balanced in

TABLE 2 Fractal of variogram on spatial pattern of county economy in Hubei province.

year	All directions		South—north		Northeast—southwest		East—west		Southeast—northeast	
	D	R^2	D	R^2	D	R^2	D	R^2	D	R^2
2005	1.895	0.528	1.897	0.251	1.913	0.345	1.943	0.187	1.753	0.599
2010	1.918	0.297	1.956	0.056	1.963	0.049	1.925	0.184	1.763	0.535
2015	1.905	0.432	1.995	0.532	1.848	0.569	1.889	0.275	1.898	0.44
2020	1.878	0.617	1.873	0.408	1.685	0.84	1.902	0.372	1.917	0.232

Note: D stands for fractal dimension; R^2 stands for coefficient of determination.

the southeast to northwest direction, and the spatial difference in economic development in other directions is increasing. The main reason is that after 2010, with the rapid social and economic development of Hubei province, the construction of Wuhan metropolitan area, Western Hubei eco-cultural tourism circle, and Hanjiang ecological economic belt has been proposed successively. The absolute difference in the total amount of county economy is expanding, which shows that the fractal dimensions in all directions are decreasing. With the opening of Wuhan–Xiangyang–Shiyan high-speed railway and the development of economy around the axis, the economic spatial difference from southeast to northwest is relatively balanced, and the economic development is of good homogeneity.

4) The kriging interpolation 3D fitting chart was depicted based on per capita GDP of each county in Hubei province. By observing the graph (Figure 9), we can see that there is a certain regularity in the development pattern of the county economy in Hubei province, and the distribution pattern has structural characteristics. The 4 years showed obvious correlation in all directions. There were obvious multi peak slope structures in the 4 years, among which the slope in 2010 was relatively steep, and the slope in the other 3 years was relatively gentle. Over time, the number of peaks tends to be stable, and the height gradually increased. The spatial structure of slope was consistent with the spatial structure of urban and economic development of "One Main and Two Deputy (One is Wuhan, the provincial central city; two are Yichang and Xiangyang, the provincial sub-central cities)" in Hubei Province, and the vast county periphery presents a plain structure. The results showed that the county economic pattern of Hubei province was relatively stable, and the peak value was mainly concentrated in three regions: the east, including Wuhan City and its surrounding counties; the northwest, including Xiangyang, Jingmen, and Shiyan; and the southwest, including Yichang city and its surrounding counties. The peak growth rate of Yichang city and its surrounding counties in the southwest was relatively high.

In general, the spatial structural difference in county economic development in Hubei province is very significant.

4 Discussion

4.1 Driving forces of economic spatial pattern evolution

The driving factors of county economy development refer to the internal and external factors and their mutual relations that promote the occurrence and development of county economy. The county economic system is a complex and open regional system. Its formation, development, and evolution involve a variety of influencing factors, and different influencing factors have different modes of action. This article uses statistical analysis methods to analyze the spatial pattern and mechanism of county economy in Hubei province. In terms of driving factors, this article mainly starts from systematic theory, combined with the aforementioned empirical research, and finds that the driving forces of county economy evolution is mainly the result of four important factors: historical and cultural background, economic development, traffic location, and policy system, which together form a relatively complete driving system (Figure 10).

Historical and cultural background is a basic driving force. The evolution of spatial pattern of economic development has its inherent historical basis. Since the economic reform starting in 1978, the regional economic pattern of Hubei is centered on Wuhan, the hot spots are mainly distributed stably in Wuhan and its neighboring counties. With the approval of the State Council, Wuhan also implements separate plans and is endowed with provincial economic management authority. The starting point of development of eastern Hubei represented by Wuhan is obviously higher than that of central and western Hubei, which makes eastern Hubei rely on good material foundation and achieve rapid economic development, while central and western Hubei develop relatively slow due to their own basic

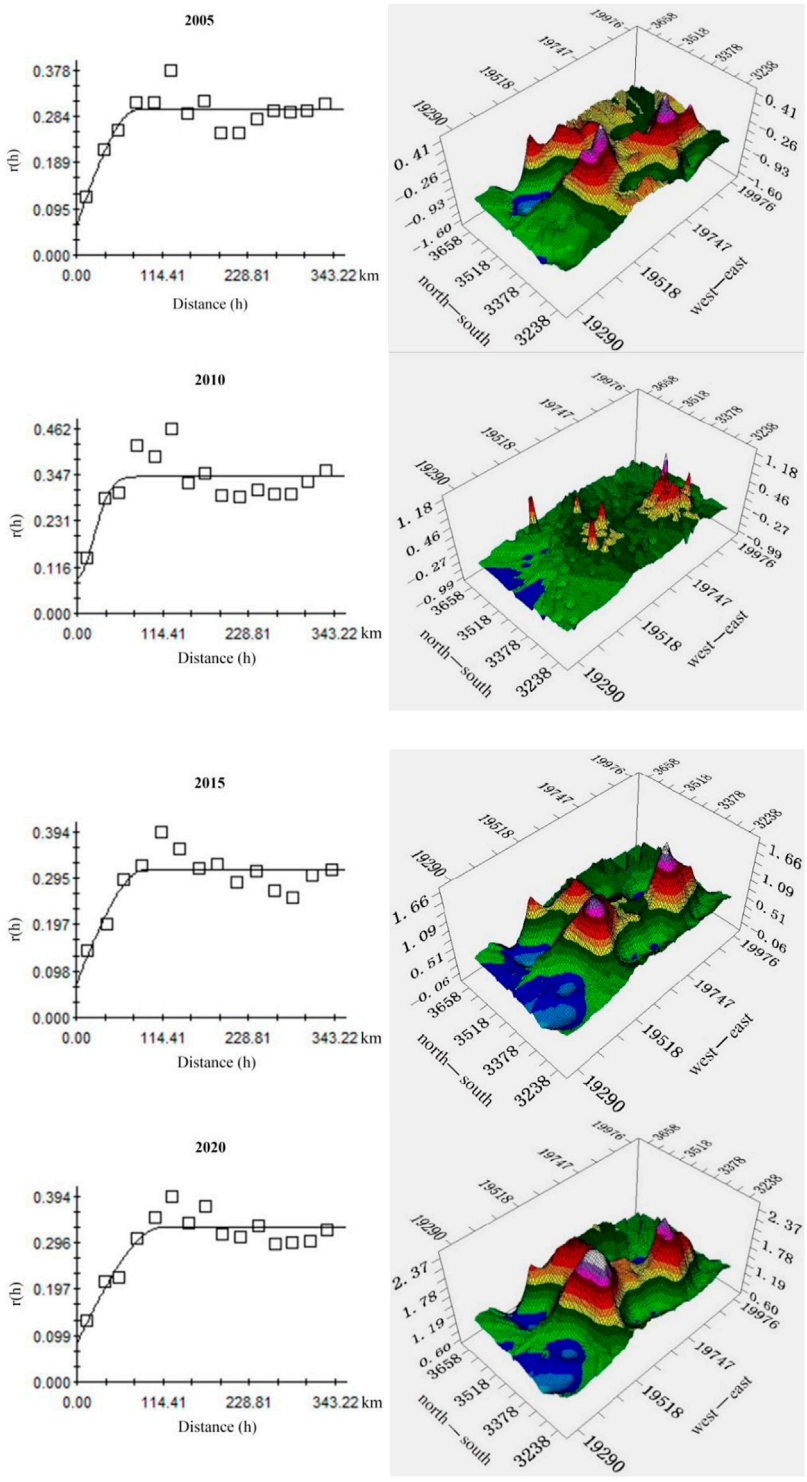


FIGURE 9
Evolvement of variogram on spatial pattern of county economy in Hubei province (left: covariance fitting chart and right: kriging interpolation 3D fitting chart).

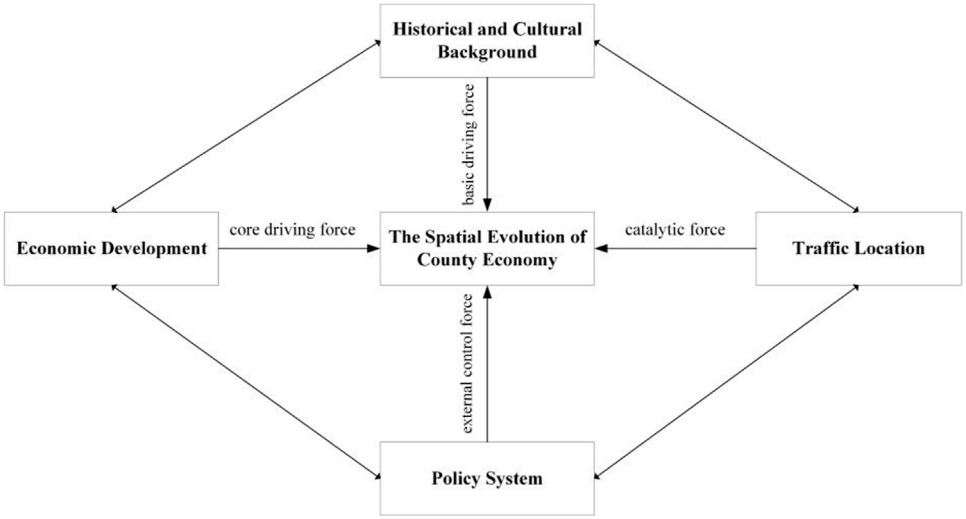


FIGURE 10
Driving forces of the spatial evolution of the county economy in Hubei province.

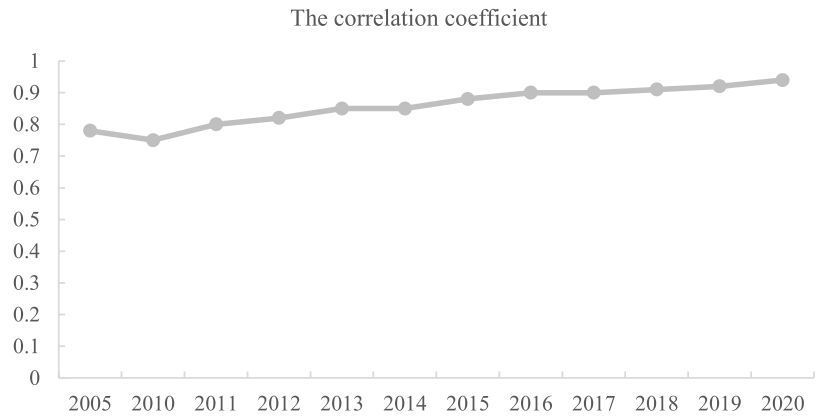


FIGURE 11
Correlation coefficient between per capita GDP and the urbanization level in Hubei province, 2005–2020.

problems. A cold spot for economic development has been formed for a long time.

Economic development is a core driving force. Economic development is the most important factor affecting the evolution of economic pattern in Hubei. The advantages and disadvantages of the economic level will lead to the difference of development order and degree, as well as the difference of other economic growth factors. The higher the level of regional economic development is, the more employment opportunities can be provided and the more population can be accommodated, thus promoting the expansion of production space and living space and the continuous evolution of the urban and rural settlement system. Selecting urbanization level indicators to represent economic development and by analyzing the correlation between per capita GDP and urbanization level in each county, it is found that there is a significant correlation between the two at the level of 5%, and the overall trend is increasing year by year (Figure 11). The changes of population scale and structure directly promote the transformation of county economy and the reconstruction of settlement space and directly affect the different spatial changes of county (district) area, town area, and village land, thus promoting the

continuous evolution of the county economic spatial pattern. In addition, under market economy conditions, the industries in the southeast coastal areas are transferred to the central and western regions, and Hubei is at the node of transfer. Guided by the concept of unbalanced development, in accordance with the principle of combining the regional tilt with industrial tilt, regional production specialization with comprehensive development, together with "opening up and developing the Hubei Yangtze River Economic Belt." The economic pattern extends to northwest and southwest Hubei, forming the "big triangle" regional economic layout of Wuhan–Huangshi, Yichang–Jingzhou, and Xiangyang–Shiyan.

Traffic location is a catalytic force. The transportation network constitutes an important channel for the flow of elements and resource exchange between counties, and it is also an important guarantee for the external connection of urban–rural integration. A developed transportation network means more convenient logistics, flow of people, and more development opportunities. The land rent difference in the county area is gradually changed to be determined by the transportation location. As a result, the county area usually expands along the two sides of the transportation line or around the transportation hub. The transportation network and its convenience have an important catalytic effect on the county settlement system and spatial evolution. The eastern Hubei represented by Wuhan is located in the Jianghan Plain, with a dense transportation network and frequent flow of various elements, making it always the core driving area of economic development. Except for Xiangyang and Yichang in central and western Hubei, the traffic conditions are relatively backward, the location is not good, and the economic development is in the cold spot.

Policy system is an external control force. Regional development positively relates to effective local state initiatives and favorable local conditions. The macro policy of the national and provincial governments is a special instrumental "resource," which itself contains a specific right of preemption, especially the direct input of the government shows the orientation of its administrative behavior. It reflects the decision maker's intention or goal of regional economic development, that is, balanced or unbalanced regional development. The spatiotemporal pattern evolution of economy is related closely to a series of Hubei regional development policies. From 2005 to 2015, Hubei province put forward the development strategy of "Wuhan metropolitan area" and "One Main and Two Deputy," the spatial agglomeration trend is relatively stable. Although there is a switch between hot spots and cold spots in the economy, the switch has an obvious "trajectory." The eastern Hubei region, represented by Wuhan, is attributable to their incomparable policy privileges and superior socioeconomic development advantages to other regions, thus their economic development reached the top level. Yichang and Xiangyang took special advantage of their optimal physical conditions, location, and socioeconomic factors, becoming the key areas of industrial construction in Hubei. In 2015, the GDP of two cities was 164.732 billion yuan and 163.830 billion yuan, respectively, accounting for 10.8% of the GDP of Hubei province, respectively, which is the intermediate bridge to promote the regional cascade development. Together with the adoption of trickle-down growth strategy by the government, the county economic differences in Hubei province have eased; From 2015 to 2020, under the guidance of the concept of balancing urban and rural development, Hubei province put forward the construction of the Yangtze River Economic Belt and the Han River ecological economic belt, as well as the construction of Jingzhou "waist-strengthening project," which have significant spatial spillover effects. The hot spots and sub-hot spots of economic growth move to the central Hubei and concentrated in contiguous distribution, further narrowing the county economic differences in Hubei province.

4.2 Optimization of the economy spatial pattern guided by the "point-axis" development model

The point-axis theory is an extension of the growth pole theory. It not only attaches importance to the role of "point" of growth pole but also emphasizes the role of "axis" to drive regional economic growth. With the formation of growth poles in different regions, the diffusion of multiple growth poles will form the aggregation of elements facing each other between growth poles, thus generating relatively dense element flow on the "axis" between growth poles, forming a relatively developed axis in the region relative to the growth poles. This axis is generally developed along the trunk lines of traffic, once it is formed, it is more favorable to develop the development potential of areas along the trunk lines of traffic. The axis area with its location advantage, adsorption enterprises, and economic sectors form secondary development advantage, which will eventually become a wider growth zone.

This article shows that the "point-axis" development model may be an effective form of economic spatial organization in Hubei province in the future. The spatial continuity and self-organization of the county economic development are strengthened. The aforementioned analysis showed that Hubei province has formed Wuhan, Yichang, and Xiangyang three economic growth points. Several economic growth points and key development axes can be identified in eastern, central, southwest, and northwest Hubei, so as to gradually spread and develop, and promote coordinated regional development in an all-round way. Specifically, east Hubei to southwest Hubei can form a southern horizontal development axis along the Yangtze River, Hu-Han-Rong expressway, and Hu-Han-Rong expressway and national highway 318, the area has a high density of towns and cities with close links to transportation and

industry. The northern horizontal development axis from eastern Hubei to northwest Hubei can be formed along Han-Shi Expressway, Handan–XiangYu railway, and 316 national road. The central and southern part of Hubei province is centered on Jingmen and Jingzhou, with the Jianghan Plain as its vast hinterland and highway 318 as its core development axis. In addition, Xiangyang and Shiyan are the centers in the northwest of Hubei, which can take the Handan line and the Han River as the development axis. The southwest of Hubei province is sparsely populated, with uneven distribution of traffic lines and traffic jams. We should take Yichang as the center and the Yangtze River as the development axis, expand westward along the Qingjiang River and speed up traffic construction.

The A-shaped point-axis structure is a reliable transportation framework for regional development in Hubei province. It also has a significant spatial effect surface, which is conducive to giving full play to the role of central cities at all levels and realizing the optimal spatial combination between the production layout and linear infrastructure. It is beneficial to realize the specialization and cooperation between regions and urban and rural areas, and form an organic regional economic network, so as to make the region get the best development.

4.3 Limitations of the study

This article only takes per capita GDP as the measurement index, which cannot fully reflect the comprehensive economic strength of the county, and it only takes the county as the research unit, rather than smaller scales, such as township economic differences. In addition, the economic development has the spatial spillover effect, and the county economy will also be affected by the surrounding areas. Therefore, in future research, we can use the multi index evaluation system of economic differences, extend the research time range, try to explore the spatial differentiation of regional economic differences on a smaller scale, and further explore the effect intensity and regional differences in the influencing factors of economic development in Hubei province and its different regions.

5 Conclusion and policy implications

Taking Hubei province as an example and based on spatial autocorrelation and a spatial variogram, this article revealed the spatial pattern evolution and mechanism of the county economy in Hubei province. The conclusions are as follows:

1) The absolute economic differences among the counties in Hubei province expanded from 2005 to 2020, the relative differences tended to narrow, and the number of counties larger than the average per capita GDP of the province gradually increased. On the whole, this reflects the trend of narrowing economic differences among counties. There has been significant spatial agglomeration in the development of the county economy, and the degree of agglomeration has increased. From the perspective of local spatial autocorrelation analysis, HH-type counties are mainly concentrated in the Wuhan urban area and surrounding areas, LL-type counties are mainly distributed in most counties in western Hubei, and the types are scattered in northeastern and south Hubei. The distribution of HL- and LH-type in counties is very small, and there is no obvious law.
2) The county economy in Hubei province presents a circular structure as a whole. From the Wuhan urban area to the edge of the province, a hot area, sub-hot area, sub-cold area, and cold area have successively formed. The Yichang urban area and its surrounding counties form another hot area of economic development in the province, and the sub-hot counties in the middle of the province have increased. The cold area concentrates in the edge counties of western, northeastern, and southeastern of Hubei. During the study period, the sub-hot area and the sub-cold area changed greatly, and they transformed each other.
3) The spatial self-organization and spatial autocorrelation of the county economic pattern in Hubei province have strengthened; the structural differentiation trend caused by the spatial autocorrelation has also strengthened, and the economic spatial differentiation is obvious. The southeast to northwest direction has the largest quantile, the county economy is relatively balanced in the southeast to northwest direction, and the spatial differences in other directions have increased. In addition, having multiple peaks is a feature of the evolution of spatial patterns of the county economy. Over time, the number of peaks has tended to be stable.
4) Our results have significant policy implications for county economic development in Hubei province. Through a spatiotemporal perspective, we find that spatial agglomeration is obvious, which indicates regional inequality will be continued. Therefore, we proposed some suggested countermeasures in this article. ① Build the economic development axis of “Wuhan–Yichang–Xiangyang” in central Hubei. On the axis of economic development, the first is to support Wuhan to speed up the construction of a national central city, optimize and adjust the administrative divisions in a timely manner, incorporate the corresponding surrounding cities (counties) into the territory of Wuhan, and expand the “waist circumference” of Wuhan; the second is to lead the construction of the “Xiang–Shi–Sui” urban belt and promote regional cooperation in the Han River basin; the third is to lead the construction of the “Yi–Jing–Jing” urban agglomeration and promote the ecological and economic cooperation of the Three Gorges, and forming an “A-shaped point-axis” spatial structure of economic development. ② The

local county-level government in the western region should take more measures to adjust the industrial structure and improve farmers' income such as enhancing infrastructure, developing ecological tourism, and selling agricultural products through e-commerce. Enhancing regional cooperation and coordinated development across counties is helpful to reinforce the spatial spillover effect between counties and then narrow the localized regional inequality.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

LJ: methodology, writing—original draft, writing—review, and editing; JC: software; CZ: validation, data curation, and project administration; YT: formal analysis, investigation, resources, and visualization; and GW and JL: conceptualization and supervision.

Funding

This research was funded by the National Natural Science Foundation of China (Grant Numbers: 41871176, 42001185, and 42101209) and the Humanities and Social Science Foundation of Ministry of Education of China (Grant Number: 20YJCZH147).

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

Anselin, L. (2013). *Spatial econometrics: Methods and models*, 4. Berlin, Germany: Springer Science & Business Media.

Atkinson, A. (1970). On the measurement of inequality. *J. Econ. Theory* 2, 244–263. doi:10.1016/0022-0531(70)90039-6

Cai, A. N., Zhuang, L., and Liang, J. S. (2011). Analyse of regional economic inequality in Jiangsu province by decomposition of Gini. *Econ. Geogr.* 31, 1995–2000. doi:10.15957/j.cnki.jjdl.2011.12.009

Candelaria, C., Daly, M., and Hale, G. (2009). Beyond kuznets: Persistent regional inequality in China. *Fed. Reserve Bank San Franc. Work. Pap. Ser.* 7, 1.000–28.000. doi:10.24148/wp2009-07

Chen, L., Zhu, X. G., and Li, X. H. (2016). Disparities of country economy at Yunnan province in China base on the perspective of industrial structure. *Sci. Geogr. Sin.* 36, 384–391. doi:10.13249/j.cnki.sgs.2016.03.009

Chen, P. Y., and Zhu, X. G. (2012). Regional inequalities in China at different scales. *Acta Geogr. Sin.* 67, 1085–1097. doi:10.11821/xb201208007

Czako, K., Fekete, D., and Poreisz, V. (2014). Economic differences of countries by the river danube. *Procedia Econ. Finance* 9, 163–175. doi:10.1016/s2212-5671(14)00017-3

Dou, Y., Luo, X., Dong, L., Wu, C. T., Liang, H. W., Ren, J. Z., et al. (2016). An empirical study on transit-oriented low-carbon urban land use planning: Exploratory Spatial Data Analysis (ESDA) on Shanghai, China. *Habitat Int.* 53, 379–389. doi:10.1016/j.habitatint.2015.12.005

Esteban, J., Gradín, C., and Ray, D. (2007). An extension of a measure of polarization, with an application to the income distribution of five OECD countries. *J. Econ. Inequal.* 5, 1–19. doi:10.1007/s10888-006-9032-x

Esteban, J., and Ray, D. (1994). On the measurement of polarization. *Econometrica* 62, 819. doi:10.2307/2951734

Ezcurra, R., Gil, C., Pascual, P., and Rapun, M. (2006). Inequality, polarisation and regional mobility in the European Union. *Urban Stud.* 42, 1057–1076. doi:10.1080/00420980500120824

Fang, Y. L., Huang, Z. F., Chen, W. D., and Xie, H. W. (2013a). Spatial evolution of county economy in Anhui province during 2001-2010. *Prog. Geogr.* 32, 831–839. doi:10.11820/dlkxjz.2013.05.014

Fang, Y. L., Huang, Z. F., Tu, W., and Wu, L. M. (2013b). Research of spatial differences of county econ-omy in Anhui based on Geostatistical analysis. *Econ. Geogr.* 33, 33–38. doi:10.15957/j.cnki.jjdl.2013.02.005

Fedorov, L. (2002). Regional inequality and regional polarization in Russia, 1990-1999. *World Dev.* 30, 443–456. doi:10.1016/s0305-750x(01)00124-3

Feng, X. H., Zhong, Y. X., Li, J. X., and Huang, H. (2015). Spatial distribution of the comprehensive level of county-urbanization and its driving factors in urban agglomeration in the middle reaches of Yangtze river. *Resour. Environ. Yangtze Basin* 24, 899–908. doi:10.11870/cjlyzyyhj201506001

Friedman, J. R. (1966). *Regional development policy—a case study of Venezuela*. Cambridge: The MIT Press.

Fu, B., J. (2014). The integrated studies of geography: Coupling of patterns and processes. *Acta Geogr. Sin.* 69, 1052–1059. doi:10.11821/dlxb201408002

Fujita, M., Krugman, P., and Venables, A. J. (1999). *The spatial economy*. Cambridge: MIT press.

Gasparini, L., Horenstein, M., and Olivieri, S. (2006). *Economic polarisation in Latin America and the caribbean: What do household surveys tell us?* La Plata: Universidad Nacional de La Plata.

Goodchild, M. F. (2009). What problem? Spatial autocorrelation and geographic information science. *Geogr. Anal.* 41, 411–417. doi:10.1111/j.1538-4632.2009.00769.x

He, S., Fang, C., and Zhang, W. (2017). A geospatial analysis of multi-scalar regional inequality in China and in metropolitan regions. *Appl. Geogr.* 88, 199–212. doi:10.1016/j.apgeog.2017.08.017

He, S. W., Felix, H. L., and Li, G. D. (2019). A spatiotemporal analysis of county economy and the multi-mechanism process of regional inequality in rural China. *Appl. Geogr.* 111, 102073. doi:10.1016/j.apgeog.2019.102073

Hu, Y., and Zhang, J. X. (2010). Economic space evolution of Yangtze river delta city group based on county scale. *Econ. Geogr.* 30, 1112–1117. doi:10.15957/j.cnki.jjdl.2010.07.009

Jin, C., and Lu, Y. Q. (2009). Evolvement of spatial pattern of economy in Jiangsu province at county level. *Acta Geogr. Sin.* 64, 713–724. doi:10.11821/xb200906008

Jin, C., Xu, J., and Lu, Y. Q. (2007). Study on disparity of tourism scale for cities in Yangtze—Delta and construction for the system of rank—Size distribution. *Econ. Geogr.* 27, 676–680. doi:10.3969/j.issn.1000-8462.2007.04.035

Kanbur, R., and Zhang, X. (2005). Fifty years of regional inequality in China: A journey through central planning, reform and openness. *Rev. Dev. Econ.* 9, 87–106. doi:10.1111/j.1467-9361.2005.00265.x

Ke, W. Q., and Lu, Y. Q. (2011). Evolvement of spatial pattern of economy at county level in Fujian Province. *Econ. Geogr.* 31, 1081–1087. doi:10.15957/j.cnki.jjdl.2011.07.005

Ke, W. Q., Lu, Y. Q., Yu, Z. Y., Chen, W., and Wang, H. (2013a). Study on economic factor relation of Jiangsu counties and evolution process based on quantile regression. *Geogr. Geo-Information Sci.* 29, 72–77.

Ke, W. Q., Lu, Y. Q., Yu, Z. Y., Wang, H., and Ma, Y. Y. (2013b). Spatial evolution of economic development at county level in Jiangsu based on multivariate drive. *Acta Geogr. Sin.* 68, 802–812.

Krugman, P. (1991). Increasing returns and economic geography. *J. Political Econ.* 99, 483–499. doi:10.1086/261763

Lewis, P. H. (1996). *Tomorrow by design: A regional design process for sustainability*. New York: Wiley, 36–37.

Li, F., and Li, G. D. (2018). Agglomeration and spatial spillover effects of regional economic growth in China. *Sustainability* 10, 4695. doi:10.3390/su10124695

Li, G. D., and Fang, C. L. (2014). Quantitative measure and influencing mechanism of land intensive use in China at the county level. *Acta Geogr. Sin.* 69, 1739–1752. doi:10.11821/dlxb201412001

Li, R., Cao, M. M., Hu, S., Liu, W., and Ji, Q. J. (2014). The coordinated development about eco-environment and economy: A case study of yulin city of shanxi. *Hum. Geogr.* 29, 101–107. doi:10.13959/j.issn.1003-2398.2014.05.017

Li, X. J., and Qiao, J. Q. (2001). County level economic disparities of China in the 1990s. *Acta Geogr. Sin.* 68, 136–145. doi:10.3321/j.issn:0375-5444.2001.02.002

Liao, F. H. F., and Wei, Y. D. (2012). Dynamics, space, and regional inequality in provincial China: A case study of guangdong province. *Appl. Geogr.* 35, 71–83. doi:10.1016/j.apgeog.2012.05.003

Liao, F. H., and Wei, Y. D. (2015). Space, scale, and regional inequality in provincial China: A spatial filtering approach. *Appl. Geogr.* 61, 94–104. doi:10.1016/j.apgeog.2014.12.022

Liu, C. M., and Zeng, J. X. (2011). The calculating method about the comprehensive transport accessibility and its correlation with economic development at county level: The statistical analysis of 79 counties in Hubei province. *Geogr. Res.* 30, 2210–2220.

Liu, J. B., Yao, Y. J., and Zhang, Y. F. (2007). The analysis of multi-scale spatial patterns of vegetation based on fractal geometry. *Res. Soil Water Conservation* 14, 155–158. doi:10.3969/j.issn.1005-3409.2007.01.049

Liu, X. Z., Zhang, A. D., and Li, J. Z. (2009). *Geography teaching methods*. Beijing: Science Press.

Liu, Y. S., and Yang, R. (2012). The spatial characteristics and formation mechanism of the county urbanization in China. *Acta Geogr. Sin.* 67, 1011–1020. doi:10.11821/xb201208001

Lu, D. D., Liu, Y., and Fan, J. (1999). The regional policy effects and regional development states in China. *Acta Geogr. Sin.* 66, 496–508. doi:10.11821/xb199906003

Lu, F., and Xu, J. H. (2005). Chinese regional economic inequalities based on the two-stage nested theil decomposition method. *Sci. Geogr. Sin.* 25, 401–407. doi:10.3969/j.issn.1000-0690.2005.04.003

Ma, G. X., Xu, Y., and Tian, Y. J. (2007). Spatial analysis of economic growth convergence mechanism in Beijing-Tianjin-Hebei Metropolitan Region. *Geogr. Res.* 26, 590–598. doi:10.3321/j.issn:1000-0585.2007.03.020

Massey, D. (1985). "New directions in space," in *Social relations and spatial structures*. Editors D. Gregory and J. Urry (London: Macmillan), 9–19.

Ortega, E., Monzon, A., and Lopez, E. (2018). The influence of spatial data allocation procedures on accessib-ility results: The case of high-speed rail networks. *Appl. Geogr.* 94, 241–250. doi:10.1016/j.apgeog.2018.03.020

Perroux, F. (1950). Economic space: Theory and applications. *Q. J. Econ.* 64, 89. doi:10.2307/1881960

Rey, S. J., and Le Gallo, J. (2009). Spatial analysis of economic convergence. *Palgrave Handb. Econ.* 2, 1251–1290. doi:10.1057/9780230244405_27

Rey, S. J., and Sastre Gutierrez, M. L. (2015). Comparative spatial inequality dynamics: The case of Mexico and the United States. *Appl. Geogr.* 61, 70–80. doi:10.1016/j.apgeog.2015.01.012

Rey, S. J. (2016). Space–time patterns of rank concordance: Local indicators of mobility association with application to spatial income inequality dynamics. *Ann. Am. Assoc. Geogr.* 106, 788–803. doi:10.1080/24694452.2016.1151336

Rozelle, S., Huang, J., and Zhang, L. (2002). Emerging markets, evolving institutions, and the new opportunities for growth in China's rural economy. *China Econ. Rev.* 13, 345–353. doi:10.1016/s1043-951x(02)00091-3

Sun, P. J., Xiu, C. L., and Dong, C. (2013). Quantitative analysis of economic spatial polarization and driving factors in the northeast of China. *Hum. Geogr.* 28, 87–93. doi:10.13959/j.issn.1003-2398.2013.01.015

Sun, S. S., Zhu, C. G., and Li, Z. J. (2009). Economic development disparities of huaihai economic zone. *Acta Geogr. Sin.* 64, 924–934.

Tian, X. Y., and Zhao, Y. T. (2013). Economic spatial disparity in Guizhou Province based on county. *Resour. Environ. Yangtze Basin* 22, 27–31.

Van Oort, F. G., Burger, M. J., Knoben, J., and Raspe, O. (2012). Multilevel approaches and the firm-agglomeration ambiguity in economic growth studies. *J. Econ. Surv.* 26, 468–491. doi:10.1111/j.1467-6419.2012.00723.x

Wan, G. H., and Zhou, Z. Y. (2005). Income inequality in rural China: Regression based decomposition using household data. *Rev. Dev. Econ.* 9, 107–120. doi:10.1111/j.1467-9361.2005.00266.x

Wang, F., and Gao, X. L. (2014). Spatial pattern evolvement of the economy in inner Mongolia at the county level. *Sci. Geogr. Sin.* 34, 818–824. doi:10.13249/j.cnki.sgs.2014.07.012

Ward, P. S. (2016). Transient poverty, poverty dynamics, and vulnerability to poverty: An empirical analysis using a balanced panel from rural China. *World Dev.* 78, 541–553. doi:10.1016/j.worlddev.2015.10.022

Wei, Y. D., and Kim, S. (2002). Widening inter-county inequality in Jiangsu province, China, 1950–95. *J. Dev. Stud.* 38, 142–164. doi:10.1080/00220380412331322621

Wei, Y. D. (2015). Spatiality of regional inequality. *Appl. Geogr.* 61, 1–10. doi:10.1016/j.apgeog.2015.03.013

World Bank (2009). *World development report 2009: Reshaping economic geography*. Washington, DC: World Bank.

Wu, D. T. (2001b). Systematic analysis on regional difference between east, middle and west of China. *Areal Res. Dev.* 20, 10–15. doi:10.3969/j.issn.1003-2363.2001.02.003

Wu, D. Y. (2001a). A study on north-south differences in economic growth. *Geogr. Res.* 20, 238–246. doi:10.3321/j.issn:1000-0585.2001.02.015

Wu, J. M., and Ding, J. H. (2011). A research on the spatial differences of the economic performance on the county level and the reasons: Taking Wuji county, Hebei Province for instance. *Areal Res. Dev.* 30, 33–37+44. doi:10.3969/j.issn.1003-2363.2011.02.007

Xiong, W., Xu, Y. L., and Wang, Y. G. (2011). Temporal-spatial change of the county-level economic disparities in Jiangsu Province. *Prog. Geogr.* 30, 224–230. doi:10.11820/dlkxjz.2011.02.012

Xu, J. H., Lu, F., Su, F. L., and Lu, Y. (2005). Spatial and temporal scale analysis on the regional economic dispariti-es in China. *Geogr. Res.* 24, 57–68. doi:10.3321/j.issn:1000-0585.2005.01.007

Xu, M., and Liu, C. L. (2012). Performance evaluation and correlation trend analysis of urban-rural integrated development in China. *J. Nat. Resour.* 27, 734–746. doi:10.11849/zrzyxb.2012.05.003

Xu, Y. N., Li, E. L., Xu, Y. H., and Wei, S. H. (2015). Spatial pattern change of the county scale economic development in Central Plains Economic Region. *Econ. Geogr.* 35, 33–39. doi:10.15957/j.cnki.jjdl.2015.04.005

Ye, X., and Wei, Y. H. D. (2005). Geospatial analysis of regional development in China: The case of zhejiang province and the wenzhou model. *Eurasian Geogr. Econ.* 46, 445–464. doi:10.2747/1538-7216.46.6.445

Yu, D. L., and Wei, Y. H. D. (2003). Analyzing regional inequality in post-mao China in a GIS environment. *Eurasian Geogr. Econ.* 55, 514–534. doi:10.2747/1538-7216.44.7.514

Zhang, J., and Deng, W. (2016). Multiscale spatio-temporal dynamics of economic development in an interprovincial boundary Region: Junction area of Tibetan plateau, hengduan mountain, yungui plateau and sichuan basin, southwestern China case. *Sustainability* 8, 215. doi:10.3390/su8030215

Zhang, J. Z., Liang, J. S., and Zhu, Y. X. (2010). The characteristics, reasons, and measures of county economic division of Xinjiang. *Sci. Geogr. Sin.* 30, 536–543. doi:10.13249/j.cnki.sgs.2010.04.024

Zhang, X., and Kanbur, R. (2005). Spatial inequality in education and health care in China. *China Econ. Rev.* 16, 189–204. doi:10.1016/j.chieco.2005.02.002

Zhang, X., and Kanbur, R. (2001). What difference do polarisation measures make? An application to China. *J. Dev. Stud.* 37, 85–98. doi:10.1080/00220380412331321981

Zhao, Y. Z., and Dong, P. (2012). Analysis on the characteristics of county economic disparities and its causes in Jiangxi Province. *Hum. Geogr.* 27, 87–91. doi:10.3969/j.issn.1003-2398.2012.01.017

Zhong, Y. X., and Lu, Y. Q. (2010). County level economic disparities in the red-triangle economic zone. *Econ. Geogr.* 30, 568–573. doi:10.15957/j.cnki.jjdl.2010.04.012

Zhou, J. W., and Zhang, L. (2011). Spatial scale effects of the regional economic disparities in the central China area. *Geogr. Geo-Information Sci.* 27, 49–52.

Zhu, S. J., and Wu, J. M. (2015). Fluctuation and structure characters of county economic growth quality: A case study of hebei province. *Geogr. Geo-Information Sci.* 31, 81–87.

frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 08 August 2022
DOI 10.3389/fenvs.2022.929707

Check for updates

OPEN ACCESS

EDITED BY
Shaolong Sun,
Xi'an Jiaotong University, China

REVIEWED BY
Siroos Shojaei,
University of Sistan and Baluchestan, Iran
Swapan Talukdar,
Jamia Millia Islamia, India

*CORRESPONDENCE
Nadhir Alansari,
nadhir.alansari@ltu.se

SPECIALTY SECTION
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

RECEIVED 27 April 2022
ACCEPTED 06 July 2022
PUBLISHED 08 August 2022

CITATION
Maroufpoor S, Sammen SS, Alansari N, Abba SI, Malik A, Shahid S, Mokhtar A and Maroufpoor E (2022), A novel hybridized neuro-fuzzy model with an optimal input combination for dissolved oxygen estimation.
Front. Environ. Sci. 10:929707.
doi: 10.3389/fenvs.2022.929707

COPYRIGHT
© 2022 Maroufpoor, Sammen, Alansari, Abba, Malik, Shahid, Mokhtar and Maroufpoor. 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 novel hybridized neuro-fuzzy model with an optimal input combination for dissolved oxygen estimation

Saman Maroufpoor[1], Saad Sh. Sammen[2], Nadhir Alansari[3]*, S.I. Abba[4,5], Anurag Malik[6], Shamsuddin Shahid[7], Ali Mokhtar[8] and Eisa Maroufpoor[9]

[1]Department of Irrigation and Reclamation Engineering, University of Tehran, Tehran, Iran, [2]Department of Civil Engineering, College of Engineering, University of Diyala, Baqubah, Iraq, [3]Department of Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, Lulea, Sweden, [4]Department of Civil Engineering, Faculty of Engineering, Baze University, Abuja, Nigeria, [5]Interdisciplinary Research Centre for Membrane and Water Security, King Fahd University of Petroleum and Minerals Dhahran, Dhahran, Saudi Arabia, [6]Punjab Agricultural University, Regional Research Station, Bathinda, India, [7]School of Civil Engineering, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia, [8]Department of Agricultural Engineering, Faculty of Agriculture, Cairo University, Giza, Egypt, [9]Department of Water Engineering, Faculty of Agriculture, University of Kurdistan, Sanandaj, Iran

Dissolved oxygen (DO) is one of the main prerequisites to protect amphibian biological systems and to support powerful administration choices. This research investigated the applicability of Shannon's entropy theory and correlation in obtaining the combination of the optimum inputs, and then the abstracted input variables were used to develop three novel intelligent hybrid models, namely, NF-GWO (neuro-fuzzy with grey wolf optimizer), NF-SC (subtractive clustering), and NF-FCM (fuzzy c-mean), for estimation of DO concentration. Seven different input combinations of water quality variables, including water temperature (TE), specific conductivity (SC), turbidity (Tu), and pH, were used to develop the prediction models at two stations in California. The performance of proposed models for DO estimation was assessed using statistical metrics and visual interpretation. The results revealed the better performance of NF-GWO for all input combinations than other models where its performance was improved by 24.2–66.2% and 14.9–31.2% in terms of CC (correlation coefficient) and WI (Willmott index) compared to standalone NF for different input combinations. Additionally, the MAE (mean absolute error) and RMSE (root mean absolute error) of the NF model were reduced using the NF-GWO model by 9.9–46.0% and 8.9–47.5%, respectively. Therefore, NF-GWO with all water quality variables as input can be considered the optimal model for predicting DO concentration of the two stations. In contrast, NF-SC performed worst for most of the input combinations. The violin plot of NF-GWO-predicted DO was found most similar to the violin plot of observed data. The dissimilarity with the observed violin was found high for the NF-FCM model. Therefore, this study promotes the hybrid intelligence models to predict DO concentration accurately and resolve complex hydro-environmental problems.

KEYWORDS
neuro-fuzzy, grey wolf optimizer, dissolved oxygen, turbidity, California

1 Introduction

The evaluation of water assets and the administration of water quality and quantity have become a debated issue in hydroecology with population growth and environmental changes. Water pollution rather than water availability is often the main challenge due to its multifaced impact ranging from biodiversity to public health. Therefore, water-quality monitoring is one of the most emphasized topics in water research. Numerous chemical, physical, and biological parameters determine river water quality. All the water quality parameters directly or indirectly affect dissolved oxygen (DO) concentration in water bodies, and thus DO is considered an integrated water-quality indicator (Ahmed and Shah, 2017; Hameed et al., 2017). Therefore, precise forecasts of DO are prerequisites to protect amphibian biological systems and support powerful administration choices (Wen et al., 2013; Elkiran et al., 2019; Nourani et al., 2019).

Traditionally, physically based models are used for DO prediction (Radwan et al., 2003; Wu and Yu, 2021). The models generally use advection and dispersion theories to simulate the biological and chemical processes in water for DO prediction. For example, Radwan et al. (2003) used Mike11 for modeling DO in river water. Wu and Yu (2021) used a modified version of the Streeter–Phelps model coupled with the shallow water equation model to simulate mass transportation and DO distribution. The studies indicated the need for a large amount of data and computational time to predict DO reliability. Radwan et al. (2003) reported that a simplified conceptual model can provide a similar DO prediction with much less resources and time. In addition, statistical models can be used to predict DO with less amount of data and resources. Pham et al. (2020) developed several generalized linear models to predict DO. They showed that statistical models can predict DO with reasonable errors. However, conceptual models simplify real physical processes and fail to provide accurate predictions when river water DO concentration follows a non-linear and complex pattern. The linear statistical models experience similar drawbacks when DO is non-linearly related to its controlling factors (Chen and Liu, 2014; Elkiran et al., 2018).

Artificial Intelligence (AI)-based models have been utilized in recent decades in various hydro-natural investigations (Abba et al., 2017; Yavari et al., 2018; Maroufpoor et al., 2019b; Seyedzadeh et al., 2020; Meidute-Kavaliauskiene et al., 2021). The rapid evolution of AI techniques also helped in accurate DO simulations to resolve complex hydro-environmental problems. Artificial neural networks (ANNs) have helped hydrologists in predicting variations in water quality accurately. The other AI models also showed promising results in predicting water quality and DO (Zaher et al., 2015; Elkiran et al., 2018; Nourani et al., 2018; Pham et al., 2019; Banadkooki et al., 2020; Abba et al., 2021; Pham et al., 2021a; Pham et al., 2021b). For example, Chen and Liu (2014) used neuro-fuzzy (NF), back propagation neural network (BPNN), and multiple linear regression (MLR) approaches to estimate DO in the reservoir. The outcomes indicated that the NF model outperformed BPNN. Xiao et al. (2017) employed BPNN to simulate DO in Beihai, Guangxia aquaculture, using various inputs. The prediction results showed the superiority of BPNN against autoregression (AR), curve fitting (CF), grey model (GM), and SVR models. Elkiran et al. (2018) employed the combinations of NF, feedforward neural network (FFNN), and MLR to predict DO at multiple locations in India using different input combinations. Their results showed a slight prediction increment of NF over FFNN. Other recent DO prediction studies using AI-based models include Antanasijević et al. (2019), Cao et al. (2019), Liu et al. (2019), Kisi et al. (2020), and Rahman et al. (2020). The studies revealed AI-based models as promising tools owing to their capability to handle non-linear systems.

The literature overview revealed no specific AI-based models tend to be incomparable to others due to the anthropogenic nature of complex aquaculture in different geographical locations. According to Abba et al. (2020), Hadi et al. (2019), and Yaseen et al. (2020), the estimation outcomes produced by some computational models were still grieving from the degree of inadequacy, particularly when a highly chaotic hydro-environmental system is employed. Therefore, hydrologists continuously explored better AI models for DO prediction more efficiently. NF integrates neural networks and fuzzy systems to join their advantages for a better solution to complex problems. The capability of NF to learn data patterns using fuzzy rules has made it highly adaptable to different kinds of data and thus superior to many other AI algorithms in solving a wide range of problems from different fields (Atmaca et al., 2001). However, the major drawback of NF is that its performance significantly is susceptible to the selection and optimization of the input variable's fuzzy membership function. The state-of-the-art hybrid AI model displayed promising prediction results over standalone models in different hydrological studies (Maroufpoor et al., 2019a; Pham et al., 2019; Maroufpoor et al., 2020; Mohammadi et al., 2020; Ebtehaj et al., 2021; Malik et al., 2021; Sammen et al., 2021). Therefore, such models may be a suitable alternative to standalone models in DO prediction. The optimizations and chemometric approaches were introduced in several fields of science and engineering, for instance Shojaei et al. (2019), Shojaei and Shojaei, (2019), Pourabadeh et al. (2020), Shojaei et al. (2021), and Yang et al. (2022).



FIGURE 1
Location map of the selected stations.

In order to overcome the inherent limitations established by standalone models. This research aims to predict the DO concentration using three hybrid models, namely, NF-GWO, NF-SC, and NF-FCM, and compare them with the standalone NF model. The entropy method was used to evaluate each input variable's effect and uncertainty on the models' performance to select the best prediction model structure. The entropy theory, developed by Shannon (1948), has been used in a wide range of studies (Singh, 2013a; Singh, 2013b; Ellenburg et al., 2018; Maroufpoor et al., 2020). There is no technical research in which the aforementioned techniques are used for predicting DO concentration to the best of the author's knowledge.

2 Methodology and materials

2.1 Case study and data sets

This research aimed to develop an intelligent hybrid paradigm for predicting DO concentration in river water using three hybrid models, NF-GWO (neuro-fuzzy with grey wolf optimizer), NF-SC (subtractive clustering), and NF-FCM (fuzzy c-mean). The newly proposed models' efficacy was established by comparing their performance with the standalone neuro-fuzzy (NF) model.

In this study, four water quality parameters, namely, specific conductance (SC, μS.cm-1), water temperature (TE,°C), pH of the water, and turbidity (Tu, Formazin Nephelometric Units: FNU) were used as inputs for the prediction of DO. Two stations (i.e., Station-A and Station-B) in California were selected for the case study (Figure 1). Hourly water quality data at these two stations for the period 1 January 2019–31 December 2019 were collected from the United States Geological Survey (USGS). The descriptive statistics of the data are shown in Table 1. Furthermore, the entropy theory, based on statistical measurements introduced by Shannon (1948), was used to evaluate the significant input variables. Seven different input combinations of water quality variables, including water temperature (TE), specific conductivity (SC, turbidity (Tu), and pH, were used to develop the prediction models at two stations in California. Finally, data were randomly divided into two parts: training (70%) and test (30%).

2.2 Input combinations

The entropy theory, based on statistical measurements introduced by Shannon (1948), was used to evaluate the significant input variables. In this theory, "information" indicates the level of stochastic. The entropy is calculated based on the following steps:

The G matrix is introduced as:

$$\begin{bmatrix} G_{11} & G_{12} & \dots & G_{1N} \\ G_{21} & G_{22} & \cdots & G_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ G_{M1} & G_{M2} & \cdots & G_{MN} \end{bmatrix}, \quad (1)$$

where N (i = 1, 2, 3 . . . . . N) and M (j = 1, 2, 3 . . . M) represent the number of samples in each variable and the number of variables, respectively. Eq. 2 applied to normalize the G matrix:

TABLE 1 Statistical description of data sets at study stations.

Station	Variable	Mean	Maximum	Minimum	SD	CV
A	TE (°C)	15.99	23.95	7.80	4.63	3.45
	SC (μS/cm)	168.97	375.50	97.25	46.39	3.64
	pH	7.58	8.30	7.08	0.17	44.99
	Tu (FNU)	20.67	309.25	2.35	32.82	0.63
	DO (mg/L)	8.92	11.10	7.30	0.82	10.82
B	TE (°C)	16.11	23.33	8.33	4.58	3.52
	SC (μS/cm)	2104.22	14700.00	104.50	2774.84	0.76
	pH	7.68	8.30	7.20	0.20	39.28
	Tu (FNU)	22.74	174.75	4.83	19.73	1.15
	DO (mg/L)	9.08	10.85	7.75	0.71	12.86

FNU, formazin nephelometric units, SD, standard deviation, CV, coefficient of variation.

TABLE 2 Selected input combinations for DO prediction at study stations.

Combination						
1	2	3	4	5	6	7
TE	TE	TE	TE	TE	TE	TE
SC	SC	SC	pH	SC	pH	Tu
Tur	pH	Tu	Tu			
pH						

TE, water temperature; SC, specific conductance; Tu, turbidity

$$O_{ij} = \frac{G_{ij} - \left(G_{ij}\right)_{Min}}{\left(G_{ij}\right)_{Max} - \left(G_{ij}\right)_{Min}}, \quad (2)$$

where O_{ij}, $(Y_{ij})_{Max}$, and $(Y_{ij})_{Min}$ are the normalized, maximum, and minimum parameters, respectively. The next step is to calculate the probability of each parameter:

$$K_{ij} = \frac{\left(G_{ij} + 0.0001\right)}{\sum_{i=1}^{M}\left(G_{ij} + 0.0001\right)}, \quad (3)$$

where K_{ij} is the probability of each parameter.

$$E_j = -\frac{1}{ln(M)}\sum_{i=1}^{M} K_{ij}\, ln\left(K_{ij}\right), \quad (4)$$

$$EW_j = \frac{1 - E_j}{\sum_{j=1}^{N}\left(1 - E_j\right)}, \quad (5)$$

where E_j is the information entropy and EW is the entropy weight (relative importance). A variable with a weight close to one indicates more importance.

Seven scenarios of inputs were investigated to assess their influence on DO concentration prediction, as described in Table 2. The input combinations are termed C1 to C7 in Table 2. The comparison of model performance for different input combinations helped find a suitable model based on data availability. Absolute correlation coefficient and entropy

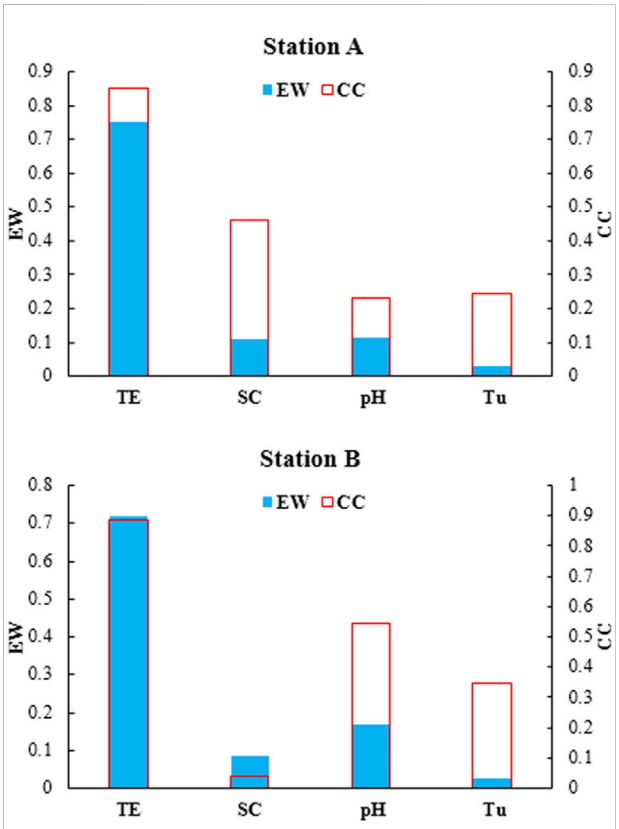


FIGURE 2
Absolute correlation coefficient and entropy weight for the input variables at Station-A and Station-B.

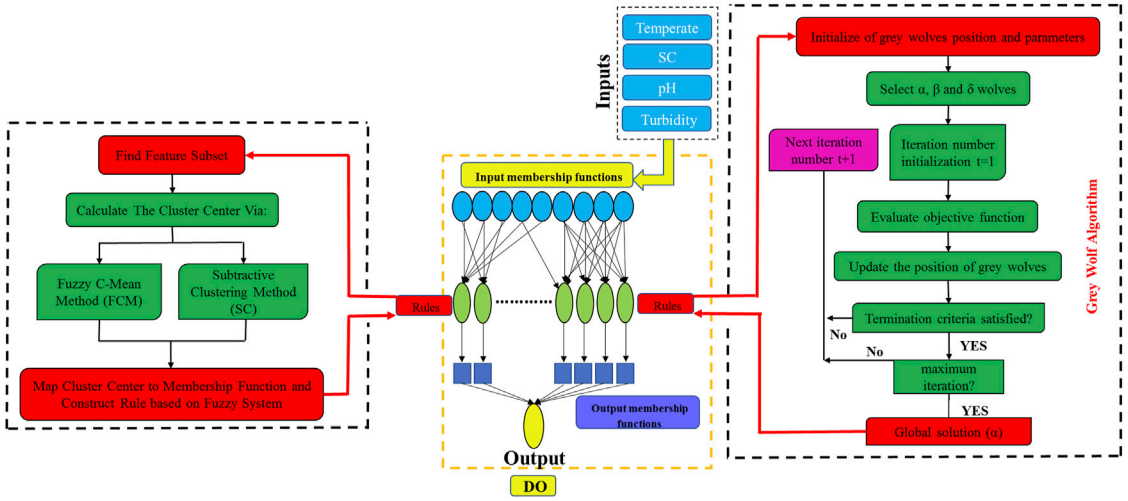


FIGURE 3
Flowchart of applied models.

weight were used to select the input combinations (Figure 2). The first combination includes the four variables (TE, SC, pH, and Tu). Other combinations include fewer inputs to find a parsimonious model for its easy application in the data scarcity regions. For Station-A, the highest correlation coefficient and entropy were found for TE, 0.85 and 0.75, respectively. The SC showed the second-highest correlation of 0.46, but a low entropy of 0.11. The lowest entropy was recorded for Tu (0.03), which showed a correlation coefficient of 0.24. For Station B, the highest correlation coefficient and entropy were also noticed for TE, 0.88 and 0.72, respectively. The pH showed the second-highest correlation, 0.54, and the entropy of 0.17. In this study, data were randomly divided into two parts: training (70%) and test (30%). It should be noted that outlier data were removed by statistical methods, and also raw data were normalized for modeling.

2.3 Applied AI models

2.3.1 Neuro-fuzzy system

The NF model, which combines ANN and fuzzy logic, was designed by Jang (1993). The most important benefit of fuzzy logic is that it can give an intermediate answer to zero-one programming problems. It can be used when there is no complete understanding of the system's physical and fundamental relationships. The NF structure is formed based on the membership functions of input and output, fuzzy rules, and the number of membership functions (Tanaka, 1997). Parameters related to membership functions need to be selected so that they are most consistent with the input–output data. Three algorithms, including SC, FCM, and GWO were used in this research to optimize the rules in the model training process. The flowchart of the proposed methodology is shown in Figure 3.

2.3.2 Subtractive clustering

In the NF model, the number of rules to determine the optimal system increases with the number of membership functions and parameters of the model. Therefore, it is necessary to optimize the NF rules to reduce computational costs. To this end, the subtractive cluster is integrated with the NF system, where the modeling process consists of two stages. First, the fuzzy inference system is determined using the subtractive clustering method. Then, NF is used to adjust the fuzzy inference system and train it based on input–output data.

In subtractive clustering, each cluster's center represents the behavior of a part of the data and represents a rule. Therefore, to determine the optimal structure, cluster information is used to determine the number of basic rules and membership functions. Choosing a small radius increase the number of rules and make the computations more complicated. In this study, the effective radius within the range of zero to one was selected based on the least root mean square error (RMSE).

2.3.3 Fuzzy C-means clustering

FCM is a clustering technique where each point belongs to a cluster with a certain degree. Bezdek (1973) introduced this technique to improve the efficiency of previous clustering methods. In FCM, a certain number of different clusters describe the data clustering in multi-dimensional space. The FCM starts from an initial hypothesis as the centers of the

10.3389/fenvs.2022.929707

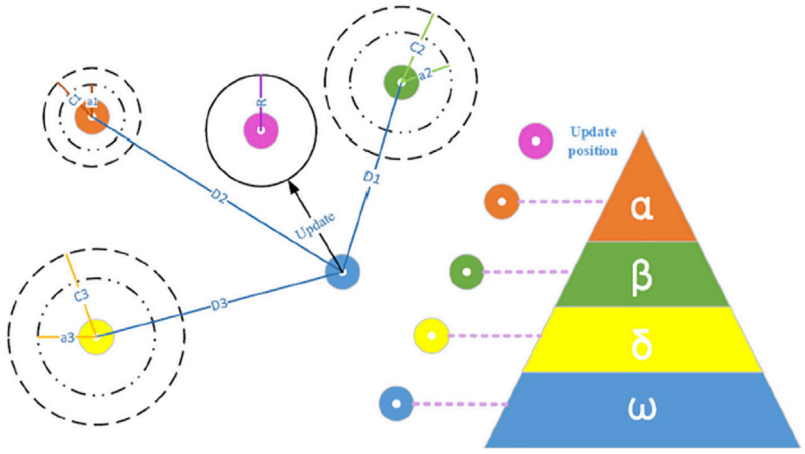


FIGURE 4
Diagram of GWO.

TABLE 3 Performance of models in predicting DO concentration in the test phase (Station-A).

Model	Metrics	C1	C2	C3	C4	C5	C6	C7
NF-GWO	MAE	0.250	0.295	0.475	0.315	0.503	0.332	0.407
	RMSE	0.320	0.365	0.565	0.391	0.625	0.438	0.508
	CC	0.869	0.812	0.590	0.800	0.381	0.791	0.610
	WI	0.908	0.894	0.657	0.872	0.594	0.826	0.763
	Structure	(30,1500)	(40,1500)	(35,1500)	(40,1500)	(20,1000)	(25,1000)	(30,1000)
NF	MAE	0.463	0.481	0.443	0.478	0.612	0.448	0.442
	RMSE	0.609	0.592	0.590	0.597	0.757	0.566	0.558
	CC	0.523	0.508	0.478	0.609	0.149	0.642	0.491
	WI	0.692	0.681	0.671	0.663	0.501	0.639	0.664
	Structure	--	--	--	--	--	--	--
NF-SC	MAE	0.637	0.452	0.449	0.477	0.545	0.456	0.596
	RMSE	0.808	0.575	0.550	0.565	0.671	0.575	0.770
	CC	0.639	0.643	0.501	0.477	-0.046	0.648	-0.074
	WI	0.683	0.662	0.653	0.662	0.304	0.667	0.345
	Radii	0.4	0.5	0.75	0.7	0.6	0.3	0.5
NF-FCM	MAE	0.504	0.348	0.463	0.577	0.521	0.412	0.485
	RMSE	0.634	0.441	0.565	0.765	0.695	0.534	0.625
	CC	0.563	0.752	0.500	0.411	0.618	0.624	0.365
	WI	0.646	0.825	0.658	0.619	0.754	0.789	0.606
	Cluster	3	5	3	3	4	3	2

clusters. Usually, this initial hypothesis is incorrect and does not specify the correct location of the centers. The FCM tries to link each point to one of the clusters by the level of its membership. The centers gradually move to their actual position in the data-set through repeated updating of the cluster centers and membership levels for clusters. These updates are based on decreasing the distance between each point to the center of the clusters. The least RMSE was the basis for selecting the optimal number of clusters in the FCM.

TABLE 4 Performance of models in predicting DO concentration in the test phase (Station-B).

Model	Metrics	C1	C2	C3	C4	C5	C6	C7
NF-GWO	MAE	0.346	0.352	0.386	0.346	0.414	0.401	0.44
	RMSE	0.413	0.416	0.459	0.417	0.475	0.454	0.508
	CC	0.563	0.453	0.215	0.497	0.280	0.359	0.387
	WI	0.713	0.607	0.438	0.644	0.440	0.486	0.53
	Structure	(20,1500)	(30,1500)	(25,1500)	(40,1500)	(20,1000)	(25,1000)	(30,1000)
NF	MAE	0.407	0.392	0.531	0.471	0.514	0.436	0.440
	RMSE	0.490	0.481	0.626	0.557	0.585	0.536	0.506
	CC	0.339	0.424	0.303	0.350	0.266	0.371	0.375
	WI	0.580	0.641	0.522	0.516	0.455	0.610	0.516
	Structure	--	--	--	--	--	--	--
NF-SC	MAE	0.434	0.332	0.607	0.454	0.383	0.396	0.537
	RMSE	0.526	0.419	0.690	0.562	0.479	0.479	0.612
	CC	0.464	0.398	0.441	0.354	0.236	0.595	0.288
	WI	0.678	0.611	0.540	0.607	0.492	0.740	0.481
	Radii	0.7	0.9	0.7	0.6	0.7	0.7	0.8
NF-FCM	MAE	0.395	0.411	0.422	0.420	0.430	0.458	0.499
	RMSE	0.469	0.521	0.510	0.500	0.487	0.537	0.584
	CC	0.465	0.446	0.433	0.409	0.270	0.419	0.103
	WI	0.642	0.643	0.653	0.634	0.432	0.629	0.388
	Cluster	2	4	3	3	4	3	2

2.3.4 Grey wolf optimizer

GWO is an evolutionary algorithm introduced by Mirjalili et al. (2014). This algorithm follows the leadership hierarchy structure, which consists of an average of 5–12 wolves. The leadership hierarchy is based on four types of grey wolves, namely, alpha (α), beta (β), delta (δ), and omega (ω). The order of the solutions in this algorithm after alpha is considered β, δ, and ω. Therefore, the major stages of grey wolf hunting are as follows. The alpha wolf (α) plays the main role which includes hunting, sleeping, and waking hours. The β is responsible for performing alpha commands over the pack. On the other hand, ω is seen as a victim in the group and therefore follows other wolves and can eat after all of them. Finally, δ wolves must follow β and α wolves, while they dominate ω wolves. Figure 4 shows the diagram of GWO.

Grey wolf hunting behavior modeling assumes that alpha, beta, and delta have sufficient knowledge of prey position. Therefore, three optimal solutions, including alpha, beta, and delta are obtained, and the other solutions (wolves) must change their position based on the optimal solutions.

2.3.5 Proposed hybrid method

The coupled NF and GWO were developed as NF-GWO to predict the DO concentration. In the NF-GWO model, GWO optimizes NF parameters for best performance. NF-GWO consists of five layers. The first layer's nodes represent the input variables. The second and third layers represent the membership functions for the input variables and the fuzzy logic rules, respectively. In the fourth layer, Takagi-Sugeno-Kang's model adjusts the performance of the nodes. Finally, the DO concentration is predicted in the last layer (output layer). During the training phase, the GWO generates the initial population of wolves and updates the solutions based on the DO concentration prediction accuracy. The solutions are continuously updated unless the algorithm reached the maximum number of iterations or errors less than the sill value. The parameters found in the last step are transferred to the structure of the NF model. The initial population and the number of iterations for each combination were determined through the trial-and-error method. The initial population and the number of iterations are listed in the structure row of the results table. Population values ranged from 20 to 40 and iterations from 1,000 to 1,500.

2.4 Performance statistics for the evaluation of the models

Four statistical metrics named root mean squared error (RMSE), mean absolute error (MAE) correlation coefficient (CC), and Willmott Index (WI), were applied to assess the performance of the applied models in predicting DO. Among the four metrics, two

10.3389/fenvs.2022.929707

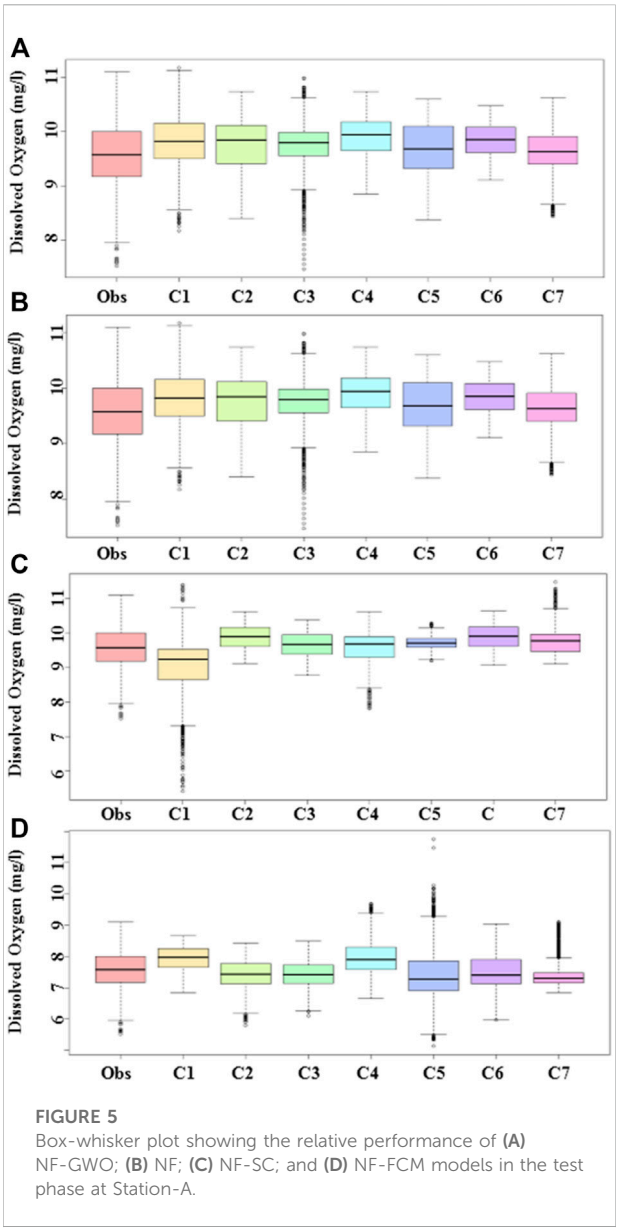


FIGURE 5
Box-whisker plot showing the relative performance of **(A)** NF-GWO; **(B)** NF; **(C)** NF-SC; and **(D)** NF-FCM models in the test phase at Station-A.

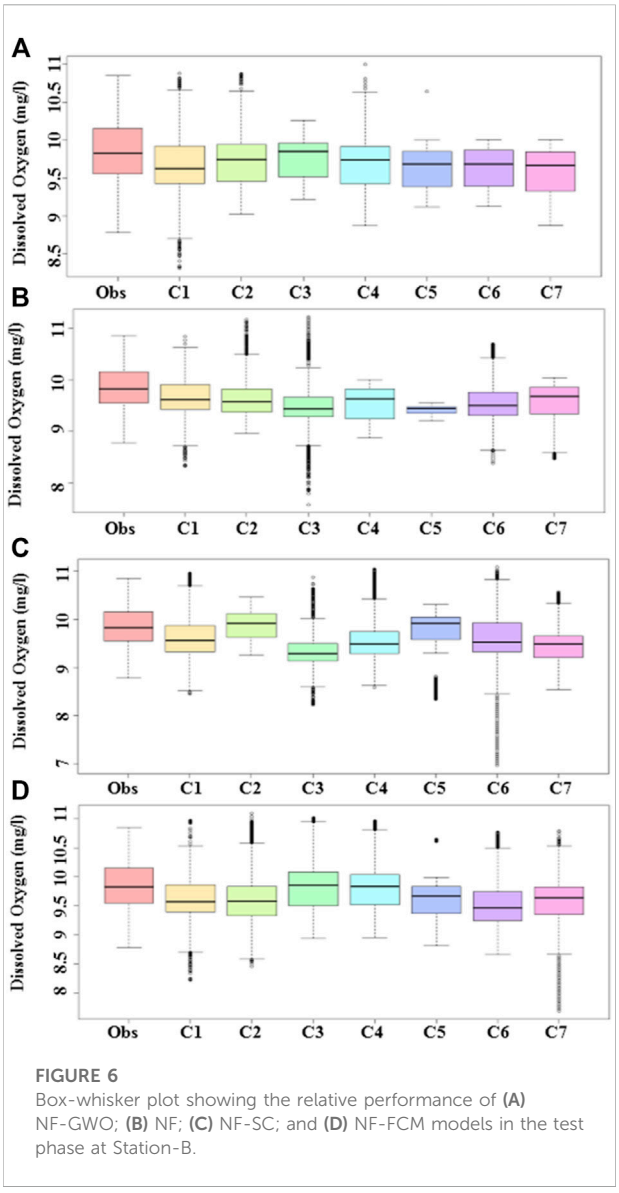


FIGURE 6
Box-whisker plot showing the relative performance of **(A)** NF-GWO; **(B)** NF; **(C)** NF-SC; and **(D)** NF-FCM models in the test phase at Station-B.

(MAE and RMSE) were used to evaluate the error in the models, and two (CC and WI) were utilized to assess the models' ability to simulate the temporal pattern of observed DO. They are defined as:

$$RMSE = \sqrt{1/n \sum_{i=1}^{N} (O_i - P_i)^2}, \quad (6)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |O_i - P_i|, \quad (7)$$

$$CC = \frac{\sum_{i=1}^{N} (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{N} (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^{N} (P_i - \bar{P})^2}}, \quad (8)$$

$$WI = 1 - \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} \left(\left| P_i - \overline{O_i} \right| + \left| |O_i - \overline{O_i}| \right| \right)^2}, \quad (9)$$

where O_i and P_i are observed, and the predicted DO value for *ith* observations. $\bar{O}$ and $\bar{P}$ are the mean of observed and the predicted DO, and n is data points. Four statistical metrics and visual inspection using scatter plot, box plot, violin plot, and Taylor diagram were used to assess the performance of the applied models. Different statistical metrics and plots provide different measures of model performance, including error, association, and distribution. Therefore, they were used in this study for a complete assessment of model performance.

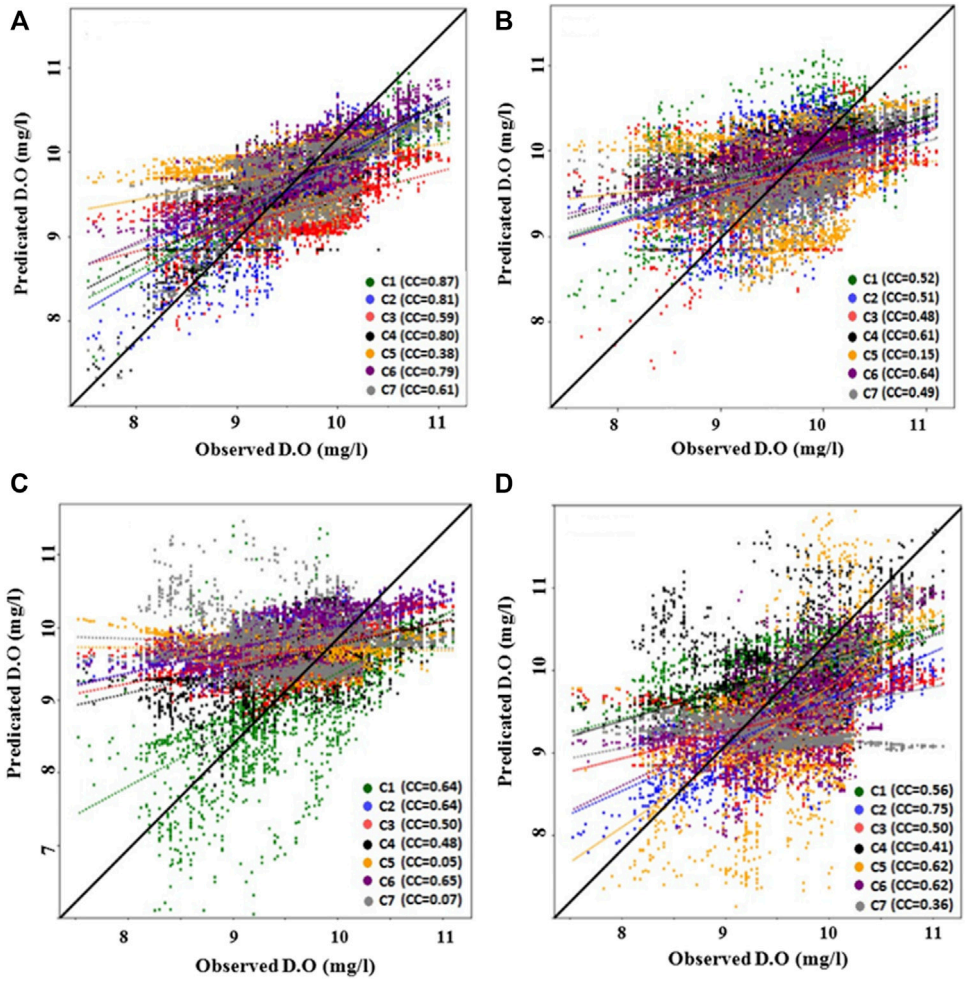


FIGURE 7
Scatter plot of observed and predicted dissolved oxygen by **(A)** NF-GWO; **(B)** NF; **(C)** NF-SC; and **(D)** NF-FCM models during the testing phase at Station-A.

3 Results and discussion

3.1 Assessment of the models

In order to assess the performance of the proposed hybrid model, its results have been compared with the results of the standalone NF model. The hybrid models are adopted in order to adjust the hyper-parameters of the standalone model. In this study, different optimization algorithms including GWO, SC, and FCM are used to adjust the parameters of the NF model where these algorithms examine different regions of the search space which has several local minima, and then minimize the range of search to the region that includes the global minima.

Obtained results at Station-A and Station-B are presented in Tables 3 and 4, respectively. The results at Station-A showed large variability in model performance for different input combinations. All the models, except NF-GWO, also showed different performances in terms of different metrics. For example, NF for input combination, C7 showed the best performance based on MAE and RMSE, while it showed the best performance in terms of CC and WI with input combination C4. Similar inconsistency was noticed for NF-SC and NF-FCM. However, NF-GWO performed best in terms of all metrics for first input combination, C1 (MAE = 0.250 mg/L, RMSE = 0.320 mg/L, CC = 0.869, and WI = 0.908). Comparison model performance revealed the better performance of NF-GWO for all input combinations compared to other models. Therefore, NF-GWO for the input combination of C1 can be considered the best model for predicting DO at Station-A. Different models showed the worst performance for different input combinations. For example, NF-SC performed the worst for C1 while NF-FCM performed the worst for C5. Overall, NF-SC performed the worst

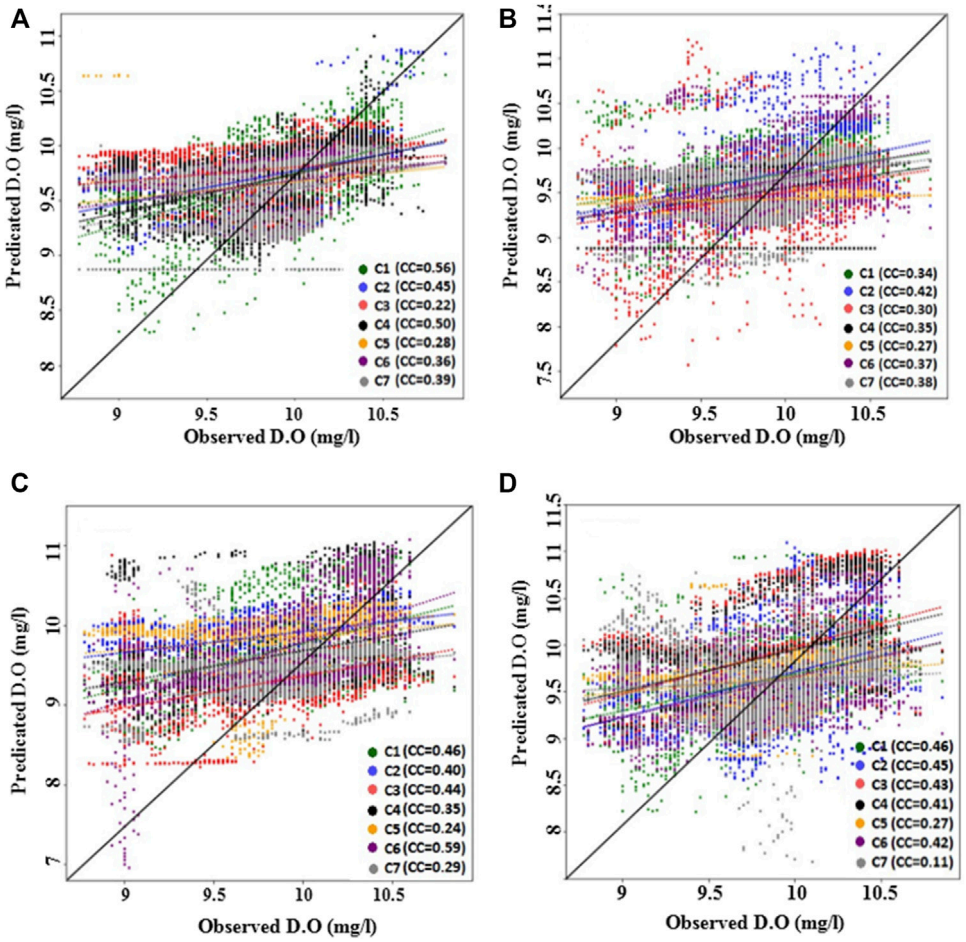


FIGURE 8
Scatter plot of observed and predicted dissolved oxygen by **(A)** NF-GWO; **(B)** NF; **(C)** NF-SC; and **(D)** NF-FCM models during the testing phase at Station-B.

for most of the input combinations and in terms of different metrics.

The performance of the models at Station-B was found very similar to that at Station-A (Table 4). Large variability in model performance for different input combinations was also noticed at this station. Only NF-GWO showed consistent performance in terms of all metrics. It also showed better performance compared to other models for all input combinations. Comparison of model performance for different input combinations revealed best performance of NF-GWO for C1 (MAE = 0.346 mg/L, RMSE = 0.413 mg/L, CC = 0.563, and WI = 0.713) at this station. Like Station-A, NF-SC showed the worst performance at this station for most of the input combinations. The C1 scenario considers all variables (TE, SC, Tu, and pH) as inputs, while other scenarios omit one to more variables. The better performance of NF-GWO for C1 indicates all variables are required to consider for better prediction of DO because the use of a few water quality parameters may drive to missing the required information about the effect of these parameters on dissolved oxygen concentration. Therefore, to obtain more realistic results and in order to investigate the effect of each water quality parameter on DO concentration, all possible input combinations should be considered.

The observed and predicted DO by different models for different input combinations at Station-A and Station-B are presented using box-–whisker plots in Figures 5, 6, respectively. A box with a whisker presents mean (horizontal line within the box), 25th and 75th percentiles (lower and upper bound of the box), range without outliers (spread of the whiskers), and outliers (dots). Therefore, a comparison of the whisker–boxes provides a model performance assessment in terms of median, range, and outliers. The box plot of different models was found similar to the observed one for different input combinations. For example, the whisker–box of NF-GWO for C1 was found more similar to the observed one, while the

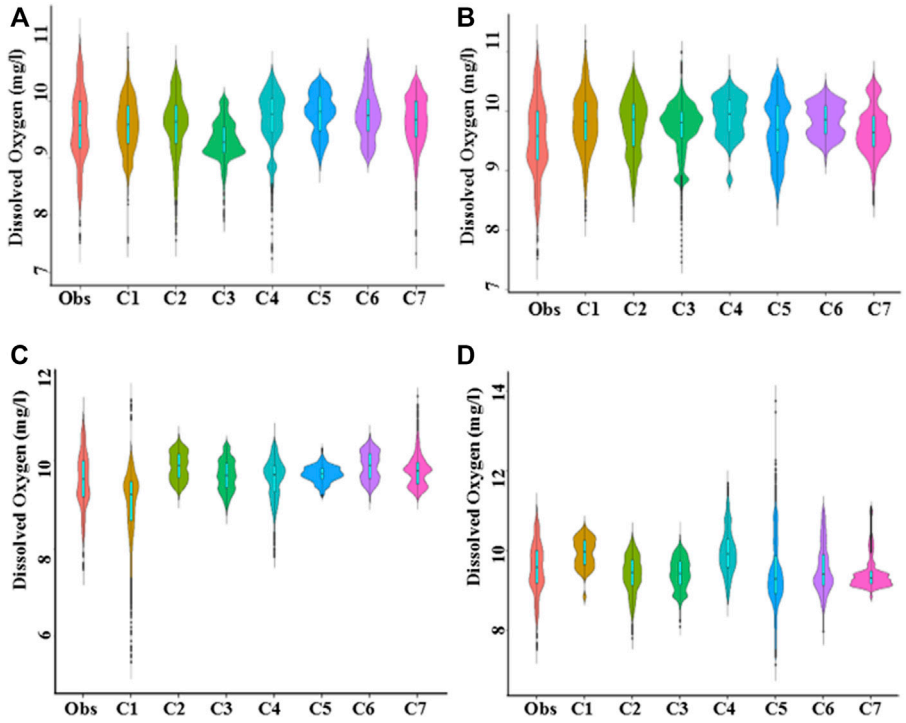


FIGURE 9
Split–violin plot showing the relative performance of **(A)** NF-GWO; **(B)** NF; **(C)** NFSC; and **(D)** NF-FCM models in replicating probability distribution of observed data during testing phase at Station-A.

whisker–box of NF was found more similar to the observed one for C5. A comparison of whisker–box plots of all models for all input combinations revealed the best performance of NF-GWO for C1. The results were found a bit different at Station-B. The best performance was noticed for NF-FCM with C4. It was able to replicate the median, interquartile range, and the range of the data more accurately. However, the simulated data by NF-FCM for C4 was found a bit right-skewed compared to observed data. The NF-GWO for C1, which showed the best performance at Station-A, also performed well at this station but underestimated the DO values.

The scatter plots of observed and predicted DO are presented in Figures 7, 8, respectively. Different plots in each figure show the performance of a model for different input combinations. Different colors are used to show the model performance for different input combinations. The scatter plots at Station-A (Figure 7) showed better performance of NF-GWO for C1, NF for C1, NF-SC for both C1 and C2, and NF-FCM for C2. Overall, most of the models showed better performance for the input combination (C1). The performance comparison of the models revealed a much higher performance of NF-GWO compared to other models. The NF-GWO for C1 replicated the observed DO with R^2 (determination coefficient) = 0.75. The NF-GWO predicted DO for C1 was found more aligned to the plots' diagonal line compared to other models. Though an underprediction for high values and overprediction for low values was noticed, the model could still predict most of the high and low values.

The scatter plots of the models' predictions at Station-B (Figure 8) showed more inconsistency than at Station-A. The NF-GWO performed best for C1, NF for C2, NF-SSC for C6, and NF-FCM for C1 at this station. All models showed a lower performance at this station compared to that found at Station-A. However, NF-GWO for C1 showed the best performance at this station like Station-A. Though it underestimated many observed values, it was most aligned to the scatter plot's diagonal line compared to other models for different input combinations.

The capability of the models to reconstruct the distribution of observed DO was estimated using violin plots. The plots for the models at Station-A and Station-B are presented in Figures 9, 10, respectively. If the shape of a model's violin is similar to the violin of observed DO, the model is considered good. The violins of different models' predicted DO showed large variability in shape and size (Figure 9). None of the models was able to replicate the shape of the observed violin accurately.

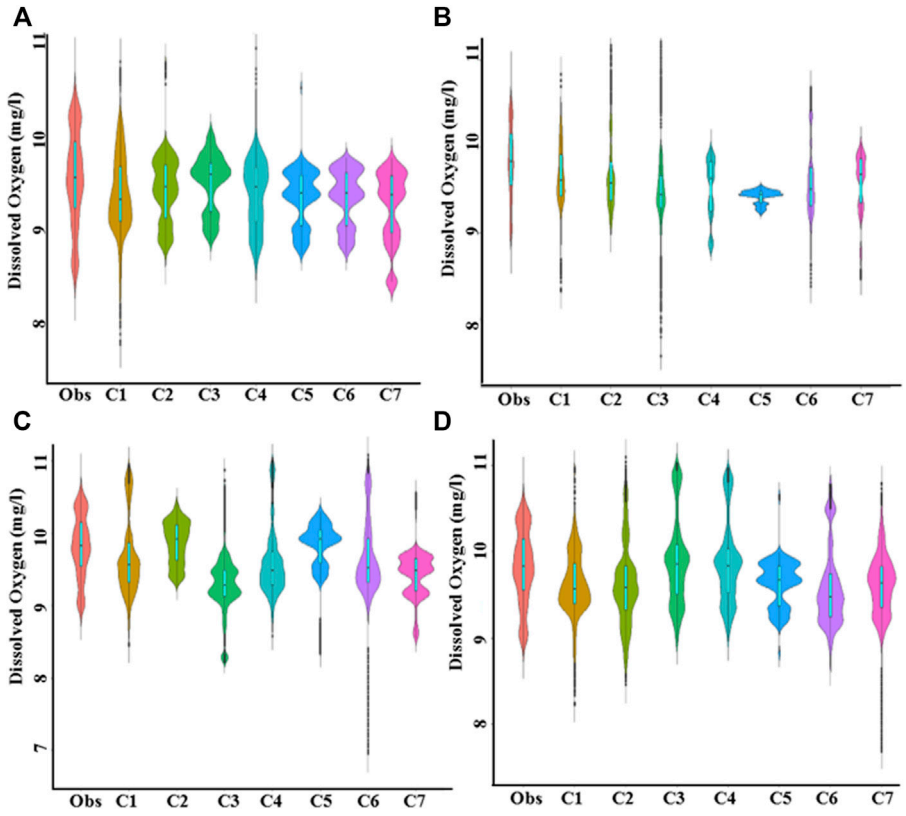


FIGURE 10
Split–violin plot showing the relative performance of **(A)** NF-GWO; **(B)** NF; **(C)** NF-SC; and **(D)** NF-FCM models in replicating the probability distribution of observed data during the testing phase at Station-B.

Overall, the violin of NF-GWO for C1 was most similar to the violin of the observed data. The dissimilarity with the observed violin was more for NF-FCM. The dissimilarity between models' predicted and observed violins was more at Station-B (Figure 10) than that noticed at Station-A. In most of the cases, the models failed to reconstruct the distribution of observed DO. Overall, NF-GWO for C1 has the best performance at this station.

Finally, the Taylor diagram was developed to show the relative accuracy of model predictions. Taylor diagram graphically compared association, the similarity invariance, and the mean difference between observed and model output. Therefore, it is considered a composite way to compare model performance. Taylor diagrams of the models for Station-A and Station-B are presented in Figures 11, 12, respectively. The black dot on the diagram's *x*-axis represents observed data, while different colors present model performance for different input combinations. A model nearest to the observed point indicates better performance. The Taylor diagram also showed large variability in models' performance for different input combinations at Station-A (Figure 11). However, all the models, except NF-FCM, performed best for the first input combination (C1). The NF-FCM performed best for C6. The NF-GWO models were nearest to the observation compared to other models for different input combinations. Overall, the results identified NF-GWO for C1 as the best model at Station-A. Inconsistency in model performance was noticed at Station-B (Figure 12), similar to that noticed using the scatter plot and the box plot. The NF-GWO showed the best performance for C1, NF for C2, NF-SC for C6, and NF-FCM for C1. The performance of the models at Station-B was poor compared to that found at Station-A. However, the best model was still NF-GWO for C1 at this station. It showed a correlation of 0.58 and an RMSE of less than 0.40. The variability of predicted DO by NF-GWO for C1 was 0.38, which was very near the observed DO variability (0.44).

3.2 Discussion

The performance of NF-GWO compared to standalone NF was improved by 24.2–66.2% and 14.9–31.2% in terms of

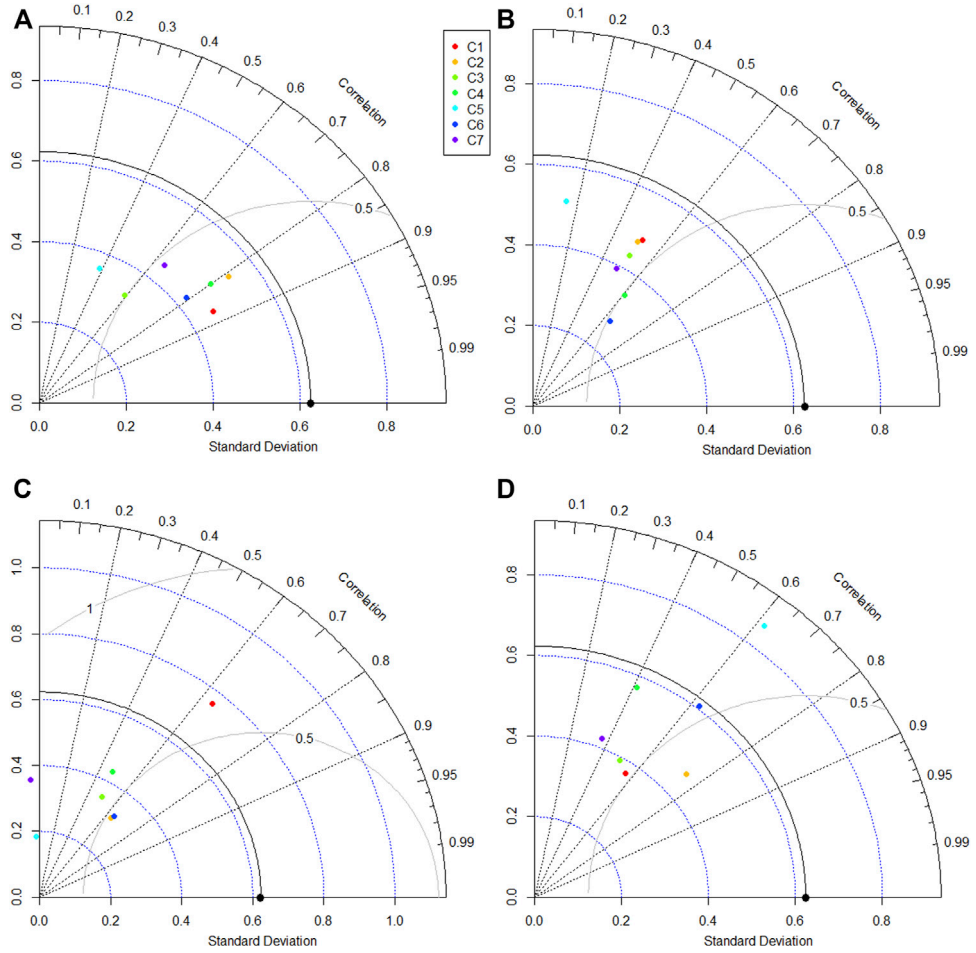


FIGURE 11
Taylor diagram of the models used to compare the capability of **(A)** NF-GWO; **(B)** NF; **(C)** NF-SC; and **(D)** NF-FCM models in the testing phase at Station-A.

CC and WI for different input combinations. The MAE and RMSE of the NF model were reduced using the NF-GWO model by 9.9–46.0% and 8.9–47.5%, respectively. A similar improvement in DO concentration prediction was achieved using NF-GWO models compared to NF-SC and NF-FCM for most input combinations. However, the improvement was not consistent in terms of all statistics, as mentioned earlier. The performance of NF-GWO compared to NF-SC was improved for all input combinations except for C3 at Station-A. The improvement in MAE and RMSE was in the range of 7.7–60.7 and 6.8 to 60.4, respectively. However, MAE and RMSE were decreased for C3 by 5.7% and 2.7%. Significant improvement in NF-GWO prediction compared to NF-FCM was also noticed for all input combinations, except C3 and C5 at Station-A. The MAE and RMSE were reduced by 3.4–50.3% and 10.1–49.5%, while CC and WI were increased by 7.9–94.6% and 4.6–40.8%. The most significant improvement in NF-GWO performance was for C1. The MAE, RMSE, CC, and WI values of NF-GWO prediction were improved by 46.0, 47.5, 66.1, and 31.2% compared to standalone NF, 60.7, 60.3, 35.9, and 32.9% compared to NF-SC, and 50.4, 49.5, 54.4 and 40.5% compared to NF-FCM at Station-A.

Improvement in DO prediction using NF-GWO was also noticed compared to other models for most of the input combinations at Station-B. However, the improvement was not as great as it was noticed for Station-A. For the best input combination (C1), the improvement in NF-GWO prediction compared to NF was 14.9, 15.7, 66.1, and 22.9% in MAE, RMSE, CC, and WI, respectively. Those values were 20.3, 21.5, 21.3 and -5.2%, respectively for NF-SC, and 12.4, 11.9, 21.1, and 11.1%, respectively, for NF-FCM. Here, it should be noted that, like Station-A, the biggest improvement in the NF-GWO model was not for C1 for all input combinations.

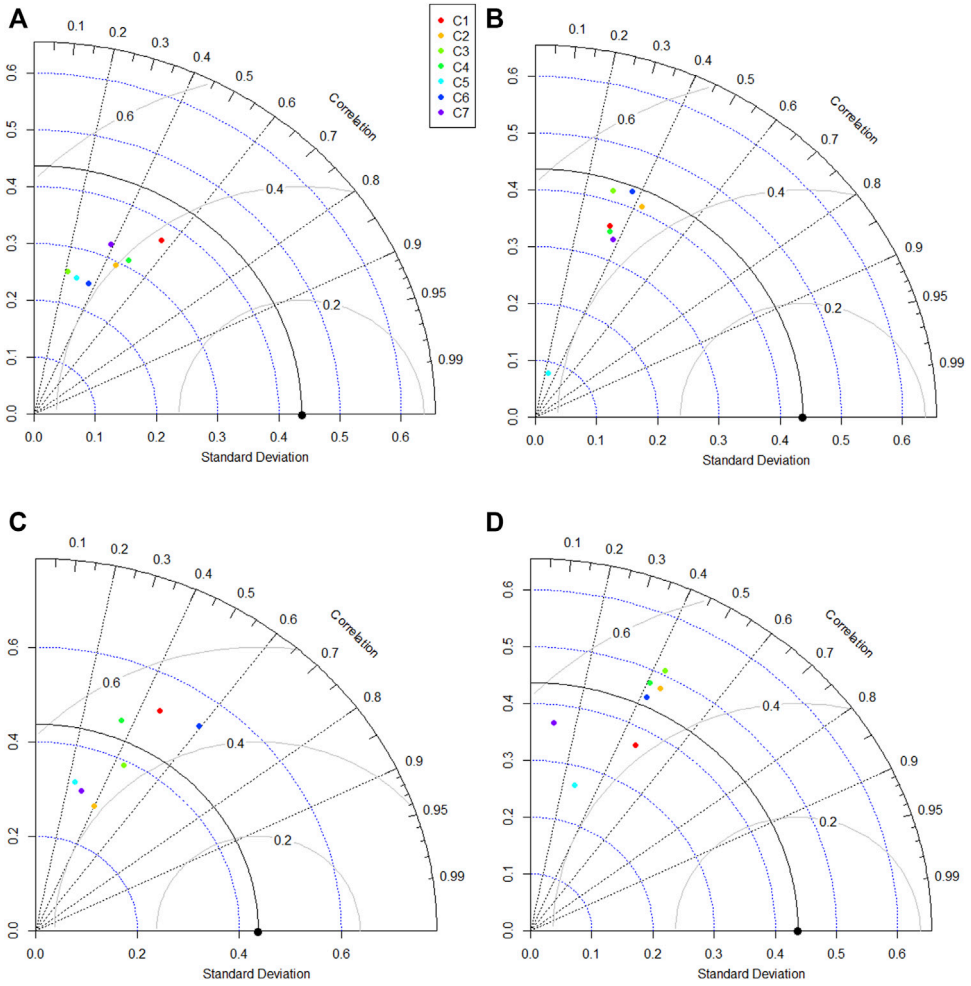


FIGURE 12
Taylor diagram of the models used to compare the capability of **(A)** NF-GWO; **(B)** NF; **(C)** NF-SC; and **(D)** NF-FCM models in the testing phase at Station-B

For example, the most significant improvement in NF-GWO compared to NF in MAE, RMSE, and WI were for C4. However, the highest improvement in CC was for C1.

Overall, the results revealed that only NF-GWO for C1 showed consistent DO concentration prediction improvement in all statistics. It means all the statistical metrics used in this study showed better performance of NF-GWO for C1. The NF-GWO also showed improvement in prediction for almost all input combinations. The best performance of NF-GWO for C7 was also consistent for both locations. Therefore, the results presented in this study revealed the ability of the GWO to enhance the performance of the standalone NF model and this agrees with results obtained by Ewees and Elaziz (2018) and (Dehghani et al., 2019). In addition, the results of the proposed model undoubtedly establish the efficacy of the NF-GWO model in predicting DO concentration.

NF integrates neural networks and fuzzy systems to join their advantages for a better solution to complex problems. The capability of NF to learn data patterns using fuzzy rules has made it highly adaptable to different kinds of data, and thus, superior to many other AI algorithms in solving a wide range of problems from different fields (Atmaca et al., 2001). However, the major drawback of NF is that its performance significantly is susceptible to the selection and optimization of the input variable's fuzzy membership function. In this study, it was solved by integrating NF with optimization algorithms. Therefore, the performance of NF-GWO was found better than the other version of NFs (Sremac et al., 2019). The major challenge in selecting and adjusting NF

hyperparameters is finding the global optimum solution (Negi et al., 2021).

Among varieties of optimization algorithms developed so far, the GWO, a metaheuristics optimization technique developed based on wolf behavior for preying, has shown its capability in solving complex optimization problems. This population-based or trajectory-based algorithm can search for a solution over a large complex space, thus less susceptible to being trapped in local minima. The recent review of GWO by Negi et al. (2021) showed the capability of GWO in optimizing a wide variety of engineering problems. The optimization of NF intern parameters using GWO has made the NF-GWO highly capable of predicting DO concentration.

It should be noted that an AI algorithm's performance depends on the problem to be solved and the kinds of input data used as predictors. Similarly, a particular optimization algorithm is always not the best for the optimization of an AI model hyperparameters. Different optimization algorithms can perform differently in the optimization of an AI algorithm in solving different problems. In this study, optimization of NF parameters using GWO made it highly capable in DO prediction. However, it does not guarantee that the NF-GWO model performs well in predicting other hydrological variables or the prevision of DO in other regions. Therefore, it is always suggested to compare different AI and optimization algorithms' performance to find the best empirical model for selecting the best model.

The sensitivity analysis of different input variables was conducted through the evaluation of best models' (NF-GWO) performance for different input combinations. First, the performance for C1 with the three input models (C2 to C4) was analyzed. The model performance for C2 to C4 was reduced due to the drop of an input variable. The highest drop in prediction accuracy was for C3 when the pH was dropped from the input combination. The MAE and RMSE of C3 were 90 and 76.6% higher, and CC and WI were 32.1 and 27.6% lower than C1 at Station-A. A similar result was also obtained at Station-B. The increase in prediction error was the highest for C3 compared to the other three-input models. This could be attributed to that the pH has a valuable impact on DO concentration and the absence of pH in C3 caused a large drop in model prediction accuracy. Therefore, pH can be considered the most sensitive input after TE in predicting DO concentration. The analysis of two input NF-GWO models' performance (C5 to C7) with the NF-GWO model for C1 showed the highest decrease in model performance for C5 at both stations. The results indicate that the reduction of prediction accuracy was due to the absence of Tur as input. This indicates Tur is the third most crucial variable for the prediction of DO concentration. Overall and according to the sensitivity of DO prediction accuracy to different input variables, it can be said that the TE has the highest influence on DO, followed by the pH, Tur, and SC variables, respectively.

4 Conclusion

The present study assessed the ability of hybrid NF models in predicting DO concentration at two stations located in California. Different combinations of water quality parameters including temperature, specific conductivity, turbidity, and pH parameters were formulated and used as input to these models. Entropy and the correlation coefficient were used to evaluate these parameters in order to obtain the optimum input combination. The result showed that the best-input combinations are four input variables, namely, TE, SC, Tu, and pH. Among the four models developed in this study, only the NF-GWO showed consistent performance for all input combinations at both stations in terms of all metrics. The NF-GWO attained the highest performance (MAE = 0.256 mg/L, RMSE = 0.320 mg/L, CC = 0.869, and WI = 0.908) at Station-A and (MAE = 0.346 mg/L, RMSE = 0.413 mg/L, CC = 0.563, and WI = 0.713) at Station B. Also, the performance of NF-GWO compared to standalone NF was improved by 24.2–66.2% and 14.9–31.2% in terms of CC and WI for different input combinations and the MAE and RMSE of the NF model were reduced using the NF-GWO model by 9.9–46.0% and 8.9–47.5%, respectively, for Station-A, while for Station B, the improvement in NF-GWO prediction compared to NF was 14.9, 15.7, 66.1, and 22.9% in MAE, RMSE, CC, and WI, respectively. The sensitivity analysis of input parameters revealed that water temperature followed by pH and specific conductivity is the most important for DO concentration prediction in the study area. Using direct methods to measure DO concentration is costly and time-consuming. The hybrid AI models can be used for reliable estimation of DO concentration in such a situation. Finally, although the developed hybrid AI models in this study achieve high performance, there are still some limitations due to practical factors. The most important one is the data used, where the data of only 1 year was used in the study. Therefore, the capability of the suggested model should be evaluated using long-term data. In addition, other metaheuristics optimization algorithms can be used for the optimization of the NF model to evaluate their relative performance in improving prediction accuracy.

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 authors.

Author contributions

All the authors contributed to conceptualizing and designing of the study. Data were gathered by SSS; the modeling was carried out by SM, AM, and SS; an initial draft of themanuscript was prepared by SSS, AM, SA, and the manuscript was repeatedly

revised to generate the final version by SS and EM; the work was supervised by NA.

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

Abba, S., Hadi, S. J., and Abdullahi, J. (2017). River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. *Procedia Comput. Sci.* 120, 75–82. doi:10.1016/j.procs.2017.11.212

Abba, S., Hadi, S. J., Sammen, S. S., Salih, S. Q., Abdulkadir, R., Pham, Q. B., et al. (2020). Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. *J. Hydrol.* 587, 124974. doi:10.1016/j.jhydrol.2020.124974

Abba, S. I., Abdulkadir, R. A., Sammen, S., S., Usman, A. G., Meshram, S G., Malik, A., et al. (2021). Comparative implementation between neuro-emotional genetic algorithm and novel ensemble computing techniques for modelling dissolved oxygen concentration. *Hydrol. Sci. J.* 66 (10), 1584–1596. doi:10.1080/02626667.2021.1937179

Ahmed, A. M., and Shah, S. M. A. (2017). Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River. *J. King Saud Univ. - Eng. Sci.* 29 (3), 237–243. doi:10.1016/j.jksues.2015.02.001

Antanasijević, D., Pocajt, V., Perić-Grujić, A., and Ristić, M. (2019). Multilevel split of high-dimensional water quality data using artificial neural networks for the prediction of dissolved oxygen in the Danube River. *Neural comput. Appl.* 32, 3957–3966. doi:10.1007/s00521-019-04079-y

Atmaca, H., Cetisli, B., and Yavuz, H. S. (2001). "The comparison of fuzzy inference systems and neural network approaches with ANFIS method for fuel consumption data," in Second international conference on electrical and electronics engineering papers ELECO.

Banadkooki, F. B., Ehteram, M., Panahi, F., Sammen, S. S., Othman, F. B., El-Shafie, A., et al. (2020). Estimation of total dissolved solids (TDS) using new hybrid machine learning models. *J. Hydrol.* 587, 124989. doi:10.1016/j.jhydrol.2020.124989

Bezdek, J. C. (1973). Cluster validity with fuzzy sets. *J. Cybern.* 3 (3), 58–73. doi:10.1080/01969727308546047

Cao, W., Huan, J., Liu, C., Qin, Y., and Wu, F. (2019). A combined model of dissolved oxygen prediction in the pond based on multiple-factor analysis and multi-scale feature extraction. *Aquac. Eng.* 84, 50–59. doi:10.1016/j.aquaeng.2018.12.003

Chen, W.-B., and Liu, W.-C. (2014). Artificial neural network modeling of dissolved oxygen in reservoir. *Environ. Monit. Assess.* 186 (2), 1203–1217. doi:10.1007/s10661-013-3450-6

Dehghani, M., Riahi-Madvar, H., Hooshyaripor, F., Mosavi, A., Shamshirband, S., Zavadskas, E. K., et al. (2019). Prediction of hydropower generation using grey wolf optimization adaptive neuro-fuzzy inference system. *Energies* 12, 289. doi:10.3390/en12020289

Ebtehaj, I., Sammen, S. S., Lariyah, M. S., Malik, A., Sihag, P., Al-Janabi, A. M. S., et al. (2021). Prediction of daily water level using new hybridized GS-GMDH and ANFIS-FCM models. *Eng. Appl. Comput. Fluid Mech.* 15 (1), 1343–1361. doi:10.1080/19942060.2021.1966837

Elkiran, G., Nourani, V., Abba, S., and Abdullahi, J. (2018). Artificial intelligence-based approaches for multi-station modelling of dissolve oxygen in river. *Glob. J. Environ. Sci. Manag.* 4 (4), 439–450. doi:10.22034/GJESM.2018.04.005

Elkiran, G., Nourani, V., and Abba, S. (2019). Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *J. Hydrol.* 577, 123962. doi:10.1016/j.jhydrol.2019.123962

Ellenburg, W. L., Cruise, J., and Singh, V. P. (2018). The role of evapotranspiration in streamflow modeling–An analysis using entropy. *J. Hydrol.* 567, 290–304. doi:10.1016/j.jhydrol.2018.09.048

Ewees, A. A., and Elaziz, M. A. (2018). Improved adaptive neuro-fuzzy inference system using gray wolf optimization: A case study in predicting biochar yield. *J. Intell. Syst.* 29 (1), 924–940. doi:10.1515/jisys-2017-0641

Hadi, S. J., Abba, S. I., Sammen, S. S., Salih, S. Q., Al-Ansari, N., and Yaseen, Z. M. (2019). Non-linear input variable selection approach integrated with non-tuned data intelligence model for streamflow pattern simulation. *IEEE Access* 7, 141533–141548. doi:10.1109/ACCESS.2019.2943515

Hameed, M., Sharqi, S. S., Yaseen, Z. M., Afan, H. A., Hussain, A., and Elshafie, A. (2017). Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. *Neural comput. Appl.* 28 (1), 893–905. doi:10.1007/s00521-016-2404-7

Jang, J.-S. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man. Cybern.* 23 (3), 665–685. doi:10.1109/21.256541

Kisi, O., Alizamir, M., and Gorgij, A. D. (2020). Dissolved oxygen prediction using a new ensemble method. *Environ. Sci. Pollut. Res.* 27, 9589–9603. doi:10.1007/s11356-019-07574-w

Liu, Y., Zhang, Q., Song, L., and Chen, Y. (2019). Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction. *Comput. Electron. Agric.* 165, 104964. doi:10.1016/j.compag.2019.104964

Malik, A., Tikhamarine, Y., Al-Ansari, N., Shahid, S., Sekhon, H. S., Pal, R. K., et al. (2021). Daily pan-evaporation estimation in different agro-climatic zones using novel hybrid support vector regression optimized by Salp swarm algorithm in conjunction with gamma test. *Eng. Appl. Comput. Fluid Mech.* 15, 1075–1094. doi:10.1080/19942060.2021.1942990

Maroufpoor, S., Bozorg-Haddad, O., and Maroufpoor, E. (2020). Reference evapotranspiration estimating based on optimal input combination and hybrid artificial intelligent model: Hybridization of artificial neural network with grey wolf optimizer algorithm. *J. Hydrol.* 588, 125060. doi:10.1016/j.jhydrol.2020.125060

Maroufpoor, S., Maroufpoor, E., Bozorg-Haddad, O., Shiri, J., and Yaseen, Z. M. (2019a). Soil moisture simulation using hybrid artificial intelligent model: Hybridization of adaptive neuro fuzzy inference system with grey wolf optimizer algorithm. *J. Hydrol.* 575, 544–556. doi:10.1016/j.jhydrol.2019.05.045

Maroufpoor, S., Sanikhani, H., Kisi, O., Deo, R. C., and Yaseen, Z. M. (2019b). Long-term modelling of wind speeds using six different heuristic artificial intelligence approaches. *Int. J. Climatol.* 1, 3543–3557. doi:10.1002/joc.6037

Meidute-Kavaliauskiene, I., Jabehdar, M. A., Davidavičienė, V., Ghorbani, M. A., and Sammen, S. S. (2021). A simple way to increase the prediction accuracy of hydrological processes using an artificial intelligence model. *Sustainability* 13, 7752. doi:10.3390/su13147752

Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). Grey wolf optimizer. *Adv. Eng. Softw.* 69, 46–61. doi:10.1016/j.advengsoft.2013.12.007

Mohammadi, B., Linh, N. T. T., Pham, Q. B., Ahmed, A. N., Vojtekova, J., Guan, Y., et al. (2020). Adaptive neuro-fuzzy inference system coupled with shuffled frog leaping algorithm for predicting river streamflow time series. *Hydrol. Sci. J.* 65 (10), 1738–1751. doi:10.1080/02626667.2020.1758703

Negi, G., Kumar, A., Pant, S., and Ram, M. (2021). GWO: a review and applications. *Int. J. Syst. Assur. Eng. Manag.* 12 (1), 1–8. doi:10.1007/s13198-020-00995-8

Nourani, V., Elkiran, G., and Abba, S. (2018). Wastewater treatment plant performance analysis using artificial Intelligence–an ensemble approach. *Water Sci. Technol.* 78 (10), 2064–2076. doi:10.2166/wst.2018.477

Nourani, V., Elkiran, G., and Abdullahi, J. (2019). Multi-station artificial Intelligence based ensemble modeling of reference evapotranspiration using pan evaporation measurements. *J. Hydrology* 577, 123958. doi:10.1016/j.jhydrol.2019.123958

Pham, D. T., Ho, L., Espinoza-Palacios, J., Arevalo-Durazno, M., Van Echelpoel, W., and Goethals, P. (2020). Generalised linear models for prediction of dissolved oxygen in a waste stabilisation pond. *Water* 12, 1930. doi:10.3390/w12071930

Pham, Q. B., Abba, S. I., Usman, A. G., Linh, N. T. T., Gupta, V., Malik, A., et al. (2019). Potential of hybrid data-intelligence algorithms for multi-station modelling of rainfall. *Water Resour. manage.* 33 (15), 5067–5087. doi:10.1007/s11269-019-02408-3

Pham, Q. B., Mohammadpour, R., Linh, N. T. T., Mohajane, M., Pourjasem, A., Sammen, S. S., et al. (2021a). Application of soft computing to predict water quality in wetland. *Environ. Sci. Pollut. Res.* 28, 185–200. doi:10.1007/s11356-020-10344-8

Pham, Q. B., Sammen, S. S., Abba, S. I., Mohammadi, B., Shahid, S., and Abdulkadir, R. A. (2021b). A new hybrid model based on relevance vector machine with flower pollination algorithm for phycocyanin pigment concentration estimation. *Environ. Sci. Pollut. Res.* 28, 32564–32579. doi:10.1007/s11356-021-12792-2

Pourabadeh, A., Baharinikoo, L., Shojaei, S., Mehdizadeh, B., Davoodabadi Farahani, M., Shojaei, S., et al. (2020). Experimental design and modelling of removal of dyes using nano-zero-valent iron: a simultaneous model. *Int. J. Environ. Anal. Chem.* 100 (15), 1707–1719. doi:10.1080/03067319.2019.1657855

Radwan, M., Willems, P., El-Sadek, A., and Berlamont, J. (2003). Modelling of dissolved oxygen and biochemical oxygen demand in river water using a detailed and a simplified model. *Int. J. River Basin Manag.* 1 (2), 97–103. doi:10.1080/15715124.2003.9635196

Rahman, A., Dabrowski, J., and McCulloch, J. (2020). Dissolved oxygen prediction in prawn ponds from a group of one step predictors. *Inf. Process. Agric.* 7 (2), 307–317. doi:10.1016/j.inpa.2019.08.002

Sammen, S. S., Ehteram, M., Abba, S. I., Abdulkadir, R. A., Ahmed, A. N., and El-Shafie, A. (2021). A new soft computing model for daily streamflow forecasting. *Stoch. Environ. Res. Risk Assess.* 35, 2479–2491. doi:10.1007/s00477-021-02012-1

Seyedzadeh, A., Maroufpoor, S., Maroufpoor, E., Shiri, J., Bozorg-Haddad, O., and Gavazi, F. (2020). Artificial intelligence approach to estimate discharge of drip tape irrigation based on temperature and pressure. *Agric. Water Manag.* 228, 105905. doi:10.1016/j.agwat.2019.105905

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27 (3), 623–656. doi:10.1002/j.1538-7305.1948.tb00917.x

Shojaei, S., Shojaei, S., Nouri, A., and Baharinikoo, L. (2021). Application of chemometrics for modeling and optimization of ultrasound-assisted dispersive liquid-liquid microextraction for the simultaneous determination of dyes. *Npj Clean. Water* 4, 23. doi:10.1038/s41545-021-00113-6

Shojaei, S., and Shojaei, S. (2019). Optimization of process variables by the application of response surface methodology for dye removal using nanoscale zero-valent iron. *Int. J. Environ. Sci. Technol. (Tehran).* 16 (8), 4601–4610. doi:10.1007/s13762-018-1866-9

Shojaei, S., Shojaei, S., and Pirkamali, M. (2019). Application of box–behnken design approach for removal of acid black 26 from aqueous solution using zeolite: Modeling, optimization, and study of interactive variables. *Water Conserv. Sci. Eng.* 4 (1), 13–19. doi:10.1007/s41101-019-00064-7

Singh, V. P. (2013a). *Entropy theory and its application in environmental and water engineering*. Oxford, United Kingdom: John Wiley & Sons. doi:10.1002/9781118428306

Singh, V. P. (2013b). SCS-CN method revisited using entropy theory. *Trans. ASABE* 56 (5), 1805–1820. doi:10.13031/trans.56.10236

Sremac, S., Kazimieras Zavadskas, E., Matić, B., Kopić, M., and Stević, Ž. (2019). Neuro-fuzzy inference systems approach to decision support system for economic order quantity. *Econ. research-Ekonomska istraživanja* 32 (1), 1114–1137. doi:10.1080/1331677X.2019.1613249

Tanaka, K. (1997). *An introduction to fuzzy logic for practical applications*. New York, NY: Springer.

Wen, X., Fang, J., Diao, M., and Zhang, C. (2013). Artificial neural network modeling of dissolved oxygen in the Heihe River, Northwestern China. *Environ. Monit. Assess.* 185 (5), 4361–4371. doi:10.1007/s10661-012-2874-8

Wu, J., and Yu, X. (2021). Numerical investigation of dissolved oxygen transportation through a coupled SWE and Streeter–Phelps model. *Math. Probl. Eng.* 2021, 6663696. doi:10.1155/2021/6663696

Xiao, Z., Peng, L., Chen, Y., Liu, H., Wang, J., and Nie, Y. (20172017). The dissolved oxygen prediction method based on neural network. *Complexity*, 1–6. doi:10.1155/2017/4967870

Yang, J., Shojaei, S., and Shojaei, S. (2022). Removal of drug and dye from aqueous solutions by graphene oxide: Adsorption studies and chemometrics methods. *Npj Clean. Water* 5 (1), 5. doi:10.1038/s41545-022-00148-3

Yaseen, Z. M., Faris, H., and Al-Ansari, N. (20202020). Hybridized extreme learning machine model with salp swarm algorithm: a novel predictive model for hydrological application. *Complexity*, 1–14. doi:10.1155/2020/8206245

Yavari, S., Maroufpoor, S., and Shiri, J. (2018). Modeling soil erosion by data-driven methods using limited input variables. *Hydrology Res.* 49 (5), 1349–1362. doi:10.2166/nh.2017.041

Zaher, M. Y., Ahmed, E.-s., Othman, J., Haitham, A. A., and Khamis, N. S. (2015). Artificial intelligence based models for stream-flow forecasting: 2000-2015. *J. Hydrol. (Amsterdam)* 530, 829–844. doi:10.1016/j.jhydrol.2015.10.038

frontiers | Frontiers in Environmental Science

ORIGINAL RESEARCH
published: 29 August 2022
doi: 10.3389/fenvs.2022.937405

Evidence of the Middle-Income Trap in Latin American Countries: Factor Analysis Approach Using Regression and the ARDL Model

Lisette Paola Cervantes Galvan[1]*, *Carlos Carmona Campo*[2], *Savo Stanojevic*[1] and *David Villada Alzate*[3]

[1]School of Economics and Management, Southeast University, Nanjing, China, [2]Department of Economic Science, Management and Accounting, Universidad De La Costa, Barranquilla, Colombia, [3]Corporación Universitaria Americana, Facultad de Ciencias Administrativas Económicas y Contables, Barranquilla, Colombia

OPEN ACCESS

Edited by:
Wendong Yang,
Shandong University of Finance and Economics, China

Reviewed by:
Muhammad Kamran Khan,
Bahria University, Pakistan
Vishal Dagar,
Great Lakes Institute of Management, India

***Correspondence:**
Lisette Paola Cervantes Galvan
233169927@seu.edu.cn

Specialty section:
This article was submitted to Environmental Economics and Management, a section of the journal Frontiers in Environmental Science

Received: 06 May 2022
Accepted: 13 June 2022
Published: 29 August 2022

Citation:
Galvan LPC, Campo CC, Stanojevic S and Alzate DV (2022) Evidence of the Middle-Income Trap in Latin American Countries: Factor Analysis Approach Using Regression and the ARDL Model. Front. Environ. Sci. 10:937405. doi: 10.3389/fenvs.2022.937405

The middle-income trap (MIT) is often accompanied by the decline or stagnation of economic growth, unreasonable domestic industrial structure, and serious polarization between the rich and the poor. However, due to different international environments, different specific national conditions, and different development policies adopted by each country, how to get out of the MIT varies. This study carries out an analysis of different economic growth factors of Latin American countries (we selected 19 MIT countries out of 33) and compared them with Singapore and Korea, which are in a high-income range. We used a regression model to find the relationship of variables in each country and the impact on the economic growth due to these variables. The study finds using correlation and regression analysis, that trade and foreign direct investment (FDI) play a major role in avoiding the MIT by having a strong regression (R^2 = 1.481*** for S. Korea, R^2 = 0.65 for Singapore) with the gross domestic product (GDP) for high-income countries while having a weak regression in Latin American countries. Another factor is that industrialization and services export play a vital role in avoiding the MIT in Singapore and South Korea, and the same model should be used in Latin American countries to avoid the MIT. Furthermore, using the panel ARDL model we validated the results of a regression model and established that similar factors are impacting Latin American countries' MIT. Correlation analysis is used to determine the relationship of selected factors and their impacting strength on the growth of an economy. In the final section, we present Latin American countries, and their main policy gaps according to their unique characteristics and recommend a policy with suggestions for avoiding the MIT by comparing their economies with those of high-income countries.

Keywords: middle-income trap, Latin America, ARDL model, regression, factor analysis

1 INTRODUCTION

The concept of the "middle-income trap" (MIT) has gradually become familiar to development agencies and policymakers. MIT refers to a state in which the per capita income level of a country (region) cannot change smoothly after reaching the middle-income level, resulting in economic stagnation (Yavuz Tiftikçigil et al., 2018). Middle-income countries caught in the trap lose their competitive edge with low-income countries in industries that require a lot of cheap labor, can't compete with high-income countries in R&D-intensive industries, and lose their economic growth momentum (Dui, 2020). A country must avoid or escape this trap, otherwise, it cannot become a fully developed economy. Among the 101 middle-income economies in 1960, only 15 entered the high-income economy by 2014, while other countries or regions remained in the middle-income stage, and some even returned to the low-income stage (Cherif and Hasanov, 2019).

Many economies can easily grow from a low-income country to middle-income country, but it is difficult to cross the middle-income stage and become a high-income country. Latin American countries such as Brazil, Argentina, and Mexico jumped from low-income countries to middle-income countries in a very short period, creating a miracle of growth in the national economy and per capita income, but they experienced economic regressions one after another between 1970 and 1980 (Silva, 2018). Among Asian economies, only Japan has successfully overcome the "middle-income trap" and achieved a leap in economic development. After being hit by the Asian financial turmoil, Indonesia, the Philippines, India, and other countries have remained in the middle-income ranks because they could not resume their previous prosperous economic development (Paus, 2018). MIT is a common economic phenomenon that occurs in countries with different historical, cultural, and economic backgrounds, accompanied by declining economic growth and fragile financial systems. To help countries avoid falling into or escape from the MIT many research institutions and scholars have studied its development theories, causes, and avoidance mechanisms (Kotarski and Petak, 2019). Keep in view the significance of MIT the countries should focus on different economic parameters.

A similar issue of MIT has been observed in Latin America due to a low focus on growth factors. Latin American countries were all historically colonies of European countries, and it was not until the 19th and 20th centuries that independent governments were gradually established. Therefore, the economic development of these countries started relatively late. Despite a late start, Latin American countries have their economic advantages. In the process of economic construction, Latin American countries can rely on the advantages of low labor costs and abundant natural resources to gain benefits in the international market. Relying on unique natural conditions and cheap labor, many Latin American countries have achieved excellent economic results in the export of agricultural products, mineral mining, machinery industry, etc., and completed the process from poverty and adverse conditions to poverty alleviation and prosperity (Ortiz et al., 2018). However, as the nation's wealth grows, so does the labor income, and increasingly, machines could replace human labor. Thus, the labor advantage of Latin American countries is gradually disappearing which is a major cause of MIT in Latin American countries.

Several studies highlight different causes of MIT in Latin American countries. Failure to develop a high-quality labor force is one of the reasons why Latin American economies are gradually losing their edge. To improve the quality of the labor force, it is necessary to strengthen investment in education and scientific research and encourage high technology. Nevertheless, following the economic development of Latin American countries, little attention has been paid to education. The illiteracy rate remains high, and ordinary people who wish to go to school cannot afford tuition fees. Conversely, child labor is rampant in Latin America (Máttar, 2019). Many people cannot afford to send their children to school, so their only option is to let their children go out to work early. In addition, economic growth leads to a widening gap between the rich and the poor. The rich occupy most of society's property and monopolize a large number of social resources and relationships. They can quickly realize "money begets money" through investment and other means. The poor can only earn income by selling their labor, which remains constant. Therefore, if matters continue in such a way, the gap between the rich and the poor will widen.

Many Latin American countries have managed to improve the national economy but have failed to narrow the gap between the rich and the poor and provide social welfare. Latin America, where child labor is rampant, has never had strict labor laws, and ordinary people generally have to pay for services such as education, pensions, and public facilities. The widening gap between the rich and the poor can easily create a social crisis. Under such social conditions, left-wing governments representing the interests of the middle and lower classes usually gain more support and replace the original government. The left-wing government may not perform well in driving economic growth, and it may gradually lose the trust of voters.

The MIT is not only an issue for Latin American countries; it is also a great challenge worldwide. **Figure 1** shows the change of the MIT in the last few years, and it highlights higher-income countries decreasing in the MIT and middle-income countries sustaining the MIT together with similar and a consistent number of countries. A country must avoid or escape the MIT, otherwise, it will be unable to become a fully developed economy. Among 101 middle-income economies in 1960, only 15 had entered a high-income economy by 2014, while other countries or regions remained in the middle-income stage, and some even returned to the low-income stage (Rowe et al., 2018). Many economies can easily grow from a low-income countries to middle-income country; however, it is difficult to bridge the middle-income stage and become a high-income country (Canuto, 2019). Latin American countries such as Brazil, Argentina, and Mexico escalated from low-income countries to middle-income countries in a very short period, creating a miraculous growth in the national economy and per capita income; nevertheless, they experienced economic regressions continually between 1970 and 1980 (Lee and Kim, 2018). Among Asian economies, only Japan has successfully overcome the "middle-income trap" and achieved a leap in economic development (Hartwell, 2018). After being hit by the Asian financial turmoil in 1997, Indonesia, the Philip-pines, India, and other countries have remained in the middle-income ranks as they could not resume their previous prosperous economic development (Estrada et al., 2018).

Different researchers around the globe highlight these MIT issues in different studies with causes and avoidance mechanisms. Oreiro et al. (2020) proposed the development of a list model that highlights how natural resources and an external savings growth strategy may

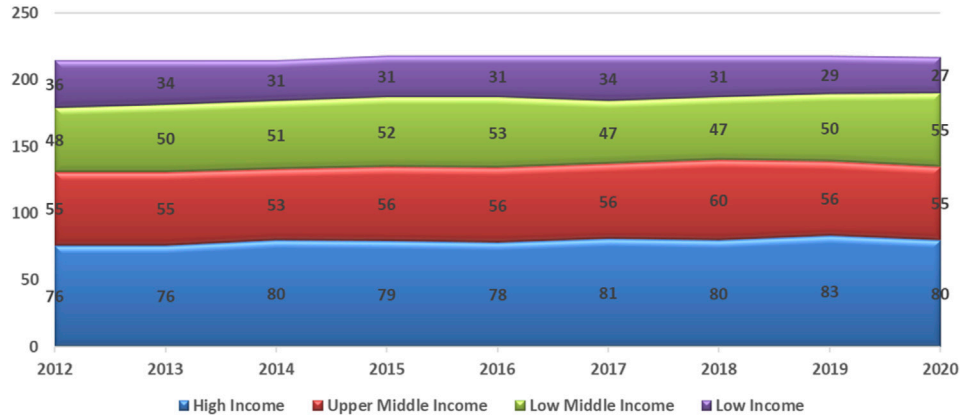


FIGURE 1 | Change of income level in all countries from 2012 to 2020.

lead to an exchange rate overvaluation. Lebdioui et al. (2021) find how Chile and Malaysia escaped from the MIT by focusing on foreign direct investment (FDI), trade (petroleum, rubber, etc.), and research and development (R&D). Justine et al. highlight that industrialization and increasing production have a huge impact on escaping from the MIT in different countries (Lin and Wang, 2020). Vidra et al. focused on the analysis of the impact of science and technology (S&T) development in high-income countries and shows that S&T growth is one of the major components that can help in escaping from the MIT (Klingler-Vidra and Wade, 2020). The MIT is a ubiquitous economic phenomenon and even if not all countries or regions fall into the MIT, they will always be affected by the MIT to varying degrees in the process of their economic development. In addition to the MIT occurring in developing countries, similar phenomena have also occurred in some developed countries, such as the United States and the United Kingdom, in the early stages of development (Ozturk, 2016).

The fundamental reason for the occurrence of the MIT is the mechanism that supported economic growth in the past was unsustainable. Eichengreen et al. (2013) show that if a country wishes to make a breakthrough in economic development and enter the ranks of high-income countries, it needs to change its stagnant or even retrogressive economic state, adjust the backward growth mechanism, and inject new impetus into economic development. From the perspective of international trade, Han and Wei (2017) highlight to avoid MIT the countries should focus on rapid social and economic development, middle-income countries have ushered in development opportunities. Once the economy grows to a certain level, the advantages of labor no longer exist, and labor remuneration continues to increase; however, the economic structure has not been upgraded, and international competitiveness has been further weakened. Therefore, the international trade situation has not only failed to improve year-on-year, but has deteriorated, and the industrial structure has not been improved. To optimize and adjust, economic development has fallen into difficulties, which has caused social turmoil and intensified social contradictions, and the economic development of middle-income countries has fallen into difficulties. Ohno (2009) suggest that foreign trade is the best way to connect the internal economy with the external economy. Improving the foreign trade structure can help solve the problems encountered in the process of industrial structure optimization.

Different researchers used different statistical methods for relationship extraction, e., Irfan et al. (2022) used AHP and G-TOPSIS approach to finding energy barriers. Tang et al. (2022) used the ARDL model for finding the relationship between natural resources and financial growth. Xie et al. (2022) used a frequency-domain approach to get the relationship between economic performance on forest resources. Irfan et al. (2022) in another work investigate the impact of trilemma energy by using correlated panel corrected standard errors (PCSEs) (Khan et al., 2022a). Other studies (Muhammad and Khan, 2021; Khan et al., 2022b; Shahzad et al., 2022; Zhang et al., 2022) also highlight the statistical approaches to get the relationship between different variables. Therefore, this study used regression, correlation, and panel ARDL models to find the relationship between variables of MIT. Bhatti et al. (2022), Bhatti et al. (2021) used regression and correlation models to extract the relationship between variables. Therefore, this study used Correlation, regression and panel ARDL model to extract the relationship.

In the context of the globalized economy, it is of practical significance to actively turn the perspective to the external economy and explore ways to optimize the industrial structure through trade structure optimization. Keeping in view the significance of MIT and how to escape that in Latin American countries, this study's main contributions and objectives are:

- This study attempts to study the factors that affect countries falling into the MIT from the perspective of trade, health, education, and investment (FDI), and puts forward policy recommendations based on the research conclusions to help the country's economy achieve sustainable growth.
- This study finds the influence mechanism of different economic factors such as health, FDI, education, and trade structure on the income level of trapped (Latin American) countries and non-trapped countries

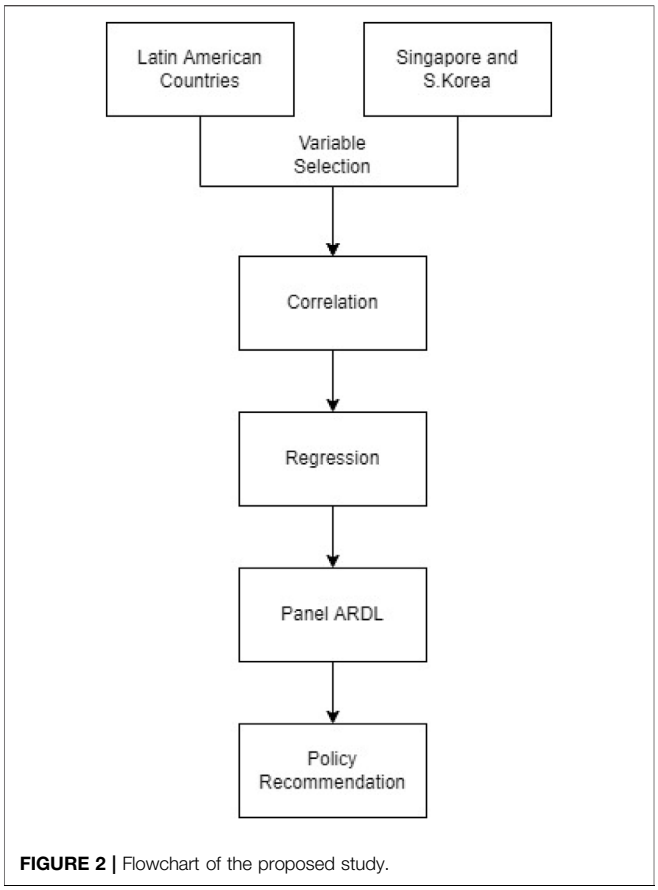


FIGURE 2 | Flowchart of the proposed study.

(Singapore and Korea). This study helps affected countries to optimize trade structure, transform industrial structure, adjust economic development mode, and achieve sustainable growth.
- Based on different statistical analyses like regression, correlation, and autoregressive models this study compares the high-income countries such as Singapore and Korea with MIT countries of Latin America.

1.1 Research Questions

This research addresses the sustainability and current economic performance of Latin American countries by highlighting the indicators that need to be focused on for avoiding the MIT. The following research gaps were identified, which are lacking in other studies, and our study kept a focus on these gaps:

- What are the basic indicators for economic development that need to be focused on by Latin American countries to escape the MIT?
- Which indicators of high-income countries are resulting in a better impact for avoiding the MIT?
- What is the economic development progress in Singapore and Korea towards sustainability in the MIT?
- How do correlation, regression, and the Auto Regressive Distributed Lag (ARDL) model help to determine the most impacting factors among all Latin American countries?
- What are our policy recommendations for governments, stakeholders, business owners, etc.?

TABLE 1 | List of selected countries with income level and region.

Country name	Income level	Region
Argentina	Upper middle income	Latin America
Belize	Lower middle income	
Bolivia	Lower middle income	
Brazil	Upper middle income	
Colombia	Upper middle income	
Costa Rica	Upper middle income	
Dominica	Upper middle income	
Dominican Republic	Upper middle income	
Ecuador	Upper middle income	
Grenada	Upper middle income	
Guatemala	Upper middle income	
Guyana	Upper middle income	
Honduras	Lower middle income	
Jamaica	Upper middle income	
St. Lucia	Upper middle income	
Mexico	Upper middle income	
Nicaragua	Lower middle income	
Panama	Upper middle income	
Peru	Upper middle income	
Paraguay	Upper middle income	
El Salvador	Lower middle income	
Suriname	Upper middle income	
Venezuela, RB	Lower middle income	
Korea, Rep	High income	
Singapore	High income	Southeast Asia

2 MATERIALS AND METHODS

Complete flowchart of this study implementation as shown in **Figure 2**:

2.1 Study Area

Latin America refers to Central America, the Caribbean, and South America, all south of the United States. It is named for the fact that most countries use Spanish and Portuguese, both of which belong to the Latin family, as their national languages (Heinicke et al., 2007). Located in the central and southern parts of the Western Hemisphere, it is bordered by the Atlantic Ocean in the east, the Pacific Ocean in the west, bordering North America (United States) in the northwest, and facing Antarctica across the Drake Passage in the south. There are a total of 33 countries in Latin America, out of which we selected a total of 19 countries and categorized them according to their income level i.e., upper middle-income and lower middle-income (Goldie et al., 2008). We selected two higher income countries, Singapore, and Korea, as models for comparison as they had escaped the MIT. **Table 1** shows the list of the selected countries. **Figure 3** shows the geographic location of countries.

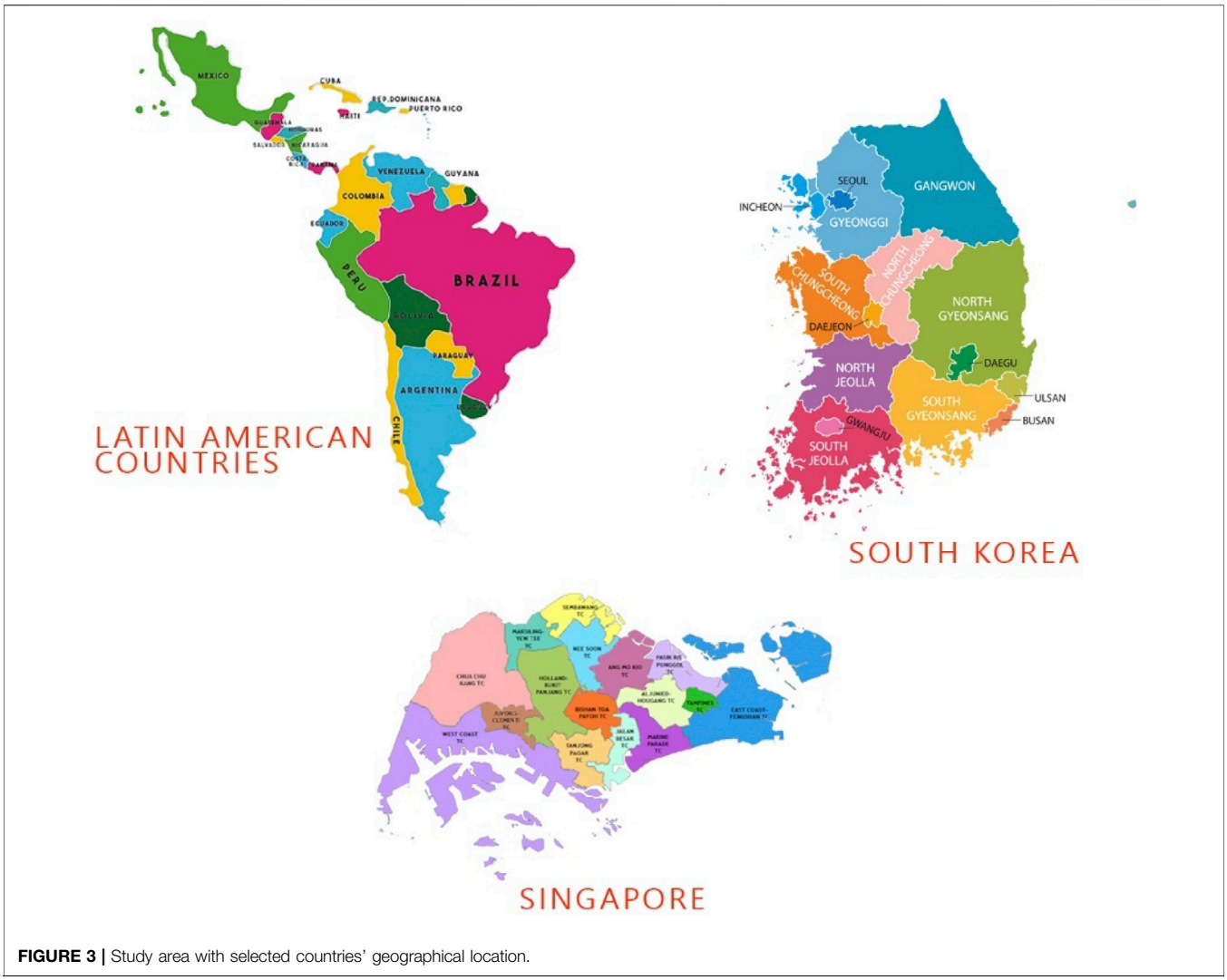


FIGURE 3 | Study area with selected countries' geographical location.

2.2 Data Selection

Data for different factors has been selected from the World Bank website for the period from 2000 to 2020 for the selected countries (World Bank). The focused factors that impact MIT are education, health, GDP, travel services/tourism, trade [high-technology exports, information and communication technology (ICT) exports etc.], and foreign direct investment (FDI). Descriptive analysis of the data is shown in **Table 2**. Statistical analysis of the data was performed using SPSS software (version 25; IBM).

2.3 Regression Model

A regression model is a method of statistical analysis that studies the dependence of a dependent variable on an independent variable in regression, with the aim of estimating or predicting the mean of the dependent variable from a given value of the independent variable. It can be used for forecasting, modeling time series, and discovering causal relationships between various variables (Florax and de Graaff, 2004). This study utilized the regression model to predict the impact of variables on the growth of economic development. The benefits of using regression analysis are as follows:

1) It indicates a significant relationship between the independent variable and the dependent variable.
2) It indicates the strength of the influence of multiple independent variables on a dependent variable.

Regression analysis can also be used to compare the interaction between variables measured by different measures, such as the link between price changes and the number of promotions. These benefits allow market researchers, data analysts, and data scientists to exclude and measure the best set of variables for building predictive models.

The simplest regression model can be represented as the data object to be fitted is $X = \{x_1, x_2, \ldots, x_m\}$, the corresponding real value is $Y = \{y_1, y_2, \ldots, y_m\}$, the linear model can be written as:

$$\hat{y} = Xw \tag{1}$$

TABLE 2 | Descriptive statistical data.

Countries	Descriptive statistics	Education (US$)	Health expenditure	GDP (current US$)	Travel services	High-technology exports (US$)	Service exports (US$)	FDI (US$)
Latin American Countries	Mean	93.13	6.16	1.71E + 11	8.34	43891276	47534758	−5485593
	Median	95.76	5.85	2.23E + 10	4.19	0.44	542.41	95.05
	Standard Deviation	7.08	1.43	4.03E + 11	17.05	25708751	2.34E + 09	2.09E + 09
	Minimum	60.89	3.06	3.01E + 10	−3.09	0	145.81	−1.4E + 10
	Maximum	100	10.44	33.01E + 10	254.95	2.015E + 09	1.55E + 10	100
Singapore	Mean	94.26	3.52	2.302E + 11	1.49	0.32	2573.43	88.15
	Median	95.67	3.35	2.4E + 11	0.96	0.32	2321.71	87.8
	Standard Deviation	3.64	0.48	1.025E + 11	1.92	0.05	950.44	3.08
	Minimum	87.75	2.84	8.979E + 10	−0.53	0.24	1384.79	83.37
	Maximum	97.84	4.39	3.76E + 11	6.63	0.43	4102.27	93.78
Korea	Mean	85.93	5.72	1.168E + 12	2.33	2.94	1851.67	82.25
	Standard Error	0.17	0.27	8.237E + 10	0.27	0.19	185.6	0.19
	Median	85.93	5.85	1.173E + 12	2.26	2.93	1792.55	82.25
	Standard Deviation	0.24	1.21	3.775E + 11	1.22	0.84	830.02	0.26
	Maximum	86.1	8.16	1.725E + 12	4.67	4.21	3521.33	82.44

Where w is the regression coefficient, we use a square error to measure the fitting error:

$$L(X) = \sum_{i=1}^{m} \left(y_i - x_i^T w\right)^2 = \left(y - Xw\right)^2 \tag{2}$$

The above formula is equal to 0 to the w.

$$\frac{\partial L(X)}{\partial w} = \frac{\partial \left(y^T y - w^T X^T y - y^T X w - w^T X^T X w\right)}{\partial w} = 2X^T \left(y - Xw\right) = 0 \tag{3}$$

It can determine:

$$\hat{w} = \left(X^T X\right)^{-1} X^T y \tag{4}$$

The above is easy to interact with training data; a good solution is partial weighted linear regression, increasing a weight w_i for each error (here w is not the above $\hat{w}$), at this time, the error function can be written:

$$L(X) = \sum_{i=1}^{m} w_i \left(y_i - x_i^T w\right)^2 = \left[W\left(y - Xw\right)\right]^2 \tag{5}$$

Among them, W is a diagonal matrix, also called the core; the type of core can choose freely, and the most common is the Gaussian nucleus. The weight corresponding to the Gaussian nucleus is as follows:

$$W(j, j) = \exp\left(\frac{\left\|x^i - x^j\right\|^2}{-2k^2}\right) \tag{6}$$

Similarly, the new error function $L\ (X)$ is governed to obtain the regression coefficient at this time:

$$\hat{w} = \left(X^T W X\right)^{-1} X^T W y \tag{7}$$

The W here is actually $W^T W$, but using W replaces the same meaning and is simple.

2.4 Granger Causality Test

This analysis is to test the causality of the time series data, and this model was established by Granger (1969) (Engle and Granger, 1987). A variable x_t is caused by y_t, incase if it can forecast the x_t with greater number of accuracy with the usage of y_t past value. The first step of this test is the calculation by the VAR model and given as follows:

$$y_t = \alpha_1 + \sum_{i=1}^{n} \beta_i x_{t-i} + \sum_{j=1}^{m} \beta_j y_{t-j} + e_{1t} \tag{8}$$

$$x_t = \alpha_2 + \sum_{i=1}^{n} \delta_i x_{t-i} + \sum_{j=1}^{m} \delta_j y_{t-j} + e_{2t} \tag{9}$$

Where e_{1t} and e_{2t} are known as the uncorrelated white noise error expressions. Here null hypothesis shows that x_t is not caused by y_t. If there is a situation that value of F computed > F critical value, thenceforth, the null hypothesis rejected, and it is finalized that x_t is caused by y_t.

2.5 Panel Auto Regressive Distributed Lag Model

ARDL model has been in use for decades to model the relationship between economic variables in a single-equation time series setup. Its popularity also stems from the fact that cointegration of nonstationary variables is equivalent to an error correction (EC) process, and the ARDL model has a reparameterization in EC form (Hassler and Wolters, 2006). The existence of a long-run/cointegrated relationship can be tested based on the EC representation. A bounds testing procedure is available to draw conclusive inference without knowing whether the variables are integrated of order zero or one, I (0) or I (1), respectively (Pesaran et al., 2001).

The regression model studies the analysis of univariate series, while the ARDL model, models multivariate time series. In the ARDL model, there is not only the lag part of the original data,

but also other influencing factors are added to adjust the autoregressive results. The method is simple and effective for small samples or limited sample data. Its general expression is:

$$B(L,P)y_t = \sum_{i=1}^{k}\beta_i(B,q_i)x_{it} + u_t \tag{10}$$

Among,

$$B(L,P) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p \tag{11}$$

$$\beta_i(B,q_i) = 1 - \beta_{i1}B - \beta_{i2}B^2 - \ldots - \beta_{iq_i}B^{q_i} \tag{12}$$

In **Eqs 11**, **12**, P represents the delay order of the explanatory variable y_t, q_i represents the delay order of the *i-th* explanatory variable x_{it}, $i = 1, 2, \ldots, k$. B is the delay operator, u_t is the random item. ARDL models are generally written as:

$$y_t = \sum_{i=1}^{p}\varphi_i y_{t-i} + \sum_{j=1}^{k}\sum_{l_j=0}^{q_j}\beta_{jl_j}x_j, t - l_j + u_t \tag{13}$$

The ARDL model is widely used for a multivariate time series model, and it requires less data; it is a simple idea, and it is a good model construction method.

Traditional panel data analysis assumes that the relationship between Y and X is homogeneous for all individuals, and the heterogeneity is mainly reflected in the intercept term. When we examine the dynamic panel data model, the variable intercept model can be written as:

$$Y(it) = aY(it-1) + bX(it) + c(i) + u(it), \tag{14}$$

where i represents the i-th cross-sectional unit, t represents the observation value of the t-th period, and $c(i)$ is the intercept term that changes with individual i. The estimation method of model parameters a and b is to use the system GMM or difference GMM method introduced in the last tweet. In this article, Pesaran et al. (1999) assumed that all coefficients may vary with individual i, so not only the intercept but also the slope, such as:

$$Y(it) = a(i)Y(it-1) + b(i)X(it) + c(i) + u(it), \tag{15}$$

The reason why this heterogeneous panel model is considered is that in macroeconomic research, unobservable factors such as the institutional culture of each region or country often show systematic differences, and these differences will not only affect the intercept term $c(i)$, may also affect the sensitivity of Y to changes in X.

Panel Auto Regressive Distributed Lag (ARDL) model or Pooled Mean Group (PMG) technique for analyzing non-stationary dynamic panels was given by Pesaran and Shin (1995), Pesaran et al. (1999). PMG has both averaging and pooling, hence it is considered as an intermediate estimator between Dynamic Fixed Effect (DFE) and Mean Group (MG). PMG has an advantage over the Dynamic OLS model that it allows the short run dynamic specification to differ among cross section, while the long run coefficients are constrained to be the same. Hence, panel ARDL/PMG model is used to investigate the heterogeneous dynamic issue across cross sections as well as to estimate the long and short run relationship among variables. So, panel ARDL/PMG model can be specified as:

$$Y_{it} = \sum_{j=1}^{p}\lambda_{ij}Y_{i,t-j} + \sum_{j=0}^{q}\delta'_{ij}X_{i,t-j} + \mu_i + \varepsilon_{it} \tag{16}$$

Where, Y_{it} reports dependent variable, X_{it} represents (k × 1) vector of explanatory variables, μ_i shows the fixed effects, λ_{ij} shows the coefficient of the lagged dependent variable, δ_{ij} represents (k × 1) coefficient vector of independent variables, ε_{it} denotes the error tern, i (1, 2, . . ., N) is number of cross section, and t (1, 2, . . ., T) is number of time.

3 RESULTS AND DISCUSSION

A generally accepted view as to the cause of the MIT is: that with economic growth, labor costs increase, and the comparative advantage of cheap labor costs is lost; a new economic growth model guided by knowledge and innovation has not yet been formed, thus making economic growth momentum insufficient. Therefore, the causes of the MIT can be roughly divided into the following three levels: first, the direct impact variables of sustained economic growth, such as health (Paus, 2014), education, industrial structure (Kanchoochat, 2014) and its changes; second, the internal and external environmental variables affecting economic growth, such as trade (Engel and Taglioni, 2017), macroeconomic policies, demographic factors (Jayasooriya, 2017), and FDI (Nguyen-Huu and Pham, 2021); and the third is the fundamental reason that affects long-term economic performance, such as social and economic systems, for example, tourism and services. In view of this, education, health expenditure, travel services, high-technology exports, service exports, and FDI are selected as independent variables affecting a country's per capita GDP, and economic growth is represented by per capita GDP (y) as a dependent variable.

According to the above analysis, with *ln(free)*, *ln(open)*, *ln(con)*, *ln(ind)*, *ln(hon)*, *ln(tec)*, *inf*, *ln(lab)*, *ln(inv)* and *ln(ci)* as the independent variables and *ln(y)* as the dependent variable, establish the following panel data econometric model:

$$\begin{aligned} In(y_{it}) = {} & c_0 + \alpha_i + c_1 ln(tec_{it}) + c_2 ln(free_{it}) + c_3 inf_{it} \\ & + c_4 ln(open_{it}) + c_5 ln(hea_{it}) + c_6 ln(edu_{it}) \\ & + c_7 ln(inv_{it}) + c_8 ln(ind_{it}) + c_9 ln(hon_{it}) \end{aligned} \tag{17}$$

where i represents the countries' element, i = 1, 2, in the "trap group" country model. . . , 14 (here Singapore and Korea are also considered for evaluation); i = 1, 2, . . . , 26, t represents time, c_0 is the intercept term, and α_i is the difference intercept term. $In(y_{it})$ is the logarithm of a country's per capita GDP, $ln(tec_{it})$ is the logarithm of a country's technological level, $ln(free_{it})$ is the logarithm of economic freedom, and inf_{it} is the logarithm of the inflation rate. $ln(open_{it})$ represents the logarithm of the proportion of trade volume in GDP, $ln(hea_{it})$ represents the logarithm of health, $ln(edu_{it})$ represents the logarithm of the proportion of education in GDP, $ln(inv_{it})$ represents the logarithm of the proportion of FDI in GDP, $ln(ind_{it})$ represents the logarithm of the proportion of secondary industry output in GDP, $ln(ci_{it})$ is the logarithm of the urbanization rate, and ε_{it} is the random error term.

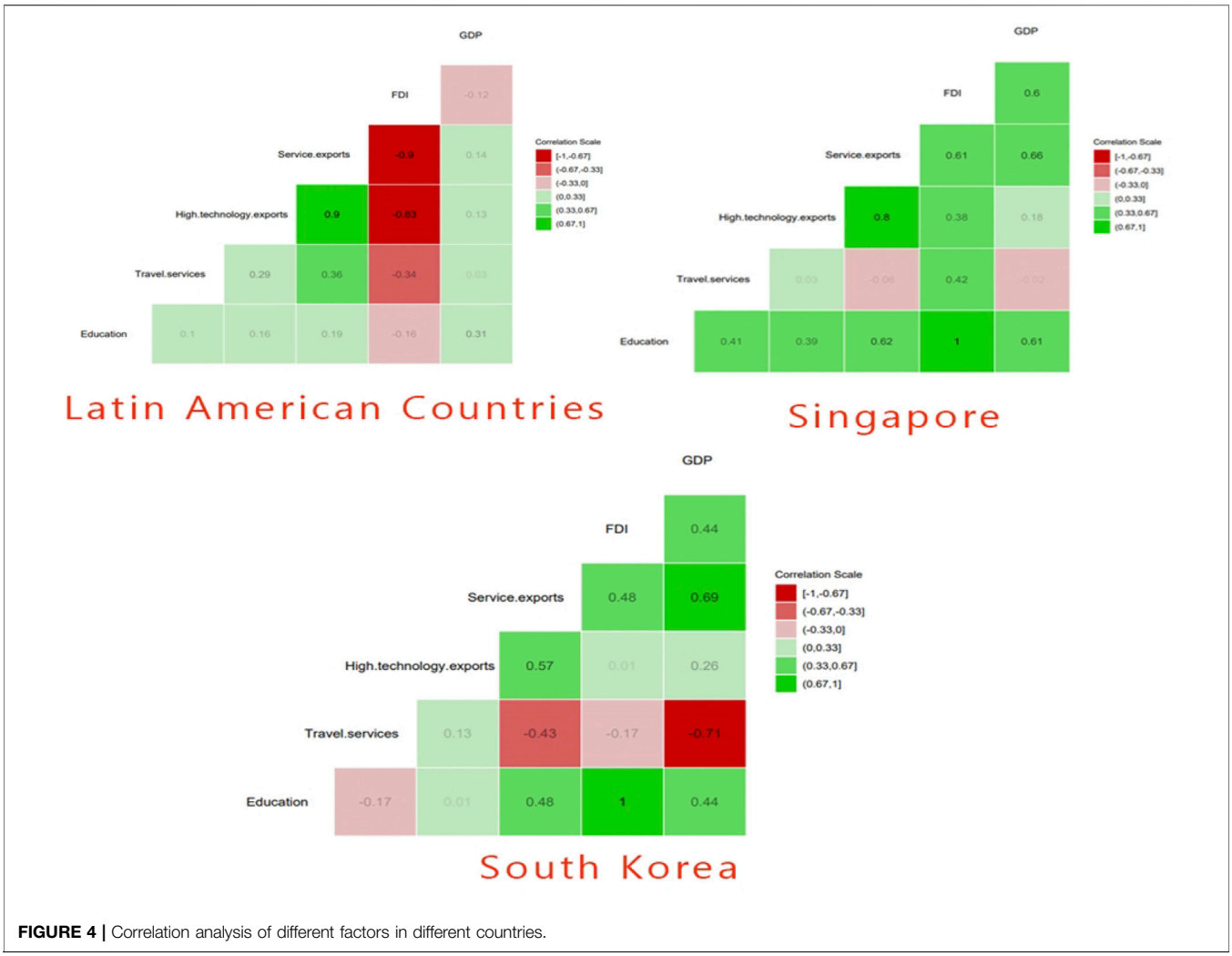


FIGURE 4 | Correlation analysis of different factors in different countries.

The results of the Hausman test found that the fixed effects model was the most suitable for the data in this paper. However, the static panel ignores the dynamic influence of the lag term of the independent variable on itself, which may lead to large deviations in the estimation results. Since the economic growth of the previous period has an impact on the economic growth of the current period, it is necessary to introduce the lag term of economic growth to reflect the dynamic lag effect. On the basis of the static panel model, the first-order lag term of the dependent variable is incorporated into the model to obtain the dynamic panel model as follows:

$$In(y_{it}) = c_0 + \alpha_i + c_1 ln(tec_{it}) + c_2 ln(free_{it}) + c_3 inf_{it} + c_4 ln(open_{it}) + c_5 ln(lab_{it}) + c_6 ln(con_{it}) + c_7 ln(inv_{it}) + c_8 ln(ind_{it}) + c_9 ln(hon_{it}) + c_{10} ln(ci_{it}) + c_{11} ln(y_{i,t-1}) + \varepsilon_{it} \quad (18)$$

Using the lagged term of the dependent variable as an independent variable will cause endogeneity problems in the regression model. In order to better deal with the correlation and endogeneity between cross-sections, regression analysis was performed on the data. First, by using correlation analysis between different factors highlights that for high income countries such as Singapore the impact of GDP is strongly correlated with FDI (0.6) and education (0.61) while travel/tourism has a weak negative correlation with GDP (−0.02). Exports of services (0.66) and technology (0.8) are highly correlated with GDP, which shows that Singapore's GDP is greatly impacted by trade services. Similarly, results are observed for South Korea where GDP has a strong positive correlation with FDI (0.44); however, services (0.69) and technology (0.57) exports also play a major role in economic development by having a positive correlation. In South Korea, travel/tourism is strongly negatively correlated with GDP, which shows that the impact of the tourism sector needs to be improved to further strengthen the economy. Nevertheless, for Latin American countries, both GDP and FDI (−0.12) have a negative correlation, which is one of the reasons that many of the remaining countries in that region are in the MIT. Another reason is the trade and exports that are weakly correlated with GDP (0.13). These are areas where Latin American countries need to focus on by increasing industrialization. **Figure 4** shows further detailed comparisons of correlation between regions or countries.

TABLE 3 | Regression model for Latin American and other countries.

Country name	Education expenditure	Health expenditure	Travel/Tourism services	High technology exports	Service exports	FDI	R^2
Argentina	−3.864*	1.085	−1.088	−0.353	1.099*	−0.000	0.986***
Belize	0.141	0.468	0.131	−0.020*	−0.226	0.001	0.971***
Bolivia	0.01	−0.702*	−0.031	−0.017	0.511*	0.154	0.998***
Brazil	−0.71	−1.63	−0.147	−0.207	0.357	3.878	0.987***
Colombia	−27.73	0.136	−0.203*	0.042	0.318	28.89	0.991***
Costa Rica	−0.305	−0.253	−0.069*	0.008	0.549*	0.745	0.997***
Dominica	0.154	−0.773***	−0.009	−0.017*	0.909***	0.131	0.990***
Dominican Republic	0.507	−0.661*	−0.047*	0.007	0.686**	-0.657	0.997***
Ecuador	0.997	−1.972	0.038	0.209	2.006	0.412	0.999*
Grenada	0.111	−0.957***	0.003	0.000	1.213***	0.131	0.994***
Guatemala	−0.159	−3.230	0.043	0.001	2.522*	0.645	0.988***
Guyana	1.785	−2.708	0.00	−0.044	10.063	0.230	0.996
Honduras	0.141	−0.748**	−0.161*	0.020	0.387	0.211	0.981***
Jamaica	0.530	0.196	−0.115	−0.005	0.048	−0.236	0.992***
S.Korea, Rep	0.512	−2.281***	−0.057*	0.062	1.481***	0.171	0.990***
St. Lucia	−0.106	−1.621	−0.009	0.015	1.421	−0.494	0.997
Mexico	1.395	0.243	0.012	0.015	−0.856***	−3.404*	0.995***
Nicaragua	0.155	−0.603	−0.034	−0.009	0.815***	0.214	0.994***
Panama	0.154	−0.755*	−0.030*	0.009	0.930***	0.131	0.997***
Paraguay	−0.558*	−0.143	0.009	0.127	0.594	0.071	0.997***
Singapore	−1.834	−0.345	−0.030	0.650	0.651	−1.452	0.998
Grenada	−0.862	−0.290	−0.032	−0.066	0.477	−2.439	0.999*
Suriname	0.31	−1.554	−0.000	0.028	1.108	0.211	0.838*
Venezuela, RB	0.112	−5.620**	0.141	0.057	5.571*	0.117	0.960***

, **, *, **** represents the Significance level.*

Next step is to apply regression analysis to further check regression results. **Table 3** shows the regression between different models with the significance between their relationship with the GDP of each country. From **Table 3**, it can be seen that health expenditure has a significant negative impact on Korea (β = −2.281***), Grenada (β = −0.957***), Dominica (β = −0.773***), Venezuela (β = −5.620**and Honduras (β = −0.748**). Observing the developing countries as models for escaping the MIT, their main focus is health expenditure, which is not focused on by many Latin American countries. Similarly, education expenditure has a significant impact on Argentina (β = −3.864*) while other countries show a weak negative relationship with GDP thus not having a major impact on economic growth. High technology exports (trade services) have a positive impact on the GDP in the high-income countries, Korea (β = 0.062) and Singapore (β = 0.65), while Latin American countries show a negative relationship with GDP such as Argentina (β = −0.353), Belize (β = −0.020*), Bolivia (β = −0.017), Jamaica (β = −0.005), Dominica (β = −0.017*), Nicaragua (β = −0.009), Grenada (β = −0.066) and Guyana (β = −0.044). Some Latin American countries have a positive impact on the GDP such as Dominican Republic (β = 0.007), Brazil (β = −0.207), Colombia (β = 0.042), Costa Rica (β = 0.008), Ecuador (β = 0.209), Guatemala (β = 0.001), Honduras (β = 0.02), St. Lucia (β = 0.015), Mexico (β = 0.015), Panama (β = 0.009), Paraguay (β = 0.127), Suriname (β = 0.028) and Venezuela, RB (β = 0.057). Services export have a significant positive impact on the GDP in Korea, Rep. (β = 1.481***) while a similar impact is also observed in some Latin American countries such as Mexico (β = −0.856***), Nicaragua (β = 0.815***), Panama (β = 0.930***), Dominica (β = 0.909***), Bolivia (β = 0.511*), Venezuela, RB (β = 5.571*) and Costa Rica (β = 0.549*) while a weak positive impact is observed in other countries. Travel and tourism has a significant negative impact on Korea (β = −0.057*) while a similar impact is also observed in some countries in Latin America such as Colombia (β = −0.203*), Costa Rica (β = −0.069*), Dominican Republic (β = −0.047*), Honduras (β = −0.161*) and Panama (β = −0.030*). However, FDI has a weak negative impact on the GDP in high-income countries Korea, Rep. (β = −1.352) and Singapore (β = −1.452). Most of the Latin American countries show a positive relationship [Brazil (β = 3.878), Colombia (β = 28.89), Jamaica (β = −0.236) and St. Lucia (β = -0.494)], which is contrary to high-income countries.

After getting significant factors from correlation and regression models next step is to validate results by checking with panel ARDL econometric model. **Table 4** shows the results of long-run estimation using PMG, MG and DFE estimators. The results of the Hausman test to choose either PMG or MG show that it is not significant, and thus PMG is better than MG. The results of the Hausman test for PMG and DFE show that PMG is still preferable as the *p*-value is higher than the significance level. The results of the three estimators (PMG, MG and DFE) in **Table 5** and **Table 5** are elaborated more in detail by keeping GDP as the dependent variable. In **Table 5** using DFE, the relationship of travel services, and FDI is negative while other factors show a positive relationship with GDP. Similarly, using PMG travel services (−1.07), services exports (−2.22) and education (−0.108) show a negative relationship while other show a positive relationship. Similarly for MG the results for FDI and Health is negative on GDP while other factors are creating a positive impact on

TABLE 4 | Panel Data Analysis using MG, DFE, and PMG for high-income Countries.

	Panel high-income countries											
	Mean group estimation (MG)				Dynamic fixed effects (DFE)				Pooled mean group estimation (PMG)			
Parameters	**Coef**	**Std. Err**	**Z**	***p*** **>** **z**	**Coef**	**Std. Err**	**Z**	***p*** **>** **z**	**Coef**	**Std. Err**	**z**	***p*** **>** **z**
Travel service	−1.93E + 10	1.93E + 10	−1.00E + 00	3.17E−01	−5.13E+09	1.15E+10	−4.50E−01	6.54E−01	−1.07E+00	4.51E−01	−2.00E−01	8.41E−01
Services exports	−3.55E + 00	2.04E + 00	−1.74E + 00	8.20E−02	2.50E−01	1.59E+00	1.60E−01	8.75E−01	−2.22E+00	2.67E+00	−1.10E−01	9.12E−01
FDI	−6.24E−01	2.21E + 00	−2.80E−01	7.78E−01	−5.81E−01	1.16E+00	−5.00E−01	6.16E−01	1.39E+00	1.57E+00	−1.90E−01	8.49E−01
High technology exports	4.14E + 00	2.18E + 00	1.90E + 00	5.80E−02	9.65E−01	8.93E−01	1.08E+00	2.80E−01	4.26E+00	3.77E+00	−1.80E−01	8.57E−01
Education	1.85E−01	3.71E−01	−3.20E−01	7.46E−01	7.40E−01	2.07E−01	3.58E+00	0.00E+00	−1.08E−01	3.96E−01	1.83E+00	6.70E−02
Health	−3.59E−02	1.11E−01	9.50E−01	3.43E−01	−2.51E−01	1.38E−01	3.08E+00	2.00E−03	6.05E−02	9.48E−02	1.15E+00	2.49E−01
Error correction estimation	2.05E−02	5.43E−02	3.80E−01	7.05E−01	−1.28E−01	1.79E−01	−7.10E−01	4.76E−01	3.99E−02	8.78E−02	−2.00E−01	8.44E−01

TABLE 5 | Panel Data Analysis using MG, DFE, and PMG for Latin American Countries.

	Panel B (Latin American countries)											
	Mean group estimation (MG)				Dynamic fixed effects (DFE)				Pooled mean group estimation (PMG)			
Parameters	**Coef**	**Std. Err**	**Z**	***p*** **>** **z**	**Coef**	**Std. Err**	**z**	***p*** **>** **z**	**Coef**	**Std. Err**	**z**	***p*** **>** **z**
Travel Service	−1.44E + 10	1.44E + 10	−1.00E + 00	3.17E−01	−5.81E + 09	1.07E + 10	−5.40E−01	5.88E−01	−1.07E + 00	4.51E−01	−2.36E + 00	1.80E−02
Services exports	−2.34E + 00	1.88E + 00	−1.24E + 00	2.14E−01	2.20E−01	1.42E + 00	1.50E−01	8.77E−01	−2.22E + 00	2.67E + 00	−8.30E−01	4.07E−01
FDI	−5.51E−01	1.57E + 00	−3.50E−01	7.25E−01	−5.04E−01	9.99E−01	−5.00E−01	6.14E−01	8.42E + 00	9.65E−01	8.90E−01	3.75E−01
High technology exports	3.03E + 00	1.90E + 00	1.60E + 00	1.10E−01	2.12E + 09	6.10E + 08	1.23E + 00	2.20E−01	1.81E−01	7.40E−01	1.13E + 00	2.59E−01
Education	9.53E + 08	9.53E + 08	1.64E + 00	1.02E−01	7.59E−01	1.75E−01	4.34E + 00	0.00E + 00	1.62E−01	−2.51E−01	6.40E−01	5.23E−01
Health	1.64E−01	2.63E−01	1.00E + 00	3.17E−01	7.59E−01	1.75E−01	−2.20E + 00	2.80E−02	1.27E−01	−1.28E−01	1.40E−01	8.86E−01
Error correction estimation	−6.16E−02	9.06E−02	−6.80E−01	4.97E−01	−1.25E−01	1.53E−01	−8.20E−01	4.15E−01	3.99E−02	8.78E−02	4.50E−01	6.50E−01

GDP. For Latin American countries using DFE, travel services and FDI producing the negative impact while other factors are having positive impact on GDP, For PMG model results for FDI are positive for GDP, while travel services and other factors are creating negative relationship. Using MG model, the travel services, services exports, FDI are having negative impact on GDP.

Our main findings can be summarised as follows: 1) there is a negative effect of the public debt ratio on economic growth, both in the short-run and long-run, 2) the negative relationship is more significant when we use common correlated factors to address the issue of cross-sectional dependence, 3) an asymmetric response of a change in public debt is found to be significantly negative in the short-run. As such, rises in short-run public debt negatively affect economic growth in the short-run but falls public debt do not have a correspondingly positive effect on economic growth in the short-run.

Kuchiki et al. (2017) highlights the same factors as our research, i.e., tourism and service exports are important components in avoiding the MIT and growing industrialization. Developing countries purchase relatively modern technical equipment from developed countries and introduce comparatively advanced management concepts to develop their somewhat backward industries, improve capital accumulation, and improve technical efficiency. However, once a country develops to a certain stage, new difficulties will appear, such as the establishment of technical barriers by advanced countries, the formulation of trade barriers, the implementation of intellectual property protection, and the change of world trade rules and regulations. Sim and Ali (1998) agree with our study, which shows that trade is important and export-oriented economies can improve production technology by imitating foreign products in the early stages. Therefore, the improvement of the R&D capabilities of enterprises and the core competitiveness of products will help them in a smooth transition. Judging from indicators such as the proportion of R&D investment in each country and the number of scientific researchers, countries such as Japan, South Korea, and Singapore have successfully broken through the MIT by having higher economic and social benefits from R&D investment, while Brazil, Mexico, and other Latin American countries are trapped due to the contribution of low scientific research for economic development within the MIT countries (Lall, 2000). The study of Dahlman et al. (1987) shows that FDI and industrialization for labor growth plays an important role in a country's development. Therefore, these countries stay in a low-end position for a long time when participating in global trade and are unable to achieve sustained and stable economic growth across the middle-income quagmire. Due to the lack of core competitiveness, the ability of low- and middle-income countries to take risks is relatively low. In the face of external risks such as short-term economic downturn, technological monopoly, and the rewriting of trade rules in developed countries, the domestic economy is turbulent, and accumulated political and social conflicts erupt (Austin, 2002). On the other hand, our results compare with Mahul et al., who highlight that the labor force and new local products are important factors for avoiding the MIT. The MIT shows that countries that rely on low value-added and high-polluting low-end products have the risk of an internal industrial structure transformation, and the external risk of increasing the division of labor in the global value chain (Mahul and Stutley, 2010).

Some countries in Latin America and Africa as well as other countries caught in the MIT have a relatively single export structure, a high proportion of low value-added primary product exports, and limited export targets (Clarke et al., 2003) (Agosin and Bravo-Ortega, 2009). Our study highlights the similar fact that the trade component is avoided in these countries. The uncertainty of the world economy and the economic instability of exporting countries will cause price fluctuations of the export products of these countries. During the downturn of the world economy, since there is no bargaining power, the foreign exchange earnings of low- and middle-income countries will be greatly weakened, and the fragile domestic economic system will definitely be implicated (Sabel et al., 2012). Such a vicious circle will cause these countries to be unable to upgrade their industrial structures and to be at the lower end of the globalized division of the labor system, thus preventing them from entering the ranks of high-income countries. For developing countries, foreign trade is an important factor in promoting economic growth. Basically, it is necessary to develop foreign trade with comparative advantages based on the national conditions of the country (Fahim et al., 2021). It is necessary to speed up the transformation of the industrial structure, vigorously develop tertiary industry, and accelerate the change of industrial focus. In the process of developing tertiary industry, the importance of secondary industry cannot be ignored. Secondary industry is still an important sector for creating social wealth and stabilizing economic growth (Besedes and Blyde, 2010).

3.1 Policy Recommendations

This study is significant for middle-income countries by focusing on different factors which can help in avoiding the MIT. A few suggestions and policy recommendations for stockholders and government are:

3.1.1 Strengthen Exports

This can be done by managing capital goods and attaching importance to the export of high-tech products. The content of capital goods is great, and its export price plays an important role in regulating the development of a country's foreign trade. It not only intuitively reflects the laws of internal industrial production, but also acutely reflects changes in the external economy. There-fore, only by strengthening reform of the capital goods export management system can economic and technological value be exerted more effectively, and greater wealth can be created. In addition, capital goods with high technology content are even more critical in current international competition. The development of economic globalization has promoted the development of the diversification of global trade. There are increasingly more trade exchanges between countries, which promote the development of world trade towards a virtuous circle and accelerate the flow of world factor resources. The research and development, production and sales of high-tech products has become a worldwide industry. High-tech products, as a means of further strengthening the competition of a country's foreign trade in capital goods, are rising in status in today's international trade. In the current world situation, developing countries such as China, which are on the disadvantaged side of

scientific and technological resources, while exerting their comparative advantages of the "catch-up effect", absorb the advanced scientific research achievements of developed countries, improve scientific literacy, increase production skills, and cultivate independent innovation capabilities, using the limited resources of the country to harvest greater output and optimize the industrial structure.

3.1.2 Consumer Goods and Industrialization

In the world market, highly competitive industrial products are mainly concentrated in labor-intensive items such as textiles, toys, and consumer goods. These products have low added value and short product chains, which are not conducive to China's competitive advantage in the international market. The export growth of consumer electronics and electrical appliances in recent years shows that China is making great efforts to improve the added value and foreign exchange earning capacity of traded products, and to participate more actively in the division of labor in the global value chain. The current global trade pattern is undergoing profound changes; on the one hand, European and American countries have implemented the strategy of "industrialization" to seize the commanding heights of international trade; on the other hand, emerging economies such as India and Vietnam have used their comparative ad-vantages in labor prices to accelerate the process of industrialization. From the perspective of long-term development, China's export of consumer goods must find a new development path, reshape new advantages in international competition, realize industrial upgrading, and become a "manufacturing power."

3.1.3 Foreign Direct Investment Involvement for Quality Products

The employment population and trade activities involved in processing trade are highly considerable. Therefore, while stabilizing the advantages of traditional processing trade, countries should enhance the innovation ability of processing trade, cultivate new advantages in processing trade, change the original pattern of relying solely on abundant labor to obtain comparative advantages, allowing enterprises to better participate in the global division of labor, and improving the awareness of independent R&D. Foreign direct investment can help in improving the quality of labor by R&D and continuous training in the latest methods. The production of high-end industrial products is of great significance to the promotion of China's trade status and the realization of sustainable economic development.

4 CONCLUSION

This study compares and analyzes different criteria for determining the MIT in Latin American countries and provides possible reasons for the MIT. Our research finds that the social and economic development of Latin American MIT countries contains some com-mon characteristics, and the findings help us understand the mechanism of the "middle-income trap." In South Korea and Singapore, through an analysis of the proportion of trade exports of various types of export commodities, a reasonable explanation for the difference in the export commodity structure between "middle-income trap" and non-"middle-income trap" countries is obtained. For non-"middle-income trap" countries, the industrial upgrading process seems to be consistent with the flying geese pattern of development. It follows that industrial upgrading through backward linkages between consumer goods and capital goods is more successful in high-income countries that are not in MITs. For countries that are in MITs, there is a tendency to rely on exports of primary products, while industrialization is driven by forward linkages of finished products. A weak industrial base is a possible factor in the MIT. These analyses can describe the MIT. The policy implications of the findings of this study are very direct. It is necessary to develop the consumer goods industry and maintain competitiveness in promoting the up-grading of industries to capital goods through backward linkages. Although there are many problems in the economic development of Latin America, these difficulties can be improved through economic and social policies. Based on the reasons analyzed above, Latin American governments can implement the following policies:

- Raise the national education level and train workers with advanced technology.
- Increase scientific research funding and encourage R&D innovation.
- Promote industrial upgrading.
- Create a favourable environment for foreign investment and encourage overseas in-vestment to stimulate the economy.
- Improve the role of taxation; allow taxation to facilitate the reduction of the gap be-tween the rich and the poor and increased social welfare.

Future work of this study is to increase the variables and find in more dept relationship to improve the economic growth substantially.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at N/A.

AUTHOR CONTRIBUTIONS

Data curation, LG, DA and CC; Formal analysis, LG and SS; Methodology LG, CC, and SS; Software, LG and SS, Writing—original draft, LG.

FUNDING

This study was funded by funds from Corporación universitaria Americana.

REFERENCES

Agosin, M. R., and Bravo-Ortega, C. (2009). *The Emergence of New Successful Export Activities in Latin America: The Case of Chile*.

Austin, J. E. (2002). *Managing in Developing Countries: Strategic Analysis and Operating Techniques*. Simon & Schuster.

Besedes, T., and Blyde, J. (2010). What Drives Export Survival? an Analysis of Export Duration in Latin America. *Inter-American Dev. Bank. mimeo* 1, 1–43.

Bhatti, U. A., Yu, Z., Chanussot, J., Zeeshan, Z., Yuan, L., Luo, W., et al. (2021). Local Similarity-Based Spatial-Spectral Fusion Hyperspectral Image Classification with Deep CNN and Gabor Filtering. *IEEE Trans. Geoscience Remote Sens*. doi:10.1109/TGRS.2021.3090410

Bhatti, U. A., Zeeshan, Z., Nizamani, M. M., Bazai, S., Yu, Z., and Yuan, L. (2022). Assessing the Change of Ambient Air Quality Patterns in Jiangsu Province of China Pre-to Post-COVID-19. *Chemosphere* 288, 132569. doi:10.1016/j.chemosphere.2021.132569

Canuto, O. (2019). *Traps on the Road to High Income*. Policy Center for the New South. Policy Brief PB-19/14, April.

Cherif, R., and Hasanov, F. (2019). The Leap of the Tiger: Escaping the Middle-income Trap to the Technological Frontier. *Glob. Policy* 10 (4), 497–511. doi:10.1111/1758-5899.12695

Clarke, G., Cull, R., Peria, M. S. M., and Sanchez, S. M. (2003). Foreign Bank Entry: Experience, Implications for Developing Economies, and Agenda for Further Research. *World Bank Res. Observer* 18 (1), 25–59. doi:10.1093/wbro/lkg002

Dahlman, C. J., Ross-Larson, B., and Westphal, L. E. (1987). Managing Technological Development: Lessons from the Newly Industrializing Countries. *World Dev.* 15 (6), 759–775. doi:10.1016/0305-750x(87)90058-1

Dui, H. J. (2020). The Middle Income Trap in China: Enlightenment from Latin American Countries. *Middle Income Trap China Enlight. Lat. Am. Ctries.* 64 (1), 6. doi:10.47119/ijrp1006411120201518

Eichengreen, B., Park, D., and Shin, K. (2013). *Growth Slowdowns Redux: New Evidence on the Middle-Income Trap (No. W18673)*. National Bureau of Economic Research.

Engel, J., and Taglioni, D. (2017). The Middle-Income Trap and Upgrading along Global Value Chains. *Glob. Value Chain Rep.*, 119–139.

Engle, R. F., and Granger, C. W. J. (1987). Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica* 55, 251–276. doi:10.2307/1913236

Estrada, G., Han, X., Park, D., and Tian, S. (2018). Asia's Middle-Income Challenge: An Overview. *Emerg. Mark. Finance Trade* 54 (6), 1208–1224. doi:10.1080/1540496x.2017.1421939

Fahim, A., Addae, B. A., Ofosu-Adarkwa, J., Qingmei, T., and Bhatti, U. A. (2021). Industry 4.0 and Higher Education: An Evaluation of Barriers Affecting Master's in Business Administration Enrolments Using a Grey Incidence Analysis. *IEEE Access* 9, 76991–77008. doi:10.1109/access.2021.3082144

Florax, R. J. G. M., and de Graaff, T. (2004). "The Performance of Diagnostic Tests for Spatial Dependence in Linear Regression Models: a Meta-Analysis of Simulation Studies," in *Advances in Spatial Econometrics* (Berlin, Heidelberg: Springer), 29–65. doi:10.1007/978-3-662-05617-2_2

Goldie, S. J., Diaz, M., Constenla, D., Alvis, N., Andrus, J. K., and Kim, S.-Y. (2008). Mathematical Models of Cervical Cancer Prevention in Latin America and the Caribbean. *Vaccine* 26, L59–L72. doi:10.1016/j.vaccine.2008.05.063

Han, X., and Wei, S.-J. (2017). Re-examining the Middle-Income Trap Hypothesis (MITH): What to Reject and what to Revive? *J. Int. Money Finance* 73, 41–61. doi:10.1016/j.jimonfin.2017.01.004

Hartwell, C. A. (2018). Old Wine and New Bottles: A Critical Appraisal of the Middle-Income Trap in BRICS Countries. *Rujec* 4 (2), 133–154. doi:10.3897/j.ruje.4.27726

Hassler, U., and Wolters, J. (2006). "Autoregressive Distributed Lag Models and Cointegration," in *Modern Econometric Analysis* (Berlin, Heidelberg: Springer), 57–72. doi:10.1007/s10182-006-0221-5

Heinicke, M. P., Duellman, W. E., and Hedges, S. B. (2007). Major Caribbean and Central American Frog Faunas Originated by Ancient Oceanic Dispersal. *Proc. Natl. Acad. Sci. U.S.A.* 104 (24), 10092–10097. doi:10.1073/pnas.0611051104

Irfan, M., Elavarasan, R. M., Ahmad, M., Mohsin, M., Dagar, V., and Hao, Y. (2022). Prioritizing and Overcoming Biomass Energy Barriers: Application of AHP and G-TOPSIS Approaches. *Technol. Forecast. Soc. Change* 177, 121524. doi:10.1016/j.techfore.2022.121524

Jayasooriya, S. P. (2017). *Growth Slowdowns, Middle-Income Trap, and Demographic Profile in South Asia (No. 736)*. ADBI Working Paper.

Kanchoochat, V. (2014). The Middle-Income Trap Debate: Taking Stock, Looking Ahead. *Kokusai Mondai Int. Aff.* (633), 1–19.

Khan, I., Zakari, A., Dagar, V., and Singh, S. (2022). World Energy Trilemma and Transformative Energy Developments as Determinants of Economic Growth amid Environmental Sustainability. *Energy Econ.* 108, 105884. doi:10.1016/j.eneco.2022.105884

Khan, I., Zakari, A., Zhang, J., Dagar, V., and Singh, S. (2022). A Study of Trilemma Energy Balance, Clean Energy Transitions, and Economic Expansion in the Midst of Environmental Sustainability: New Insights from Three Trilemma Leadership. *Energy* 248, 123619. doi:10.1016/j.energy.2022.123619

Klingler-Vidra, R., and Wade, R. (2020). Science and Technology Policies and the Middle-Income Trap: Lessons from Vietnam. *J. Dev. Stud.* 56 (4), 717–731. doi:10.1080/00220388.2019.1595598

Kotarski, K., and Petak, Z. (2019). "Croatia's Post-communist Transition Experience: The Paradox of Initial Advantage Turning into a Middle-Income Trap," in *Policy-Making at the European Periphery* (Cham: Palgrave Macmillan), 1–25. doi:10.1007/978-3-319-73582-5_1

Kuchiki, A., Gokan, T., and Maruya, T. (2017). "Railway-led Formation of the Agriculture-Food-Tourism Industry Cluster: Escaping the Middle-Income Trap," in *A Multi-Industrial Linkages Approach to Cluster Building in East Asia* (London: Palgrave Macmillan), 187–205. doi:10.1057/978-1-137-57128-1_9

Lall, S. (2000). *Technological Change and Industrialization in the Asian Newly Industrializing. Technology, Learning, and Innovation: Experiences of Newly Industrializing Economies*, 13.

Lebdioui, A., Lee, K., and Pietrobelli, C. (2021). Local-foreign Technology Interface, Resource-Based Development, and Industrial Policy: How Chile and Malaysia Are Escaping the Middle-Income Trap. *J. Technol. Transf.* 46 (3), 660–685. doi:10.1007/s10961-020-09808-3

Lee, K., and Kim, Y. K. (2018). "Comparing the National Innovation Systems in East Asia and Latin America: Fast versus Slow," in *Innovation in the Asia Pacific* (Singapore: Springer), 63–82. doi:10.1007/978-981-10-5895-0_3

Lin, J. Y., and Wang, Y. (2020). Structural Change, Industrial Upgrading, and Middle-Income Trap. *J. Ind. Compet. Trade* 20, 359–394. doi:10.1007/s10842-019-00330-3

Mahul, O., and Stutley, C. J. (2010). *Government Support to Agricultural Insurance: Challenges and Options for Developing Countries*. World Bank Publications.

Máttar, J. (2019). "Long-term Challenges of Industrial Development in Latin America and the Caribbean," in *Transforming Industrial Policy for the Digital Age* (Edward Elgar Publishing). doi:10.4337/9781788976152.00013

Muhammad, B., and Khan, M. K. (2021). Foreign Direct Investment Inflow, Economic Growth, Energy Consumption, Globalization, and Carbon Dioxide Emission Around the World. *Environ. Sci. Pollut. Res.* 28 (Issue 39), 55643–55654. doi:10.1007/s11356-021-14857-8

Nguyen-Huu, T. T., and Pham, N. S. (2021). *Escaping the Middle Income Trap and Getting Economic Growth: How Does FDI Can Help the Host Country?*

Ohno, K. (2009). Avoiding the Middle-Income Trap: Renovating Industrial Policy Formulation in Vietnam. *ASEAN Econ. Bull.* 26, 25–43. doi:10.1355/ae26-1c

Oreiro, J. L., da Silva, K. M., and Dávila-Fernández, M. J. (2020). A New Developmentalist Model of Structural Change, Economic Growth and Middle-Income Traps. *Struct. Change Econ. Dyn.* 55, 26–38. doi:10.1016/j.strueco.2020.07.008

Ortiz, I., Duran, F., Urban, S., Wodsak, V., and Yu, Z. (2018). *Reversing Pension Privatization: Rebuilding Public Pension Systems in Eastern European and Latin American Countries (2000-18)*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3275228.

Ozturk, A. (2016). Examining the Economic Growth and the Middle-Income Trap from the Perspective of the Middle Class. *Int. Bus. Rev.* 25 (3), 726–738. doi:10.1016/j.ibusrev.2015.03.008

Paus, E. (2014). *Latin America and the Middle Income Trap*. Financing for Development Series. ECLAC, 250.

Paus, E. (2018). *The Middle-Income Trap: Lessons from Latin America: Lessons from Latin America*. Avoiding The Middle-Income Trap in Asia: The Role of Trade, Manufacturing, and Finance. Tokyo: Asian Development Bank Institute, 60–105.

Pesaran, M. H., and Shin, Y. (1995). *An Autoregressive Distributed Lag Modelling Approach to Cointegration Analysis*.

Pesaran, M. H., Shin, Y., and Smith, R. J. (2001). Bounds Testing Approaches to the Analysis of Level Relationships. *J. Appl. Econ.* 16 (3), 289–326. doi:10.1002/jae.616

Pesaran, M. H., Shin, Y., and Smith, R. P. (1999). Pooled Mean Group Estimation of Dynamic Heterogeneous Panels. *J. Am. Stat. Assoc.* 94 (446), 621–634. doi:10.1080/01621459.1999.10474156

Rowe, A. K., Rowe, S. Y., Peters, D. H., Holloway, K. A., Chalker, J., and Ross-Degnan, D. (2018). Effectiveness of Strategies to Improve Health-Care Provider Practices in Low-Income and Middle-Income Countries: a Systematic Review. *Lancet Glob. Health* 6 (11), e1163–e1175. doi:10.1016/s2214-109x(18)30398-x

Sabel, C., Fernández-Arias, E., Hausmann, R., Rodriguez-Clare, A., and Stein, E. (2012). *Export Pioneers in Latin America*. IDB Book No. IDB-BK-107. Columbia Law and Economics Research Paper, 421.

Shahzad, U., Madaleno, M., Dagar, V., Ghosh, S., and Doğan, B. (2022). Exploring the Role of Export Product Quality and Economic Complexity for Economic Progress of Developed Economies: Does Institutional Quality Matter? *Struct. Change Econ. Dyn.* 62, 40–51. doi:10.1016/j.strueco.2022.04.003

Silva, M. L. F. (2018). *Modern Exchange-Rate Regimes, Stabilisation Programmes and Co-ordination of Macroeconomic Policies: Recent Experiences of Selected Developing Latin American Economies*. Routledge.

Sim, A. B., and Ali, Y. (1998). Performance of International Joint Ventures from Developing and Developed Countries: An Empirical Study in a Developing Country Context. *J. World Bus.* 33 (4), 357–377. doi:10.1016/s1090-9516(99)80080-7

Tang, C., Irfan, M., Razzaq, A., and Dagar, V. (2022). Natural Resources and Financial Development: Role of Business Regulations in Testing the Resource-Curse Hypothesis in ASEAN Countries. *Resour. Policy* 76, 102612. doi:10.1016/j.resourpol.2022.102612

World Bank (2022). World Bank Data. Available at: https://data.worldbank.org/. (Accessed March 3, 2022).

Xie, M., Irfan, M., Razzaq, A., and Dagar, V. (2022). Forest and Mineral Volatility and Economic Performance: Evidence from Frequency Domain Causality Approach for Global Data. *Resour. Policy* 76, 102685. doi:10.1016/j.resourpol.2022.102685

Yavuz Tiftikçigil, B., Güriş, B., and Yaşgül, Y. S. (2018). Does Middle Income Trap Exist?: Evidence from Emerging Economies: E7 Countries for 1969-2015. *Rge* 27 (1), 145–158. doi:10.15304/rge.27.1.5234

Zhang, C., Khan, I., Dagar, V., Saeed, A., and Zafar, M. W. (2022). Environmental Impact of Information and Communication Technology: Unveiling the Role of Education in Developing Countries. *Technol. Forecast. Soc. Change* 178, 121570. doi:10.1016/j.techfore.2022.121570

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.

Copyright © 2022 Galvan, Campo, Stanojevic and Alzate. 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership