Chaos, Solitons and Fractals 158 (2022) 112098

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization

Jujie Wang [a,b,*], Wenjie Xu [a], Yue Zhang [a], Jian Dong [a]

[a] School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
[b] Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China

ARTICLE INFO

Article history:
Received 21 January 2022
Received in revised form 29 March 2022
Accepted 6 April 2022
Available online xxxx

Keywords:
Air quality prediction and early warning
Optimal feature extraction
Combined model
Multi-step ahead forecasting
Multi-objective grey wolf optimization
Forecasting accuracy and stability

ABSTRACT

An effective air pollution prediction is of great significance to prevent and control air pollution and protect the health of residents. In order to improve the prediction accuracy of $PM_{2.5}$, an innovative $PM_{2.5}$ concentration prediction and early warning system based on optimal feature extraction and intelligent optimization is developed in this study. First, a feedback variational modal decomposition algorithm is designed to decompose the $PM_{2.5}$ concentration sequence and fuzzy entropy is used to reconstruct the patterns of similar complexity. Then, Copula entropy is used to select the influencing factors with a high impact on $PM_{2.5}$. Next, the reconstructed components and influencing factors are inputted to three individual prediction models, including long short-term memory neural network, gated recurrent unit neural network, and temporal convolutional network, for training and multi-step short-term prediction. The results of the individual prediction models are nonlinearly combined by Gaussian process regression which is optimized by the multi-objective grey wolf optimization algorithm. Finally, the prediction results of different reconstructed components are nonlinearly integrated to obtain the final $PM_{2.5}$ prediction results. In an empirical study of two Chinese cities, the combined prediction model proposed in this study outperformed the other six comparative models in terms of prediction accuracy and stability. The experimental results prove that the hybrid prediction model proposed in this paper can make an effective prediction and early warnings of air pollution.

1. Introduction

Breathing fresh air is of great importance to the healthy life of human beings. However, as the global economy develops and modern industrialization accelerates, the problem of air pollution has aroused widespread concern around the world. Severe air pollution has attracted the attention of the Chinese government, and relevant researchers have begun to pay close attention to air pollution and its effects. Many epidemiological studies have shown that various respiratory diseases, circulatory systems, and other diseases are related to air pollution [1]. Meanwhile, studies have shown that China's severe air pollution reduces public happiness levels [2]. $PM_{2.5}$, which consists of highly reactive toxic and harmful substances, is the primary pollutant of air pollution [3]. It is not only generated from natural soil dust but also from energy combustion and human industrial production [4].

At China's current stage of economic development, air pollution is unavoidable, but that does not mean it cannot be controlled. Since severe air pollution directly impacts the environmental quality and human health, accurate $PM_{2.5}$ concentration prediction is one of the primary goals of air quality research. Based on accurate $PM_{2.5}$ concentration prediction results, the government can assess the possibility of severe air pollution and provide scientific warnings. According to the predictive air pollution levels, citizens can make reasonable travel arrangements to protect their physical and mental health. However, current $PM_{2.5}$ concentration prediction systems are still poor in accuracy and stability. More attention and research are needed to develop a more accurate $PM_{2.5}$ concentration prediction

Abbreviations: ANN, Artificial neural network; ANFIS, Adaptive neuro-fuzzy inference system; ApproEn, Approximate entropy; GA, Genetic algorithm; PSO, Particle swarm optimization; AQI, Air quality index; ARIMA, Autoregressive integrated moving average; CopulaEn, Copula entropy; EMD, Empirical mode decomposition; FuzzyEn, Fuzzy entropy; GPR, Gaussian process regression; GRU, Gated recurrent unit; LSTM, Long short-term memory; MAE, Mean absolute error; MAPE, Mean absolute percentage error; MdAPE, Median of absolute percentage error; MLP, Multilayer perceptron; MLR, Multiple linear regression; MOGWO, Multi-objective grey wolf optimization algorithm; NSE, Nash Sutcliffe Efficiency; RNN, Recurrent neural networks; RMSE, Root mean square error; SampleEn, Sample entropy; SVR, Support vector regression; TCN, Temporal convolutional network; VMD, Variational mode decomposition.

* Corresponding author at: School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
E-mail address: jujiewang@126.com (J. Wang).

https://doi.org/10.1016/j.chaos.2022.112098
0960-0779/

model and generate more accurate scientific warnings for potentially severe air pollution events based on accurate prediction results. More attention and research are needed to design a more accurate $PM_{2.5}$ prediction model and generate more accurate scientific warnings for potentially severe air pollution events based on the accuracy of the prediction results.

The fluctuation of $PM_{2.5}$ is influenced by a variety of factors, making it complexly volatile and extremely sudden [5]. Accurate and stable $PM_{2.5}$ prediction can improve the system reliability of air pollution prediction and early warning systems. Currently, researchers have developed a variety of prediction techniques for $PM_{2.5}$ concentrations, which are summarized into four main categories: (1) physical models, (2) statistical methods, (3) machine learning techniques, (4) hybrid models.

Physical methods based on atmospheric physical and chemical processes use meteorological principles and mathematical methods to simulate air quality horizontally and vertically at large scales [6]. Common physical models are the Community Multiscale Air Quality Modeling System (CMAQ), Weather Research and Forecasting Model with Chemistry (WRF-Chem) [7], and the Nested Air Quality Prediction Modeling System (NAQPMS) [8]. Ge et al. [9] developed an online source tracking method combining the cloud process module in the nested air quality prediction modeling system (NAQPMS) to effectively track acidic emission precursors when dealing with non-linear secondary stratospheric pollutants. Physical models have good prediction accuracy, but they are more suitable for long-term prediction and have limited practical application in short-term prediction [10]. Physical models also have certain limitations in practical application, such as reliance on the quality of pollutant emission data, the complexity of calculations, and high uncertainty of prediction results.

Classical statistical models predict air pollutants from the perspective of time series analysis. Statistical models commonly include autoregressive integrated moving average (ARIMA) and multiple linear regression (MLR). Zhang et al. [11] applied an ARIMA model to predict $PM_{2.5}$ concentrations, and the analysis showed that $PM_{2.5}$ concentrations were significantly and positively correlated with PM_{10}, SO_2, and NO_2 concentrations. Lesar and Filipčić [12] simulated hourly $PM_{2.5}$ concentrations on sea breeze days using MLR, and the simulated hourly values were in good agreement with the measured values. Although statistical models can effectively make short-term predictions of air pollutant concentrations, their inherent linearity assumptions make them unable to solve non-linear problems. When dealing with non-linear features, statistical models have difficulty accurately capturing non-linear information, resulting in more significant prediction errors and poorer stability of the models.

Machine learning techniques have been gradually applied to air pollution studies to further increase the accuracy of air pollutant concentration prediction. Machine learning techniques are characterized by strong non-linear feature extraction, high generalization ability, and high prediction accuracy. Asadollahfardi et al. [13] used multilayer perceptron (MLP) neural networks, radial basis function (RBF) neural networks, and Markov chain models to predict $PM_{2.5}$ concentrations, respectively. They demonstrated that artificial neural networks (ANN) have better predictive performance. Rubal and Kumar [14] applied a random forest technique to air pollutant concentration prediction and obtained satisfactory results. Leong et al. [15] used a support vector machine (SVM) for $PM_{2.5}$ concentration simulation and proved that SVM with radial basis function could effectively solve the air pollutant concentration forecasting problem. Although a single machine learning prediction model has been proved to be effective in predicting the concentration of air pollutants, more researchers construct hybrid models to improve the prediction performance of the model under different data conditions [16].

Related research shows that hybrid models are mainly classified into two types: one is to mix different kinds of prediction models, and the other is to mix data processing methods and optimization algorithms with prediction models. Researchers have widely used data preprocessing methods based on decomposition and integration techniques in recent years to construct hybrid prediction models [17]. The signal decomposition algorithm can decompose a non-linear non-smooth time series into several smoother sub-series, from which more useful information can be extracted. Predicting the decomposed subsequences and then integrating the prediction results can reduce the computational complexity of the prediction model and improve the prediction performance. Therefore, data preprocessing is gradually becoming an increasingly critical technique to improve the prediction performance of hybrid models [18]. Several hybrid models based on decomposition integration techniques have been proposed for air pollutant prediction studies. Chen et al. [19] used an autoregressive integrated moving average model (ARIMA), ANN, and SVM combined with wavelet decomposition to predict $PM_{2.5}$ concentrations. They demonstrated that wavelet decomposition could capture fluctuations of $PM_{2.5}$ concentrations more accurately, which can provide early warning prediction of air pollution with effective information support. However, the performance of wavelet decomposition is hugely dependent on the subjectively selected wavelet basis functions and decomposition levels, and there is a lack of a specific and compelling theoretical basis for choosing wavelet basis functions and decomposition levels. Therefore, as a data-driven adaptive decomposition technique, empirical mode decomposition (EMD) is favoured by researchers. Zhu et al. [20] applied EMD and support vector regression (SVR) to air quality index (AQI) prediction and demonstrated that EMD helps to improve the accuracy of prediction models further. Although EMD can handle complex non-linear signals adaptively, it also suffers from the lack of rigorous mathematical theory, endpoint effects, and modal mixing [21]. The variational mode decomposition (VMD) technique can effectively handle non-linear, non-stationary complex signals while avoiding problems such as endpoint effects, spurious components, and boundary effects [22]. Wu and Lin [23] combined VMD and sample entropy (SE) and used long short-term memory (LSTM) neural networks for AQI prediction. Their results demonstrated that VMD could effectively capture the intrinsic features of the original AQI sequence and improve the prediction accuracy. Although VMD has the advantages of better decomposition performance and strong resistance to noise interference, the number of layers of signal decomposition in VMD affects the decomposition performance. It depends on human selection. As an essential module of the hybrid model, the selection and optimization of the signal decomposition algorithm still need more research. In addition, although the signal decomposition algorithm can decompose the $PM_{2.5}$ concentration sequence into several smoother subseries, there is often a similar complexity between subsequences, which is often overlooked by researchers. Modeling all decomposed subsequences would complicate the model computation and reduce the computational efficiency of the prediction model [24].

Although hybrid models based on data preprocessing techniques have proven to be practical tools for air pollution prediction, each prediction model has its own disadvantages that are difficult to overcome due to the different inherent properties of different prediction models [25]. In addition, different data have different data distribution and characteristics, and different prediction models have different prediction performances on the same data set. To address these issues, research on air pollutant concentration prediction requires further exploration of combined predictive modeling techniques. Combined forecasting models have received increasing attention from researchers since Bates and Granger proposed the theoretical foundations of combinatorial forecasting [26]. The commonly used combination prediction techniques are broadly classified into the traditional linear weighted combined model and the machine learning-based non-linear combined model. Xiao et al. [27] used multiple single-prediction models to forecast the electric load and then combined the single-prediction models and used the cuckoo search algorithm (CS) to optimize the combined models' weight coefficients. Liu et al. [28] developed a combined neural network fuzzy forecasting model for wind speed prediction and optimized the combined weights with an improved CS. It is demonstrated

that the combined forecasting technique further improves the forecasting accuracy while providing more trend variation for time series forecasting. In addition, Wang et al. [29] developed a robust combined forecasting model combining ARIMA, SVM, extreme learning machine (ELM), and least squares support vector machine (LSSVM). They used Gaussian process regression to combine the predictions of each model nonlinearly to achieve an effective short-term prediction of wind speed. It is demonstrated that the proposed non-linear combined prediction model outperforms individual models in terms of prediction performance and stability. The combined prediction technique usually obtains the optimal weights by minimizing the prediction error of the training samples [30]. In addition, a single performance evaluation metric is challenging to represent the actual predictive performance of the model entirely. It is more desirable to simultaneously provide different performance evaluation metrics to obtain better prediction accuracy when using optimization algorithms to optimize the combined model weights. However, traditional single-objective optimization algorithms such as particle swarm algorithm [31], simulated annealing algorithm [32], and genetic algorithm cannot be applied to multi-objective problems [33]. Therefore, the development of combined prediction models based on multi-objective optimization algorithms needs more attention and research.

From the above review, it can be found that the previously proposed air pollutant prediction methods are not perfect and have some unavoidable drawbacks. (1) Physical methods are suitable for long-term prediction. It is difficult to make short-term predictions effectively and has disadvantages such as reliance on data quality, high computational complexity, and long operation time. (2) Statistical methods are based on statistical assumptions and cannot effectively predict non-linear time series. (3) Machine learning algorithms can effectively extract and process complex sequences and extract non-linear features. Still, they also suffer from the problems of being prone to local optima and overfitting. (4) The importance and necessity of data preprocessing need to be given more attention. The selection and optimization of signal decomposition algorithms affect the prediction performance and accuracy. Meanwhile, the similarity complexity and similarity between the sequences obtained by decomposition are often neglected. (5) Although hybrid models based on decomposition and integration techniques can make compelling predictions, the single model's prediction accuracy and stability are still insufficient. (6) When optimizing a prediction model, a single performance evaluation metric can hardly fully represent the actual prediction performance of the model, and the traditional single-objective optimization cannot be applied to multi-objective problems.

Based on the above considerations, this paper develops a novel air pollution prediction and early warning system containing the Feedback-VMD algorithm (FVMD), Fuzzy Entropy (FuzzyEn), Copula Entropy (CopulaEn), LSTM, Gated Recurrent Unit (GRU), Temporal Convolutional Network (TCN), Gaussian Process Regression (GPR) and Multi-Objective Grey Wolf Optimization algorithm (MOGWO). First, an optimal feature extraction technique based on FVMD and FuzzyEn is developed to extract the intrinsic features of $PM_{2.5}$ concentration data and reduce the complexity of model computation. Second, the CopulaEn algorithm is employed in the selection of influencing factors that strongly impact the fluctuation of $PM_{2.5}$ concentration. Then, LSTM, TCN, and GRU are used as three individual models for multi-step short-term prediction. Next, the MOGWO-optimized GPR is used as a non-linear combination of individual prediction models. Finally, all the predictions obtained from the combined predictions are nonlinearly integrated to get the final prediction results, and the future air pollution levels are evaluated and warned. This paper selects actual $PM_{2.5}$ concentration data of two Chinese cities, Shanghai and Guangzhou for empirical analysis. The empirical results prove that the combined prediction model proposed in this paper has better prediction performance and stability than other comparative models. The air pollution prediction and early warning system are constructed based on accurate and reliable early warning effects. The main innovations and contributions of this study are as follows.

(a) A Feedback-VMD algorithm is developed to adaptively determine the number of decomposition layers of $PM_{2.5}$ concentration data to extract the intrinsic information from $PM_{2.5}$ data effectively. In addition, this paper considers the complex relationship between different decomposition patterns. It uses FuzzyEn to reorganize the decomposition patterns into several new components to reduce the computational complexity of the model.
(b) To further improve the predictive performance and robustness of the model, several influencing factors that significantly impact $PM_{2.5}$ were selected as relevant variables introduced into the $PM_{2.5}$ prediction study using the CopulaEn algorithm.
(c) The LSTM, TCN, and GRU are used as three individual forecasting models for multi-step short-term forecasting. The MOGWO-optimized GPR model is used as a non-linear combination model to absorb the advantages of different individual forecasting models to obtain better combination forecasting results.
(d) After obtaining the combined prediction results of different components, the final $PM_{2.5}$ concentration prediction results are obtained by non-linear integration. The future air pollution level is evaluated and effectively warned.
(e) This paper develops an air pollution prediction and early warning system based on the combined prediction models with optimal feature extraction and intelligent optimization. Scientific evaluation criteria are used, and simulation experiments are conducted in two cities in China. The simulation results demonstrate that the proposed hybrid framework has good prediction accuracy and stability.

The rest of the paper is organized as follows. Section 2 describes the main structure and theory of the proposed hybrid framework. Section 3 describes the combined forecasting model's data preprocessing and evaluation metrics. Section 4 shows the prediction part, comparison, and discussion of the combined prediction model. The last section concludes the whole paper and presents the outlook for future research.

2. Structure of the proposed air pollution forecasting and warning framework

This section introduces the main structure of the developed air pollution forecasting and warning system, which can be divided into three stages: optimal feature extraction and feature selection, multi-step combination forecasting with MOGWO optimization, and non-linear integration of air pollution forecasting and warning. A brief flowchart of the proposed framework is shown in Fig. 1.

2.1. Stage 1: optimal feature extraction and feature selection

The first stage can be divided into three modules: the feedback-based VMD adaptive signal decomposition method, the FuzzyEn-based reconstruction method, and the CopulaEn-based feature selection method. The specific contents and implementation methods are as follows:

2.1.1. Feedback variational mode decomposition

VMD is a signal decomposition algorithm based on Wiener filters, Hilbert transforms, and mixed frequencies for variational problems [34]. Traditional signal decomposition algorithms usually have strict data requirements. For example, the Fourier transform is suitable for processing smooth periodic signals [35], while wavelet analysis can effectively capture transient effects in non-stationary signals [36]. More, EMD does not depend on the a priori information of the data and has good adaptiveness to capture the effective features in non-linear signals. However, the problems of modal mixing and endpoint effects of EMD

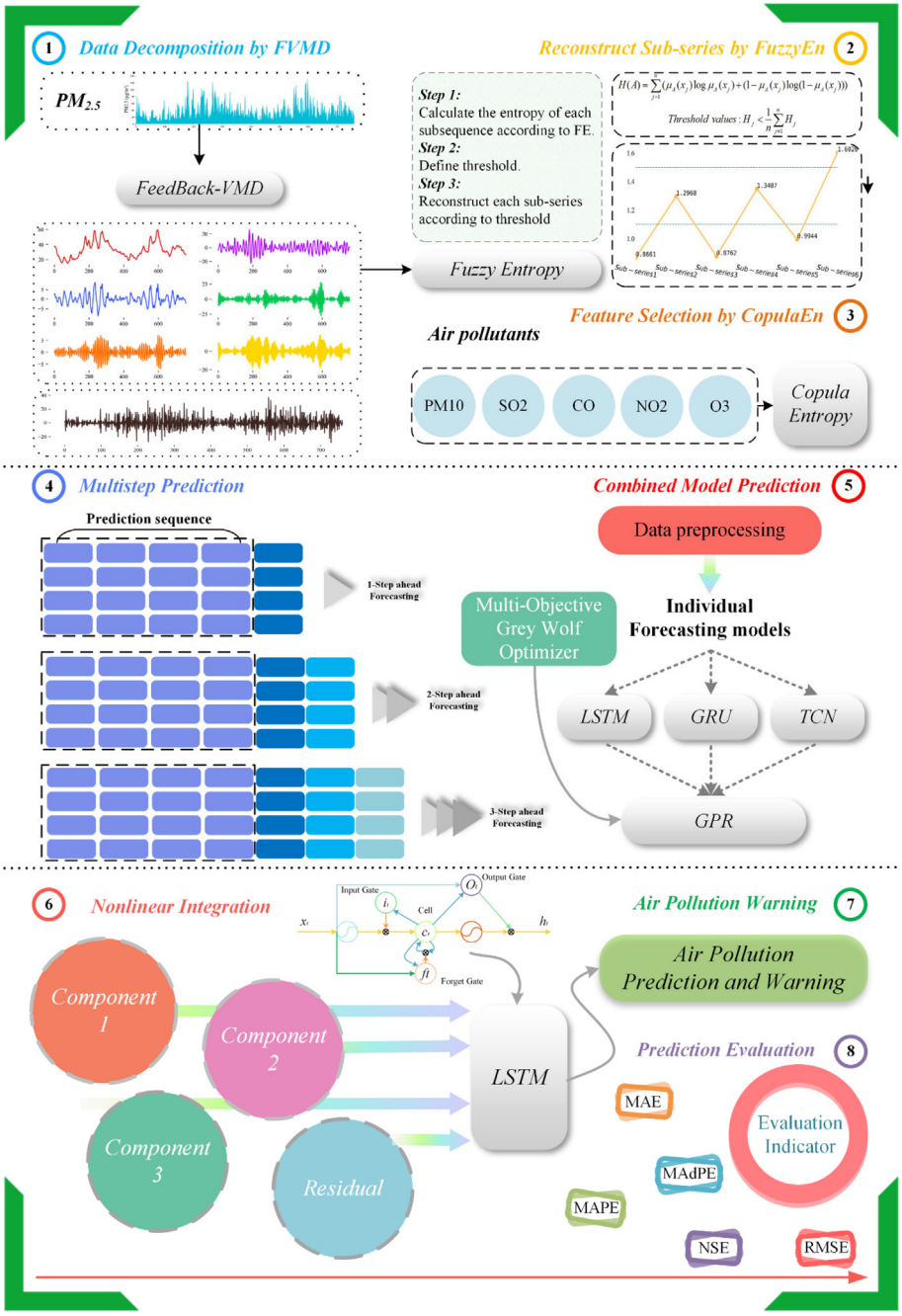


Fig. 1. Flow chart of air pollution prediction and early warning system.

limit its decomposition performance. In contrast, VMD can effectively solve the issues of modal mixing and endpoint effects compared with EMD, and it has good decomposition accuracy and better resistance to noise interference when decomposing complex nonlinear data [37]. The details of the VMD algorithm are shown in the Appendix.

Like many successful clustering and decomposition algorithms, the VMD algorithm requires a predetermined number of decomposition modes k. Related studies have proved that the number of decomposition modes affects the decomposition efficiency and noise interference resistance of VMD [38]. If the number of decomposition modes k is too tiny, information may be missing due to insufficient decomposition. If the number of decomposition modes k is set too large, too many modes may lead to phenomena such as time-frequency overlap and capturing extra noise [39].

To address the above problems, this paper proposes a Feedback-VMD algorithm (FVMD) that does not require a predetermined value k. FVMD first performs a two-mode VMD of the input signal to obtain two-mode signals. The similarity coefficients between each mode signal and the observed signal are obtained by calculating the decomposition separately, and the optimal mode component of the mode signal with the higher similarity coefficient is noted. Then, the optimal mode component is fed back to the input of the VMD, and this mode component is subtracted at the input of the VMD. Then the remaining signal is used as a new signal for the two-mode VMD again. The above process is repeated until the maximum value of the similarity coefficient of the two-mode components obtained from the nth decomposition is less than the minimum value of the similarity coefficient of the two-mode components obtained from the n-1th decomposition. The mixed signal is considered to be completely decomposed.

Let the mode components obtained from the nth decomposition be $x_{n,1}(t)$ and $x_{n,2}(t)$. The similarity coefficients of $x_{n,1}(t)$ and $x_{n,2}(t)$ to the observed signal $x(t)$ are $\zeta_{n,1}$ and $\zeta_{n,2}$. Where the similarity coefficient $\zeta_{n,i}(i=1,2)$ is calculated as shown below.

$$\zeta_{n,i}=\frac{\left|\sum_{t=1}^{N} x_{n,i}(t)\cdot x(t)\right|}{\sqrt{\sum_{t=1}^{N} x_{n,i}^2(t)\sum_{t=1}^{N} x^2(t)}} \tag{1}$$

Where N denotes the length of the observed signal $x(t)$. Then the optimal mode component is $x_{n,best}$.

$$x_{n,best}=\underset{i}{\operatorname{argmax}}\{\zeta_{n,i}\} \tag{2}$$

The discriminant condition for the FVMD stop decomposition is:

$$\max\{\zeta_{n,1},\zeta_{n,2}\}<\min\{\zeta_{n-1,1},\zeta_{n-1,2}\} \tag{3}$$

The FVMD algorithm proposed in this paper does not require a preset number of decompositions k. The observed signal is finally decomposed into $n+1$ subsequences, removing noise by adaptively n iterating times. This subseries does not overlap with each other and contain trend information of $PM_{2.5}$ concentration changes, and they will revert to the original series after adding the rejected noise. However, the noisy series are usually removed in previous time series prediction studies. $PM_{2.5}$ concentration series have complex characteristics such as non-linearity and non-smoothness, and complex factors influence the concentration changes. At the same time, the noise contains various other random factors that are difficult to measure and affect the fluctuation of $PM_{2.5}$ concentration. Some stochastic factors may include complex extreme weather and natural environmental changes (e.g., sudden wind and rainstorms and natural fires) and human activities (e.g., construction dust and traffic trips). Therefore, the excluded noise may contain information on short-term variations and extreme changes that affect $PM_{2.5}$ fluctuations. This paper obtains the residual signal $x_{residual}$ by subtracting all FVMD decompositions from the observed signal to obtain the mode component.

$$x_{residual}=x(t)-\sum_{n=1}^{n-1}x_{n,best}-x_{n,1}(t)-x_{n,2}(t) \tag{4}$$

The pseudocode for FVMD is shown below.

Algorithm 1. Feedback VMD.

Input: *PM2.5 concentration time series* $x(t)$

1. **Begin**
2. $k=2$, *Initialize the number of iterations* $n=1$
3. *Initialize the input sequence* $x_n(t)=x(t)$
4. *Decompose the signal* $x_n(t)$ *using VMD, two-mode components* $x_{n,1}(t)$ *, and* $x_{n,2}(t)$ *are obtained*
5. *Calculate the similarity coefficients* $\zeta_{n,1}$ *and* $\zeta_{n,2}$ *to obtain the optimal mode components* $x_{n,best}=\operatorname{argmax}_i\{\zeta_{n,i}\}$,
6. **IF** $\max\{\zeta_{n,1},\zeta_{n,2}\}\geq\min\{\zeta_{n-1,1},\zeta_{n-1,2}\}$ **THEN**
7. $x_{n+1}(t)=x_n(t)-x_{n,best}$ $n=n+1$, *repeat steps 4 and 5*
8. **End IF**
9. $x_{residual}=x(t)-\sum_{n=1}^{n-1}x_{n,best}-x_{n,1}(t)-x_{n,2}(t)$
10. **End**

2.1.2. Fuzzy entropy

Fuzzy entropy (FuzzyEn) can quantify and categorize the degree of uncertainty in random variables. It can be used to evaluate the complication of time series [40]. FuzzyEn, as an improved technique to SampleEn and Approximate entropy (ApproEn), introduces the concept of fuzzy set, retains the advantages of SampleEn and ApproEn, and eliminates the disadvantage of erroneous entropy analysis in the presence of minor variations and baseline drift.

As a technique for quantifying the time series' complexity, a higher entropy value of FuzzyEn indicates a higher probability of generating new patterns, i.e., higher complexity of time series. However, the similar complexity and correlation between different subsequences are often ignored. In this study, FuzzyEn is used to measure the complexity of the decomposed subsequences, and subsequences with similar complexity are reconstructed into a new component. The optimal feature extraction technique combining FVMD and FuzzyEn proposed in this paper effectively reduces the complexity of the subsequences and enhances the model's computational efficiency and prediction performance.

2.1.3. Copula entropy

Ma and Sun proposed a new concept of entropy, called Copula entropy (CopulaEn), which can be used to measure the full-order correlation between random variables [41]. The concept of correlation is a fundamental statistical idea that measures the intrinsic statistical link between random variables. The Pearson correlation coefficient, a classical correlation measure, is widely used to measure the degree of correlation between two variables [42]. However, it is only applicable to the linear case and implicitly has the defect of the assumption of Gaussian distribution, which makes it challenging to apply in practical situations.

CopulaEn, as a more advanced correlation measure, has the advantages of no model assumptions, the ability to handle non-linear relationships and monotonic transformation invariance, and applies to any type of correlation measure [43]. In this paper, several pollutant factors associated with $PM_{2.5}$ concentration changes were introduced into the prediction study, and CopulaEn was used to measure the correlation between each influencing factor and $PM_{2.5}$ concentration. The influencing factor with a higher correlation was selected for $PM_{2.5}$ concentration prediction.

2.2. Stage 2: multi-step combination prediction with MOGWO optimization

The second stage can be divided into individual prediction models and combined models with multi-objective optimization. In this

paper, three artificial neural networks, LSTM, GRU, and TCN, which have made a splash in time-series prediction, are used as individual prediction models and applied to the study of multi-step prediction $PM_{2.5}$. This paper uses GPR as the combined model and optimizes its hyperparameters with the MOGWO algorithm to nonlinearly combine the individual prediction models to improve accuracy and stability. The topology of each neural network and the flow of the MOGWO algorithm are shown in Fig. 2.

2.2.1. Long short-term memory

Hochreiter and Schmidhuber proposed LSTM in 1997, which is considered an excellent variant of RNN [44]. LSTM introduces cell states as storage units to store historical information and adds three gates: input gate, forgetting gate, and output gate to filter. This gives the LSTM a long-time memory capability, allowing it to handle non-linear time series efficiently and solving some of the shortcomings of RNN, such as gradient reduction or gradient explosion [45].

2.2.2. Gated recurrent unit neural network

Similar to LSTM, the structure of GRU combines input gates, forgetting gates, cells, and hidden states. Therefore, GRU can be a simpler gating mechanism than LSTM, requiring fewer parameters and converging more easily [46]. In addition, GRU is better at remembering recent knowledge rather than information from the distant past so that more recent data points are automatically more predictive than older ones [47]. GRU and LSTM, as excellent variants of RNN, are indistinguishable in their performance on different tasks. As individual prediction models, the structural differences between LSTM and GRU can help them learn more valid and extra information from the $PM_{2.5}$ concentration subseries, improving the accuracy and stability of the combined prediction framework.

2.2.3. Temporal convolutional network

In neural network-based timing studies, LSTM and GRU based on classical RNN structures, for example, are usually used, while convolutional network structures are often applied to image processing. Related studies have demonstrated that convolutional networks can perform better than RNN structures for tasks such as machine translation and audio processing [48]. One of the convolutional networks for time series problems, the temporal convolutional network (TCN), has been proposed.

TCN uses causal convolution to ensure that future information is not compromised. In addition, to learn long time series dependencies, TCN introduces extended convolution to reduce the depth of simple causal convolution. TCN can change the perceptual field by increasing the number of layers and changing the expansion coefficient, which provides more flexibility in the length of historical information and avoids the problems of gradient dispersion and gradient explosion in RNN. Therefore, compared with LSTM and GRU, adding TCN as an individual prediction model can learn more information from the data and improve the prediction model's accuracy and robustness.

2.2.4. Gaussian process regression

Gaussian process regression (GPR) is a machine learning method based on Bayesian and statistical theories. GPR has good adaptability and generalization performance in dealing with complex classification and regression problems such as small samples and non-linearities [49]. Therefore, this paper uses the GPR model to combine LSTM, GRU, and TCN to obtain more accurate prediction results. With the continuous development and improvement of GPR model research, researchers have found that suitable hyperparameters can reduce the number of GPR iterations and improve prediction accuracy. The covariance function in GPR, also known as the kernel function, is the main source of hyperparameters for GPR models. The common kernel functions mean exponential (SE) covariance function, quadratic rational (RQ)

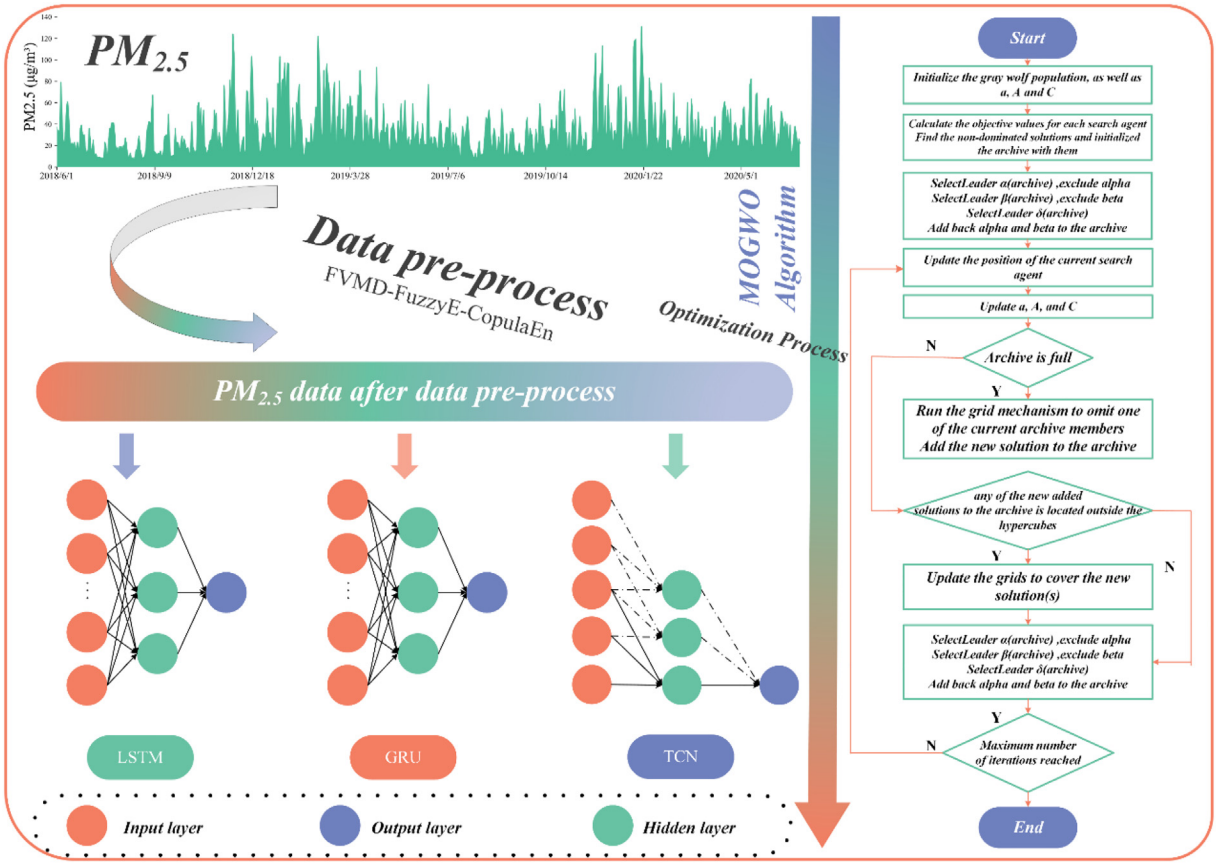


Fig. 2. Structure of individual models and MOGWO algorithm.

covariance function, and Matern covariance function are calculated as shown below.

$$k_{SE}(x_i, x_j) = \exp\left(-\frac{d(x_i, x_j)^2}{2l^2}\right) \tag{5}$$

$$k_{RQ} = \left(1 + \frac{d(x_i, x_j)^2}{2\alpha l^2}\right)^{-\alpha} \tag{6}$$

$$k_{Matern}(x_i, x_j) = \frac{1}{\Gamma(v)2^{v-1}}\left(\frac{\sqrt{2v}}{l}d(x_i, x_j)\right)^{v} K_v\left(\frac{\sqrt{2v}}{l}d(x_i, x_j)\right) \tag{7}$$

Where $d(\cdot,\cdot)$ denotes the Euclidean distance, $K_v(\cdot)$ is the modified Bessel function, and $\Gamma(\cdot)$ is the gamma function. Where α is the scale mixture parameter, which is the shape parameter of the kernel function. And l is the length scale of the kernel, which is the correlation determination hyperparameter, and the larger its value, the smaller the correlation between input and output. Compared with a single kernel function, the performance of the combined kernel function is better [50]. The combined kernel functions in this paper are shown below.

$$k_{combined}(x_i, x_j) = k_{SE} + k_{RQ} \tag{8}$$

2.2.5. Multi-Objective Grey Wolf Optimizer

Multi-Objective Grey Wolf Optimizer (MOGWO) is a multi-objective optimization algorithm based on GWO, proposed by Mirjalili et al. [51]. It is a heuristic algorithm that simulates the social hierarchy and hunting of grey wolves. GWO regards each individual in the population as a solution and treats the current optimal, superior, and suboptimal solutions as corresponding to the leading wolves: α, β, δ and the rest of the individuals defined ω, all of which are the lowest-level individuals and obey the leadership of other high-ranking grey wolves. GWO simulates the collective hunting behaviour of grey wolves, including leading, encircling prey, updating position, and hunting. GWO has the advantages of fewer parameters, easy convergence, and not easily falling into local optimality. To make GWO applicable to multi-objective optimization, the following two techniques are incorporated

(1) Pareto archive

Each iteration of GWO generates an optimal individual, and the Pareto archive is used to store these non-dominated Pareto optimal solutions. The Pareto archive may contain more and more individuals during continuous updating, so the Pareto archive usually sets an upper limit. In order to maintain the diversity of individuals, when the individuals in the Pareto archive exceed the upper limit, similar individuals are eliminated according to the magnitude of the crowding to reduce the number of individuals.

(2) Selection of lead wolves

In the original GWO, the three lead wolves can be selected based on the fitness value of the objective function. In contrast, in multi-objective optimization, the merits of individuals are determined by the Pareto dominance relationship instead of differentiating by simple function values. Therefore, the leader wolf selection mechanism is redefined using a roulette wheel to select individuals from the Pareto archive as leader wolves, while the crowding distance of individuals is associated with their probability of being selected.

2.2.6. MOGWO optimized GPR

Choosing appropriate hyperparameters for the GPR kernel function can improve the model's prediction accuracy and generalization performance. The existing literature on GPR usually uses classical single-objective optimization algorithms such as the conjugate gradient method or PSO for optimization. Single-objective optimization algorithms such as PSO have better global search capability and can find the optimal global solution with high probability. However, the evaluation of a single fitness function can hardly reflect the model's actual performance. There may be room to improve the model's prediction performance in another fitness function when the current fitness function makes the model reach the optimal performance. Therefore. In this paper, the MOGWO algorithm is introduced into the optimization problem of GPR hyperparameters. As shown in Eqs. (5) and (6), this paper uses MOGWO to search and optimize the three hyperparameters of l_{SE}, α_{RQ} and l_{RQ}, and in the combined kernel function. The search interval of all three hyperparameters is [0,10].

In this paper, two sets of fitness functions f_1 and f_2 of MOGWO are constructed with reference to two model performance evaluation metrics, the average absolute percentage error (MAPE) and the Nash efficiency coefficient (NSE), which are calculated as shown below.

$$\begin{cases} f_1 = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \\ f_2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \end{cases} \tag{9}$$

Where y_i denotes the training set observations and $\hat{y}_i$ denotes the training set fit. In addition, the population size of MOGWO is 25, the size of the Pareto archive is 20, and the number of iterations of the algorithm is set to 100.

2.3. Stage 3: non-linear integration and air pollution warning

The first and second phases of the air pollution forecasting and warning system are performed with adaptive optimal feature extraction and combined prediction for $PM_{2.5}$ concentration data, respectively. Related studies show that LSTM has powerful non-linear data processing capability and good stability. Therefore, in the third stage, LSTM is chosen as a non-linear integration model to integrate the prediction results of the reconstructed components nonlinearly.

With economic development and industrialization, the United States proposed a fine particulate matter standard in 1997 to monitor the increasing concentration of fine particulate matter ($PM_{2.5}$). Since then, the $PM_{2.5}$ concentration index has become an extremely important index to measure the degree of air pollution. To protect human health and further control air pollution, China issued the Ambient Air Quality Standard in 2012 and adopted it as a national environmental quality standard, implemented throughout China in 2016. The standard separates the ambient air functional zones into two types: the first category is for areas requiring special protection, and the second is for residential and industrial areas. It also sets out quality requirements for the ambient air functional areas, and the $PM_{2.5}$ concentration limit value standards are shown in Table 1.

After obtaining accurate $PM_{2.5}$ concentration prediction values, this paper assesses and warns the air pollution level based on the Ambient Air Quality Standards and provides reference opinions for preventing and controlling air pollution.

3. Case analysis

3.1. Data source

China, the world's largest developing country, suffers unprecedented air pollution as it accelerates its economic development. In this study, the daily average $PM_{2.5}$ concentration data of Shanghai and Guangzhou were selected as the study samples, and the data were obtained from the China Meteorological Data Network (http://data.

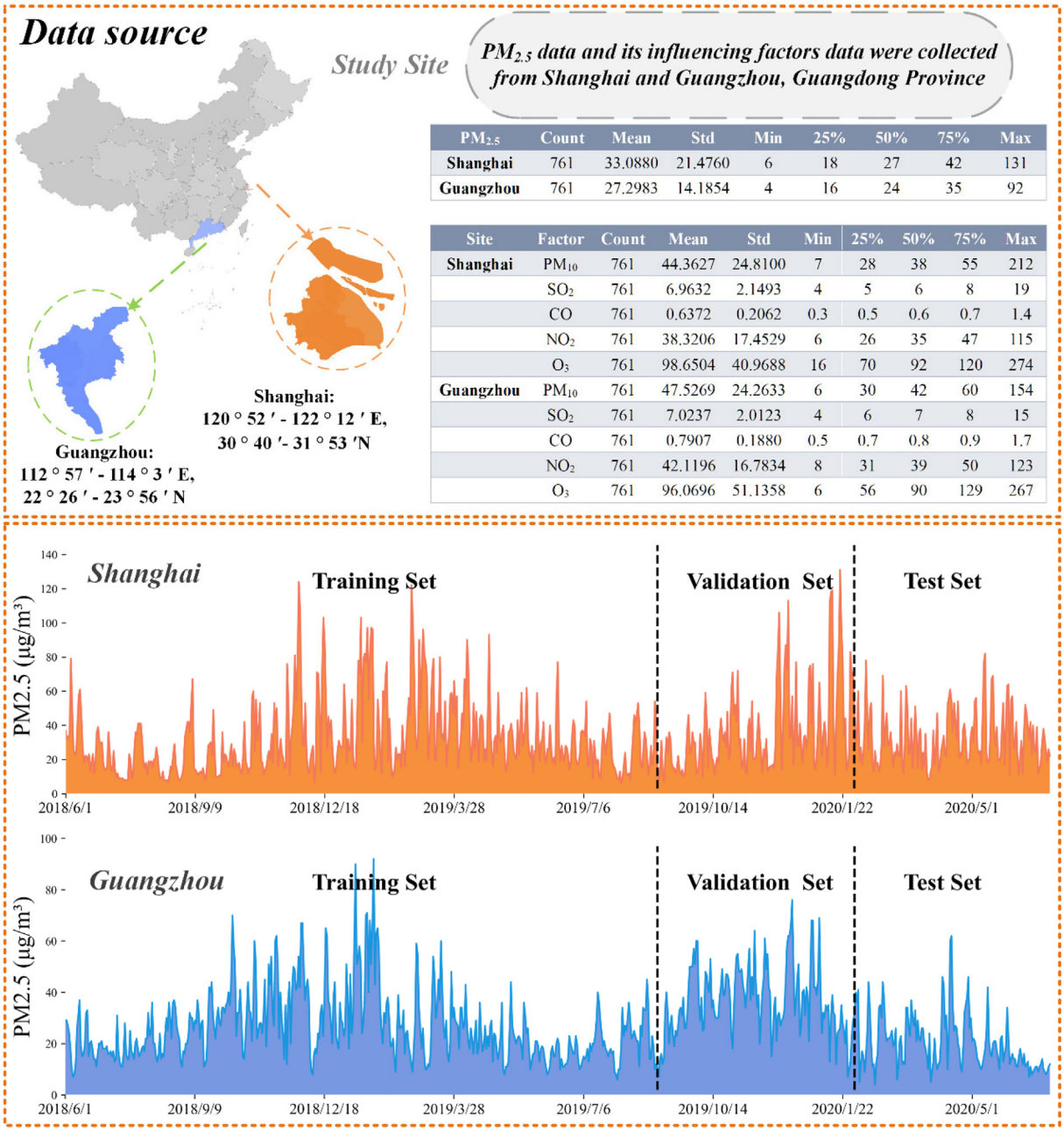
Table 1
Air pollutant concentration limit.

Air pollutants	Average time	Concentration limit		Unit
		Level I	level II	
$PM_{2.5}$	Annually average	15	35	$\mu g/m^3$
	24-hour average	35	75	

cma.cn/). These two cities are economically advanced and have a large population, so timely and accurate air pollution forecasts and warnings can effectively protect the health of city residents. The sampling range of the daily average $PM_{2.5}$ concentration data for both Guangzhou and Shanghai are from June 1, 2018, to June 30, 2020. In addition, five other air pollutants, namely PM_{10}, SO_2, CO, NO_2, and O_3, were sampled in this study as potential effects of fluctuations in $PM_{2.5}$ concentrations for prediction studies. The dataset is divided into training, validation, and test sets, which account for 60%, 20%, and 20%, respectively. Fig. 3 shows the geographical location and $PM_{2.5}$ concentration fluctuations of the two cities. Moreover, the statistical information of the sample data is also displayed in Fig. 3, including total, mean, sample difference, maximum, quantile, and minimum values. In order to better observe the correlation between the trend of $PM_{2.5}$ concentration and the trend of potential influencing factors, the line graphs are plotted in Fig. 4.

In neural networks, dimensionless can convert data of different magnitudes into data uniform scales. The dimensionless normalization can improve the speed of model convergence to a certain extent and avoiding the impact of singular values on model calculations. In this study, the typical normalization method is used to preprocess the data, and the calculation formula is as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{10}$$



$PM_{2.5}$	Count	Mean	Std	Min	25%	50%	75%	Max
Shanghai	761	33.0880	21.4760	6	18	27	42	131
Guangzhou	761	27.2983	14.1854	4	16	24	35	92

Site	Factor	Count	Mean	Std	Min	25%	50%	75%	Max
Shanghai	PM_{10}	761	44.3627	24.8100	7	28	38	55	212
	SO_2	761	6.9632	2.1493	4	5	6	8	19
	CO	761	0.6372	0.2062	0.3	0.5	0.6	0.7	1.4
	NO_2	761	38.3206	17.4529	6	26	35	47	115
	O_3	761	98.6504	40.9688	16	70	92	120	274
Guangzhou	PM_{10}	761	47.5269	24.2633	6	30	42	60	154
	SO_2	761	7.0237	2.0123	4	6	7	8	15
	CO	761	0.7907	0.1880	0.5	0.7	0.8	0.9	1.7
	NO_2	761	42.1196	16.7834	8	31	39	50	123
	O_3	761	96.0696	51.1358	6	56	90	129	267

Fig. 3. Data description for the Guangzhou and Shanghai samples.

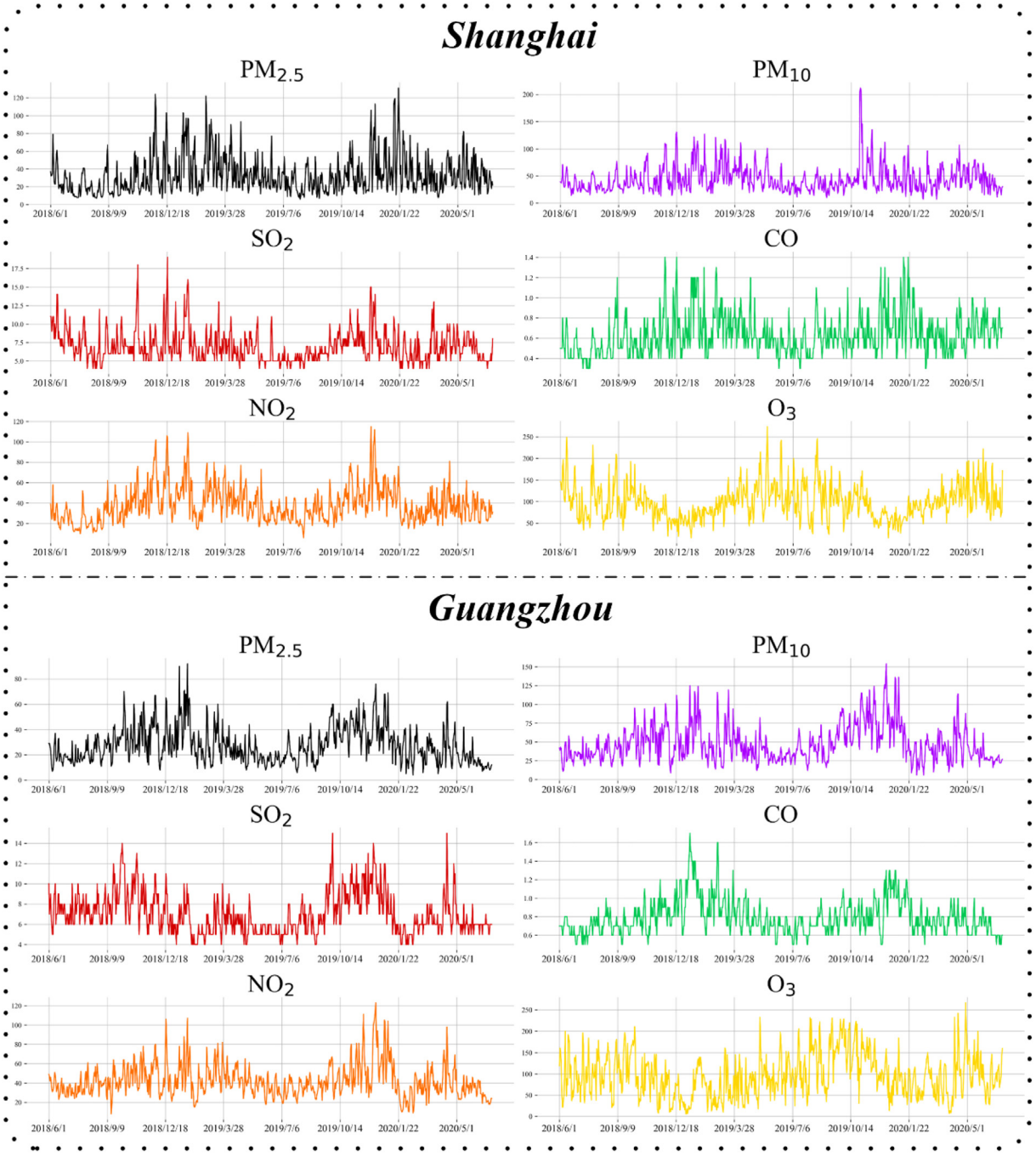


Fig. 4. $PM_{2.5}$ and related factors in Shanghai and Guangzhou.

Where x represents the original data, x' is the normalized result, and the maximum and minimum values of the original data are represented by $\max(x)$ and $\min(x)$, respectively.

3.2. Non-linearity and non-smoothness tests for time series

Before beginning the time series analysis, the non-linearity and non-stationarity of the $PM_{2.5}$ concentration must be confirmed. This paper used the two most commonly used tests, Augmented Dickey-Fuller (ADF) and Brock-Dechert-Scheinkman (BDS). The results of the ADF test are shown in Table 2, and the p-values for both city data sets are less than 0.05. Also, the original hypothesis is strictly rejected at the 1% level, so both cities' $PM_{2.5}$ series are non-stationary.

The embedding dimension of the BDS test is set to 5 in this paper. Table 3 displays the results of the BDS test. The test findings reveal that all z statistics are significantly higher than the critical values within the 95% confidence interval, and the p-values are less than 0.05. The test proves that the time series of $PM_{2.5}$ concentration has non-linear characteristics. Considering the non-linear and non-smooth characteristics of $PM_{2.5}$ concentration, LSTM, GRU, and TCN, which can effectively

Table 2
The result of ADF test in $PM_{2.5}$ concentration time series.

Cases	t-statistic	*P*-value
Shanghai	−4.352	0.0003
Guangzhou	−3.515	0.1119

Table 3
The results of BDS test in $PM_{2.5}$ concentration time series.

Cases	Statistic	Z-statistic	P-value	95%CI
Shanghai	BDS (2)	16.064	0.000	[−1.96,1.96]
	BDS (3)	15.994	0.000	[−1.96,1.96]
	BDS (4)	15.394	0.000	[−1.96,1.96]
	BDS (5)	15.367	0.000	[−1.96,1.96]
Guangzhou	BDS (2)	28.777	0.000	[−1.96,1.96]
	BDS (3)	28.764	0.000	[−1.96,1.96]
	BDS (4)	29.026	0.000	[−1.96,1.96]
	BDS (5)	29.341	0.000	[−1.96,1.96]

handle the non-linear characteristics, are selected as the three individual prediction models in this paper.

3.3. Optimal feature extraction

This study proposes a feedback VMD method to decompose $PM_{2.5}$ concentrations. As described in Section 2, the FVMD algorithm does not require a predetermined number of mode decompositions and can perform decomposition adaptively. As shown in Fig. 5-A and 6-A, FVMD decomposes the $PM_{2.5}$ concentration time series of Shanghai and Guangzhou into 6 and 5 modes, respectively, and obtains a residual series. Table 4 presents the statistical information of the original $PM_{2.5}$ series, the decomposed individual modes, and the residual series with mean and variance for the two cities.

To verify the effectiveness of the proposed FVMD method in adaptive decomposition, the number of mode decompositions is set from 2 to 10 in this paper, and the VMD decomposition is performed on the $PM_{2.5}$ concentration series of two cities. As shown in Fig. 5-B and 6-B, the histograms of the mean values of the similarity coefficients between the decomposed patterns and the original series obtained from the decomposition under different pattern decompositions are demonstrated. In addition, Fig. 5-C and 6-C show the decreasing trend of the mean value of the similarity coefficient when the number of pattern decompositions increases. Taking the Shanghai dataset as an example, as shown in Fig. 5-B, the mean value of the similarity coefficient decreases continuously when the number of VMD pattern decompositions increases. As shown in Fig. 5-C, when the number of decompositions is small ($k = 2$), the decreasing speed of the mean value of the similarity coefficient decreases from fast to slow as the number of decompositions increases. And the rate of decrease in the mean value of the similarity coefficient increases abruptly when $k \geq 7$, followed by a smoothness again. This indicates that when the number of decompositions exceeds 7, further increasing decompositions make the decomposition less efficient. It is difficult to extract more effective information and may

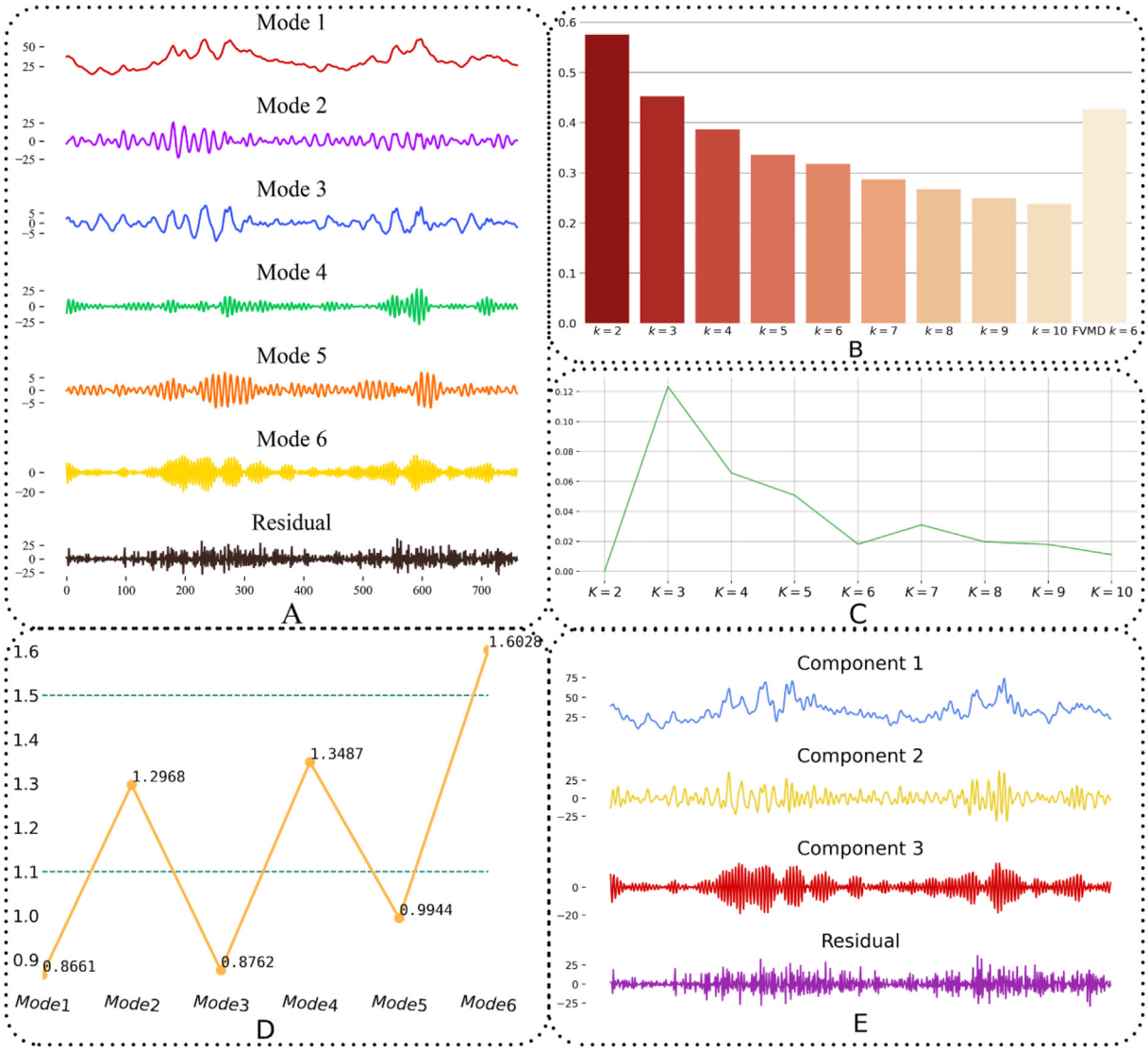


Fig. 5. The feature extraction process of the Shanghai dataset.

Table 4
The mean and variance of each mode and residual.

Cases	Statistic	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6	Residual
Shanghai	Mean	33.08	0.0008	0.002	0.0001	0.0002	0.0004	0.003
	Variance	102.80	42.27	7.83	32.19	4.55	31.99	76.45
Guangzhou	Mean	27.29	0.0005	0.0009	0.00008	0.00003	/	0.002
	Variance	81.13	20.43	2.17	6.21	9.23	/	25.67

capture additional noise. Therefore, the number of decompositions should be 5 or 6 when using VMD to decompose the Shanghai $PM_{2.5}$ series. In contrast, the FVMD algorithm proposed in this paper adaptively decomposes the Shanghai $PM_{2.5}$ series into six subsequences, consistent with the theoretical analysis. The feedback-type mechanism of FVMD makes the decomposed subsequences obtain higher mean values of similarity coefficients. Therefore, FVMD has a better decomposition performance than VMD.

According to FuzzyEn, reconstructing sequences with similar complexity can effectively reduce the computational complexity of the prediction model and improve the prediction performance. As shown in Fig. 5-D and Fig. 6-D, mode1, mode3, mode5 and mode2, mode4, and mode6 of the Shanghai dataset can be reconstructed as Component1, Component2, and Component3, respectively, according to the FuzzyEn values of different modes. Mode1, mode3 and mode2, mode4, and mode5 of the Guangdong dataset can be reconstructed as Component1, Component2, and Component3, respectively. Fig. 5-E and 6-E show the reconstruction results with residual sequences for the two cities.

3.4. Influencing factors selection

In this section, five other air pollutant data were selected as potential influencing factors for $PM_{2.5}$. The CopulaEn algorithm analyzed the degree of correlation between different features and fluctuations in $PM_{2.5}$ concentrations. Table 5 shows the values of CopulaEn for different influencing factors and $PM_{2.5}$ concentrations, and the magnitude of CopulaEn values represents the degree of correlation. In order to reduce the complexity of the model calculation, the top three influencing factors

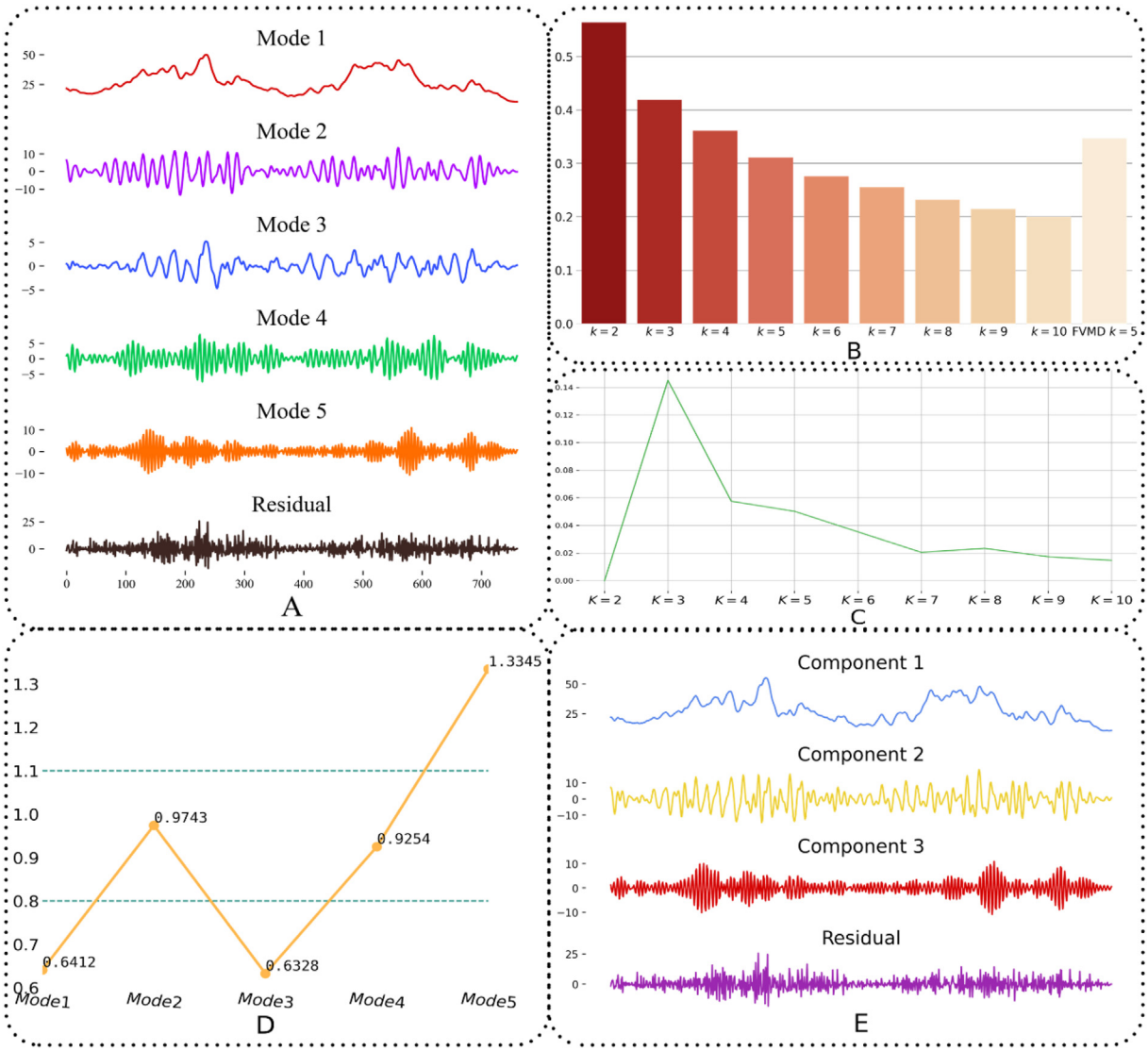


Fig. 6. The feature extraction process of the Guangzhou dataset.

Table 5
CopulaEn value of influencing factors.

Cases	PM_{10}	SO_2	CO	NO_2	O_3
Shanghai	0.4198	0.1754	0.5592	0.3255	0.0379
Guangzhou	1.1164	0.2439	0.2354	0.3304	−0.0408

in terms of CopulaEn values are selected as the input features of the model in this paper.

3.5. Evaluation indicators

In this paper, five evaluation indicators are used to assess the models' predictive performance, and these evaluations have been widely used in time series forecasting studies [52,53]. These five metrics are mean absolute error (MAE), root mean square error (RMSE), mean absolute error (MAPE), median absolute percentage error (MdAPE), and Nash Sutcliffe Efficiency (NSE). The NSE is generally used to verify the performance of model results such as hydrometeorology. If the NSE is closer to 1, it indicates that the model quality and credibility are better; if the NSE is close to 0, it indicates that the overall results of the model are credible, but the prediction error is large; if the NSE is much less than 0, it indicates that the model is not credible. Moreover, if the other four evaluation indexes are smaller, the better the model performance is indicated. The formulae for the five evaluation metrics are shown in Table 6, where $\widehat{y}$ represents the observed value, $\widehat{y}_i$ represents the predicted value, and N represents the sequence length.

4. Results and discussions

4.1. Individual prediction models

4.1.1. Time step

Time step length, also known as time lag, plays an important role in LSTM, RNN, and TCN. The time step length defines how many time-stamped data points should be incorporated as temporal neural network input data. Determining a reasonable time step can effectively improve the model's computational efficiency and prediction accuracy. Through extensive experiments and tests on three models, LSTM, GRU, and TCN, it is found that excellent prediction performance can be achieved when the time step is set to 4. Finally, the time step of all individual prediction models is set to 4.

4.1.2. Multi-step short-term forecasting

In this paper, three temporal neural networks, LSTM, GRU, and TCN, are used as individual prediction models to make short-term predictions of $PM_{2.5}$ concentrations for one, two, and three days in the future.

In terms of neural network construction, after continuous testing and experiments, it is found that the three-layer stacked temporal neural network has the optimal prediction performance. In addition, the number of neurons has a direct impact on the fitting ability of the neural network. The size of the batch size has a close relationship with the weight update of the neural network, and a suitable batch size can significantly improve the convergence speed and prediction performance of the neural. In this paper, we adjust the hyperparameters of different models and continuously optimize the models to get the individual prediction model that fits the training set best. Tables 6 and 8 show the 1-step prediction performance of the three different individual models for four subsets of series in two cities. Taking the 1-step forecasting of Shanghai data as an example, three different forecasting models behave differently in different data. In predicting Component 1, the best accuracy is GRU, and the difference in prediction performance among the three models is small. Although the LSTM does not achieve the best prediction accuracy, it maintains good accuracy in different data and has the best prediction stability. Therefore, it is difficult for a single prediction model to achieve optimal performance. The combination of multiple prediction models may further improve the prediction accuracy and stability of the model.

Table 6
Evaluation metrics.

Metrics	Definition	Equation
MAE	Mean absolute error	$MAE = \frac{1}{N}\sum_{i=1}^{N}\left\|y_i - \widehat{y}_i\right\|$
RMSE	Root mean square error	$RMSE = \sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}\left(\widehat{y}_i - y_i\right)^2}$
MAPE	Mean absolute percentage error	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\widehat{y}_i - y_i}{y_i}\right\| \times 100\%$
MdAPE	Median of absolute percentage error	$MdAPE = median\left(\left\|\frac{y_i - \widehat{y}_i}{y_i}\right\| \times 100\%\right)$
NSE	Nash Sutcliffe Efficiency	$NSE = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}_i\right)^2}$

4.2. Combined model prediction

To improve the predictive performance and robustness of the model, this paper utilizes the GPR model as a combinatorial model to perform a non-linear combination of the prediction results of the individual models. Three trained individual models are used to fit the training set, and the fitted values of the training set are used as the input values of the GPR model, and the actual values in the training set are used as the output values. The selection of hyperparameters of the GPR model greatly affects the prediction performance of the GPR model, and the covariance kernel function is the main source of hyperparameters of the GPR model. Therefore, this paper uses the MOGWO algorithm to perform multi-objective optimization of the hyperparameters of the kernel function of the GPR model to make the GPR have better combined predictive performance and stability. After training, GPR can obtain the final prediction results by adaptively and nonlinearly combining the prediction data according to the characteristics of the prediction data of different individual models. Tables 7 and 8 show the 1-step prediction performance of the three individual models and the combined 1-step prediction performance of the MOGWO-GPR model. The MOGWO-GPR model combines the prediction results of the three models nonlinearly, which has a significant improvement in different data and different evaluation indexes and effectively improves the prediction performance and stability.

4.3. Non-linear integration model

After forecasting each reconstituted subsequence using the combined forecasting model, all the forecasts for the same city are pooled for the final integration. The LSTM, with its powerful non-linear forecasting capability and good forecasting stability, is used as the integration model for the non-linear integration of the forecasts. The final multi-step forecasting results and performance for two cities are shown in Figs. 7 and 8 and Tables 9 and 10. The experimental results show that the hybrid combination prediction model has good prediction stability and accuracy in different situations and can effectively predict the concentration of $PM_{2.5}$.

4.4. Nonlinear integration and air quality warning

After forecasting each reconstructed component using the combined forecasting model, the final non-linear integration of all forecasts was performed. The final results of the two cities' multi-step predictions are shown in Figs. 7 and 8, and the prediction performance is shown in Tables 7 and 8.

The performance of the proposed combined prediction hybrid framework performs well in predicting the trend of future $PM_{2.5}$ concentrations and is suitable for 1-step ahead prediction. In multi-step prediction, less historical information, which brings problems

Table 7
Single-step prediction results of individual and combined models for the Shanghai dataset.

Sequences	Models	MAE	RMSE	MAPE	MdAPE	NSE
Component 1	LSTM	0.005097	0.005829	1.607268	0.014464	0.995456
	TCN	0.005395	0.006308	1.765825	0.015676	0.994679
	GRU	0.005711	0.005711	1.535625	0.014207	0.995638
	MOGWO-GPR	0.003249	0.003965	1.063083	0.009329	0.997898
Component 2	LSTM	0.007131	0.008309	1.626767	0.015599	0.992176
	TCN	0.005653	0.006669	1.295639	0.011569	0.994959
	GRU	0.008488	0.009449	1.969020	0.019626	0.989880
	MOGWO-GPR	0.004116	0.004969	0.947572	0.008443	0.997202
Component 3	LSTM	0.015773	0.017949	3.231277	0.030169	0.973574
	TCN	0.022099	0.025527	4.282016	0.038876	0.946557
	GRU	0.015678	0.018312	3.102545	0.030751	0.972498
	MOGWO-GPR	0.010299	0.012356	2.026295	0.017198	0.987478
Residual	LSTM	0.035957	0.040914	11.911109	0.074682	0.916611
	TCN	0.046738	0.054366	14.176702	0.099924	0.852759
	GRU4	0.037840	0.044893	9.898643	0.087130	0.899601
	MOGWO-GPR	0.020935	0.025434	5.707214	0.040975	0.967773

such as error accumulation and increased uncertainty, may reduce prediction accuracy. However, the combined prediction method proposed in this paper absorbs the advantages of the three individual models, reduces redundant information, and makes better prediction of $PM_{2.5}$ trends even in multi-step prediction, resulting in a significant improvement in prediction accuracy and stability in different data sets.

In this paper, based on accurate short-term forecasts of $PM_{2.5}$ concentrations, air quality is assessed and warned based on the Ambient Air Quality Standards implemented in China. Referring to the limitation standards in Table 1, Shanghai met the Class 1 standard on 92 days, the Class 2 standard on 53 days, and the air pollutants were below the Class 2 environmental functional area standard on three days out of the 148 days of the test data from February 4, 2020, to June 30, 2020. On the other hand, Guangzhou City met the Class 1 standard on 134 of these 148 days and the Class 2 standard on 14 days. This indicates that Shanghai's air quality needs to be improved, while Guangzhou's air quality is better. As shown in Tables 9 and 10, the hybrid model warns the air quality of Shanghai for the next one, two, and three days, and the early warning accuracy reaches 99%, 91%, and 89%, respectively. Furthermore, the short-term warning for Guangzhou air quality achieves 95%, 95%, and 90% accuracy, respectively. Therefore, the air pollutant forecasting and warning framework provide effective short-term predictions of pollutant concentrations and air quality warnings.

4.5. Comparison

To verify the effectiveness and stability of the hybrid prediction framework proposed in this paper, six comparative models are developed in this paper, with references to outstanding papers in related research fields. These models are the advanced research results in recent years and are typical. The comparison results for the data of the two cities are shown in Tables 9 and 10 and Figs. 7 and 8.

Compared with all the comparison models, the combined prediction hybrid framework proposed in this study has a significant advantage in the accuracy of air pollutant prediction and warning, and the comparison results are displayed in Tables 7 and 8. Using individual models such as Random Forest (RF) [54], LSTM [55], etc., although they can effectively predict $PM_{2.5}$ concentrations, the prediction accuracy and warning accuracy are inferior to other combined models. Moreover, the performance of individual models varies widely and is not stable in different city datasets. In contrast, PSO-SVR [56] used an optimization algorithm to optimize the SVR, which enabled the model to obtain better performance in multi-step prediction. Furthermore, the EMD-GRU [57] and EEMD-LSTM [58] models used a decomposition integration strategy, which significantly improved model prediction accuracy and performance. In the Shanghai dataset, the MAPEs of EMD-GRU multi-step prediction were 34.25%, 47.04%, and 57.08%, respectively. This indicates that the decomposition integration strategy can effectively improve the short-term 1-step forecasting accuracy, and less historical information makes the forecasting accuracy lower in multi-step forecasting. In addition, the multi-step warning accuracy of EMD-GRU in the Shanghai dataset is 48%, 45%, and 44%, respectively. While in the Guangzhou dataset, the multi-step warning accuracies are 71%, 75%, and 72%, respectively. This indicates that although the prediction accuracy of the EMD-GRU model has improved, the model does not have better stability and generalization performance. VMD-SampleEn-LSTM

Table 8
Single-step prediction results of individual and combined models for the Guangzhou dataset.

Sequences	Models	MAE	RMSE	MAPE	MdAPE	NSE
Component 1	LSTM	0.002841	0.003537	4.556866	0.017825	0.999268
	TCN	0.002623	0.002965	5.526683	0.014578	0.999486
	GRU	0.002999	0.003682	5.655767	0.018557	0.999207
	MOGWO-GPR	0.002676	0.003266	3.606455	0.016713	0.999376
Component 2	LSTM	0.006274	0.008101	2.068086	0.011765	0.996844
	TC	0.007156	0.008351	1.889907	0.015353	0.996646
	GRU	0.008385	0.010044	2.485441	0.018185	0.995149
	MOGWO-GPR	0.004301	0.005383	1.279005	0.008444	0.998606
Component 3	LSTM	0.018258	0.023292	4.420256	0.029408	0.969963
	TCN	0.007128	0.008428	1.625969	0.014934	0.996067
	GRU	0.004908	0.006558	1.100453	0.007509	0.997618
	MOGWO-GPR	0.004458	0.005563	0.975404	0.007951	0.998286
Residual	LSTM	0.043541	0.048467	11.374635	0.110915	0.755691
	TCN	0.031309	0.035525	8.198470	0.073715	0.868745
	GRU	0.036089	0.042091	9.252835	0.087672	0.815743
	MOGWO-GPR	0.019458	0.024228	5.060697	0.040113	0.938953

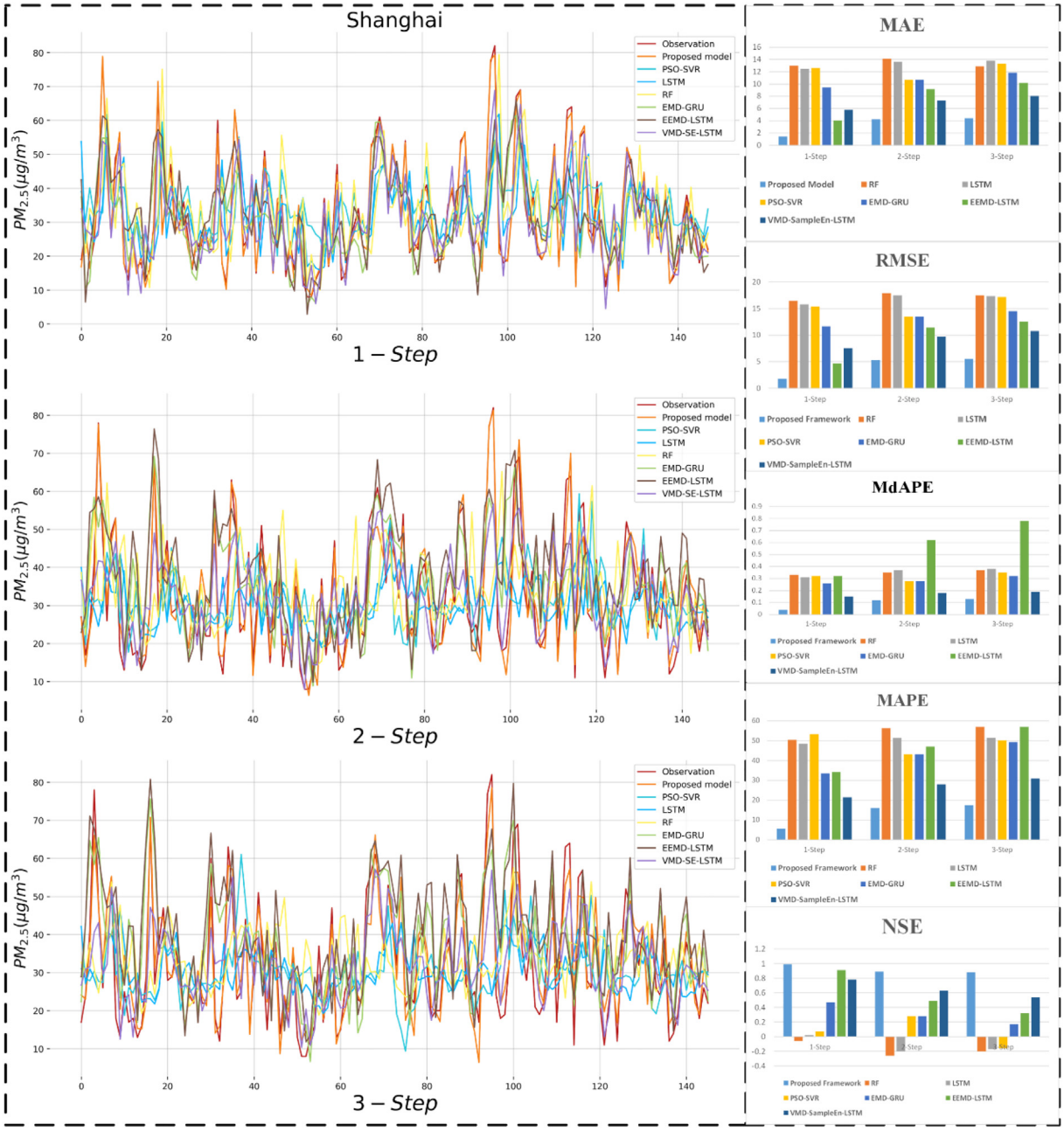


Fig. 7. Multi-step prediction results for the Shanghai dataset.

[59] has better overall performance compared with the other five comparison models. In the Guangzhou dataset, the MAPEs of multi-step prediction were 19.26%, 22.25%, and 27.54%, and the accuracy of multi-step warning was 91%, 91%, and 90%, respectively. Similar to the drawbacks of other compared models, although the performance of VMD-SampleEn-LSTM in multi-step prediction is improved, the prediction performance and stability vary widely in different datasets, and the warning accuracy in the Shanghai dataset is only 54%, 51%, and 48%.

The proposed combined forecasting hybrid framework in this paper has MAPEs of 5.57%, 16.06%, and 17.51% for the Shanghai dataset, and the accuracy of early warning is 99%, 91%, and 89%, respectively. In the Guangzhou dataset, the MAPEs are 5.91%, 8.02%, and 12.64%, and the accuracy of early warning is 95%, 95%, and 90%, respectively. This indicates that the combined forecasting hybrid framework proposed in this paper fully absorbs the advantages of each model. Although the prediction effect decreases as the number of advance prediction steps increases, It still has good prediction accuracy and robustness across various datasets.

5. Conclusion and future directions

In this paper, an air pollutant prediction and early warning system are developed based on a multi-objective optimal combined prediction hybrid framework. This combined prediction hybrid framework effectively utilizes data processing techniques and combined prediction with multi-objective optimization further to improve the prediction performance of short-term $PM_{2.5}$ concentrations. Specifically, the advantages of this paper's combined prediction hybrid framework are mainly reflected in the following aspects. (1) A feedback VMD decomposition method is proposed in this paper, which can determine the number of signal decompositions adaptively. (2) The feature extraction method based on VMD-FuzzyEn extracts effective information from the $PM_{2.5}$ concentration series, reducing the computational complexity of

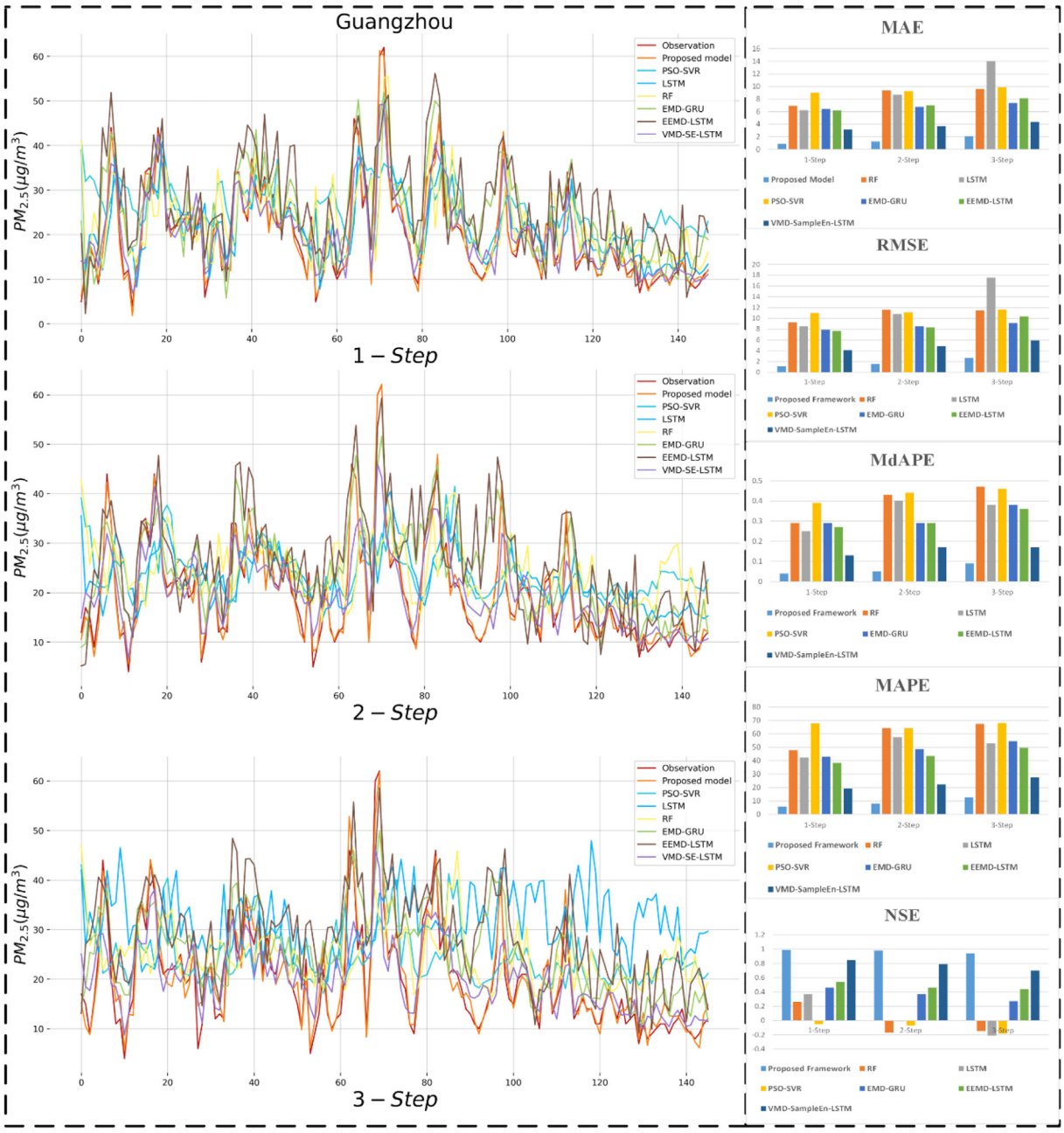


Fig. 8. Multi-step prediction results for the Guangzhou dataset.

the model and improving prediction accuracy. (3) The influencing factors of $PM_{2.5}$ are introduced into the study. The factors that strongly influence $PM_{2.5}$ are selected using CopulaEn, which reduces the redundancy and improves the model's prediction accuracy and generalization performance. (4) Three different individual models, LSTM, GRU, and TCN, were used to predict $PM_{2.5}$ effectively. The multi-objective optimized GPR model was used to absorb the advantages of the three individual models and effectively combine the prediction results of the individual models in a non-linear manner further to improve the model's predictive performance and robustness. (5) The predicted values are nonlinearly integrated using LSTM to the short-term predicted value of $PM_{2.5}$. (6) Based on the effective short-term prediction results of the combined prediction framework, air quality warnings were performed, and good warning accuracy was obtained. The experimental results demonstrate that the combined prediction hybrid framework developed in this paper has better prediction performance and robustness compared to the six comparison models. Therefore, this combined prediction hybrid framework can be effective for air pollutant prediction and early warning.

In summary, the hybrid framework developed in this paper has an excellent performance in both accuracy and stability of prediction. The application of the combined prediction hybrid model is not limited to air pollutant forecasting and early warning but also can be applied to the financial field and energy fields, such as price forecasting and wind power generation forecasting, through the processing of relevant data. More, with further research, more novel models and algorithms will be developed in the future. In future research, more novel individual models can be added to the combined prediction models to predict air pollutant concentrations. In addition, $PM_{2.5}$ is just one of the air pollutants. This study only considered the prediction of $PM_{2.5}$ concentration and incorporating more types of air pollutants into the prediction study to construct a more novel and effective air pollutant forecasting and warning can be another option for future research.

Table 9
Comparison model results for the Shanghai dataset.

Models		MAE	RMSE	MAPE	MdAPE	NSE	Accuracy
RF	1-Step	12.99	16.46	50.42	0.33	−0.06	37%
	2-Step	14.12	17.93	56.35	0.35	−0.26	34%
	3-Step	12.89	17.49	57.08	0.37	−0.20	35%
LSTM	1-Step	12.50	15.79	48.51	0.31	0.02	38%
	2-Step	13.76	17.47	50.40	0.37	−0.20	31%
	3-Step	13.78	17.33	51.48	0.38	−0.17	31%
PSO-SVR	1-Step	12.63	15.41	53.36	0.32	0.07	39%
	2-Step	13.62	17.50	51.42	0.37	−0.20	36%
	3-Step	13.32	17.18	50.16	0.35	−0.16	38%
EMD-GRU	1-Step	9.42	11.67	33.47	0.26	0.47	48%
	2-Step	10.67	13.49	43.20	0.28	0.28	46%
	3-Step	11.85	14.51	49.31	0.32	0.17	45%
EEMD-LSTM	1-Step	4.06	4.69	34.25	0.32	0.91	48%
	2-Step	9.15	11.41	47.04	0.62	0.49	45%
	3-Step	10.19	12.57	57.08	0.78	0.32	44%
VMD-SampleEn-LSTM	1-Step	5.78	7.57	21.41	0.15	0.78	54%
	2-Step	7.31	9.70	28.04	0.18	0.63	51%
	3-Step	8.04	10.76	30.94	0.19	0.54	48%
Proposed framework	1-Step	1.43	1.79	5.57	0.04	0.99	99%
	2-Step	4.25	5.32	16.06	0.12	0.89	91%
	3-Step	4.40	5.53	17.51	0.13	0.88	89%

CRediT authorship contribution statement

Jujie Wang, Wenjie Xu, Yue Zhang and Jian Dong: conceived of the presented idea, developed the theory and performed the computations, discussed the results, wrote the paper and approved the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 71971122 and 71501101).

References

[1] Hu F, Guo Y. Health impacts of air pollution in China. Front Environ Sci Eng. 2020;15:74.
[2] Zheng S, Wang J, Sun C, Zhang X, Kahn ME. Air pollution lowers Chinese urbanites' expressed happiness on social media. Nat Hum Behav. 2019;3:237–43.
[3] Huang W, Hu M. Estimation of the impact of traveler information apps on urban air quality improvement. Engineering. 2018;4:224–9.
[4] Samal KKR, Babu KS, Das SK. Multi-directional temporal convolutional artificial neural network for PM2.5 forecasting with missing values: a deep learning approach. Urban Clim. 2021.;36(100800).
[5] Lv L, Chen Y, Han Y, Cui M, Wei P, Zheng M, et al. High-time-resolution PM2.5 source apportionment based on multi-model with organic tracers in Beijing during haze episodes. Sci Total Environ. 2021.;772(144766).
[6] J A, Huang M, Wang Z, Zhang X, Ueda H, Cheng X. Numerical regional air quality forecast tests over the mainland of China. Water, Air, and Soil Pollution. 2001;130:1781–6.
[7] Tie X, Madronich S, Li G, Ying Z, Zhang R, Garcia AR, et al. Characterizations of chemical oxidants in Mexico City: a regional chemical dynamical model (WRF-Chem) study. Atmos Environ. 2007;41:1989–2008.
[8] Wang Z, Li J, Wang Z, Yang W, Tang X, Ge B, et al. Modeling study of regional severe hazes over mid-eastern China in January 2013 and its implications on pollution prevention and control. Sci China Earth Sci. 2014;57:3–13.

Table 10
Comparison model results for the Guangzhou dataset.

Models		MAE	RMSE	MAPE	MdAPE	NSE	Accuracy
RF	1-Step	6.92	9.25	47.80	0.29	0.26	68%
	2-Step	9.38	11.60	64.35	0.43	−0.17	51%
	3-Step	9.58	11.49	67.46	0.47	−0.15	52%
LSTM	1-Step	6.23	8.53	42.20	0.25	0.37	68%
	2-Step	8.67	10.80	57.51	0.40	−0.01	53%
	3-Step	14.02	17.57	52.88	0.38	−0.21	32%
PSO-SVR	1-Step	9.01	10.99	67.80	0.39	−0.05	62%
	2-Step	9.23	11.11	64.25	0.44	−0.07	44%
	3-Step	9.85	11.64	68.07	0.46	−0.18	44%
EMD-GRU	1-Step	6.41	7.87	42.98	0.29	0.46	71%
	2-Step	6.76	8.52	48.51	0.29	0.37	75%
	3-Step	7.35	9.13	54.47	0.38	0.27	72%
EEMD-LSTM	1-Step	6.21	7.67	38.32	0.27	0.54	75%
	2-Step	7.01	8.33	43.41	0.29	0.46	72%
	3-Step	8.12	10.34	49.57	0.36	0.44	44%
VMD-SampleEn-LSTM	1-Step	3.15	4.12	19.26	0.13	0.85	91%
	2-Step	3.67	4.87	22.25	0.17	0.79	91%
	3-Step	4.38	5.91	27.54	0.17	0.70	90%
Proposed Framework	1-Step	0.89	1.09	5.91	0.04	0.99	95%
	2-Step	1.23	1.57	8.02	0.05	0.98	95%
	3-Step	2.07	2.66	12.64	0.09	0.94	90%

[9] Ge BZ, Wang ZF, Xu XB, Wu JB, Yu XL, Li J. Wet deposition of acidifying substances in different regions of China and the rest of East Asia: modeling with updated NAQPMS. Environ Pollut. 2014;187:10–21.
[10] James EP, Benjamin SG, Marquis M. Offshore wind speed estimates from a high-resolution rapidly updating numerical weather prediction model forecast dataset. Wind Energy. 2018;21:264–84.
[11] Zhang L, Lin J, Qiu R, Hu X, Zhang H, Chen Q, et al. Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model. Ecol Indic. 2018;95:702–10.
[12] Lesar TT, Filipčić A. The hourly simulation of PM2.5 particle concentrations using the multiple linear regression (MLR) model for sea breeze in Split, Croatia. Water, Air, & Soil Pollution. 2021;232:261.
[13] Asadollahfardi G, Zangooei H, Aria SH. Predicting PM 2.5 concentrations using artificial neural networks and Markov chain, a case study Karaj City. Asian Journal of Atmospheric Environment. 2016;10:67–79.
[14] Rubal, Kumar D. Evolving Differential evolution method with random forest for prediction of Air Pollution. Procedia Computer Science. 2018;132:824–33.
[15] Leong WC, Kelani RO, Ahmad Z. Prediction of air pollution index (API) using support vector machine (SVM). J Environ Chem Eng. 2020;8:103208.
[16] Feng Q, Sun X, Hao J, Li J. Predictability dynamics of multifactor-influenced installed capacity: a perspective of country clustering. Energy. 2021;214:0360–5442.
[17] Zhu B, Han D, Wang P, Wu Z, Zhang T, Wei Y. Forecasting carbon price using empirical mode decomposition and evolutionary least squares support vector regression. Appl Energy. 2017;191:521–30.
[18] Wang J, Zhang W, Li Y, Wang J, Dang Z. Forecasting wind speed using empirical mode decomposition and Elman neural network. Appl Soft Comput. 2014;23:452–9.
[19] Cheng Y, Zhang H, Liu Z, Chen L, Wang P. Hybrid algorithm for short-term forecasting of PM2.5 in China. Atmos Environ. 2019;200:264–79.
[20] Zhu S, Lian X, Liu H, Hu J, Wang Y, Che J. Daily air quality index forecasting with hybrid models: a case in China. Environ Pollut. 2017;231:1232–44.
[21] Zhang W, Qu Z, Zhang K, Mao W, Ma Y, Fan X. A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting. Energ Convers Manag. 2017;136:439–51.
[22] Huang Y, Dai X, Wang Q, Zhou D. A hybrid model for carbon price forecastingusing GARCH and long short-term memory network. Appl Energy. 2021;285:116485.
[23] Wu Q, Lin H. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Sci Total Environ. 2019;683:808–21.
[24] Li J, Hao J, Feng Q. Forecasting China's sovereign CDS with a decomposition reconstruction strategy. Appl Soft Comput. 2021;105:1568–4946.
[25] Li J, Hao J, Feng Q, Sun X, Liu M. Optimal selection of heterogeneous ensemble strategies of time series forecasting with multi-objective programming. Expert Syst Appl. 2021;166:3783–99.
[26] Bates JM, Granger CWJ. The combination of forecasts. J Oper Res Soc. 1969;20:451–68.
[27] Xiao L, Wang J, Hou R, Wu J. A combined model based on data pre-analysis and weight coefficients optimization for electrical load forecasting. Energy. 2015;82:524–49.
[28] Liu Y, Zhang S, Chen X, Wang J. Artificial combined model based on hybrid nonlinear neural network models and statistics linear models—research and application for wind speed forecasting. Sustainability. 2018;10:4601.
[29] Wang J, Hu J. 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. 2015;93:41–56.
[30] Kim Y, Jeong D, Ko IH. Combining rainfall-runoff model outputs for improving ensemble streamflow prediction. J Hydrol Eng. 2006;11:578–88.
[31] Sudholt D, Witt C. Runtime analysis of a binary particle swarm optimizer. Theoretical Comput Sci. 2010;411:2084–100.
[32] Bouleimen K, Lecocq H. A new efficient simulated annealing algorithm for the resource-constrained project scheduling problem and its multiple mode version. Eur J Oper Res. 2003;149:268–81.
[33] Sun X, Hao J, Li J. Multi-objective optimization of crude oil-supply portfolio based on interval prediction data. Ann Oper Res. 2022;309:611–39.
[34] Dragomiretskiy K, Zosso D. Variational mode decomposition. IEEE Trans Signal Process. 2013;62:531–44.
[35] Takeda M, Ina H, Kobayashi S. Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry. J Opt Soc Am. 1982;72:156–60.
[36] Goswami JC, Chan AK, Chui CK. On a spline-based fast integral wavelet transform algorithm. In: Carin L, Felsen LB, editors. Ultra-Wideband, Short-Pulse Electromagnetics 2. Boston, MA: Springer US; 1995. p. 455–63.
[37] Fu W, Wang K, Li C, Tan J. Multi-step short-term wind speed forecasting approach based on multi-scale dominant ingredient chaotic analysis, improved hybrid GWO-SCA optimization and ELM. Energ Convers Manag. 2019;187:356–77.
[38] Wu Q, Lin H. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Sci Total Environ. 2019;683:808–21.
[39] Lian J, Liu Z, Wang H, Dong X. Adaptive variational mode decomposition method for signal processing based on mode characteristic. Mech Syst Signal Process. 2018;107:53–77.
[40] Chen W, Wang Z, Xie H, Yu W. Characterization of surface EMG signal based on fuzzy entropy. IEEE Trans Neural Syst Rehabil Eng. 2007;15:266–72.
[41] Ma J, Sun Z. Mutual information is copula entropy. Tsinghua Sci Technol. 2011;16:51–4.
[42] Hauke J, Kossowski T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae. 2011;30:87–93.
[43] Sun Y, Qin W, Zhuang Z. Nonparametric-copula-entropy and network deconvolution method for causal discovery in complex manufacturing systems. J Intell Manuf. 2021. https://doi.org/10.1007/s10845-021-01751-w.
[44] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80.
[45] Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: a search space odyssey. IEEE Trans Neural Netw Learn Syst. 2016;28:2222–32.
[46] Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8:6085.
[47] Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2017;24:361–70.
[48] James EP, Benjamin SG, Marquis M. Offshore wind speed estimates from a high-resolution rapidly updating numerical weather prediction model forecast dataset. Wind Energy. 2018;21:264–84.
[49] Seeger M. Gaussian processes for machine learning. Int J Neural Syst. 2004;14:69–106.
[50] Ghasemi P, Karbasi M, Zamani Nouri A, Sarai Tabrizi M, Azamathulla HM. Application of Gaussian process regression to forecast multi-step ahead SPEI drought index. Alex Eng J. 2021;60:5375–92.
[51] Mirjalili S, Saremi S, Mirjalili SM, Coelho LDS. Multi-objective grey wolf optimizer: a novel algorithm for multi-criterion optimization. Expert Syst Appl. 2016;47:106–19.
[52] Wang J, Heng J, Xiao L, Wang C. Research and application of a combined model based on multi-objective optimization for multi-step ahead wind speed forecasting. Energy. 2017;125:591–613.
[53] Wang J, Hu J. 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. 2015;93:41–56.
[54] Doreswamy KSH, Y KM, Gad I. Forecasting air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Comput Sci. 2020;171:2057–66.
[55] Yang J, Yan R, Nong M, Liao J, Li F, Sun W. PM2.5 concentrations forecasting in Beijing through deep learning with different inputs, model structures and forecast time. Atmos Pollut Res. 2021;12:101168.
[56] Chen S, Wang J, Zhang H. A hybrid PSO-SVM model based on clustering algorithm for short-term atmospheric pollutant concentration forecasting. Technol Forecast Soc Chang. 2019;146:41–54.
[57] Huang G, Li X, Zhang B, Ren J. PM2.5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. Sci Total Environ. 2021;768:144516.
[58] Bai Y, Zeng B, Li C, Zhang J. An ensemble long short-term memory neural network for hourly PM2.5 concentration forecasting. Chemosphere. 2019;222:286–94.
[59] Wu Q, Lin H. Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. Sustain Cities Soc. 2019;50:101657.