



Two-stage deep learning hybrid framework based on multi-factor multi-scale and intelligent optimization for air pollutant prediction and early warning

Jujie Wang^{1,2} · Wenjie Xu¹ · Jian Dong¹ · Yue Zhang¹

Accepted: 28 February 2022 / Published online: 26 March 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Effective prediction of air pollution concentrations is of great importance to both the physical and mental health of citizens and urban pollution control. As one of the main components of air pollutants, accurate prediction of $PM_{2.5}$ can provide a reference for air pollution control and pollution warning. This study proposes an air pollutant prediction and early warning framework, which innovatively combines feature extraction techniques, feature selection methods and intelligent optimization algorithms. First, the $PM_{2.5}$ sequence is decomposed into several subsequences using the complete ensemble empirical mode decomposition with adaptive noise, and then the new components of the subsequences with different complexity are reconstructed using fuzzy entropy. Then, the Max-Relevance and Min-Redundancy method is used to select the influencing factors of the different reconstructed components. Then, a two-stage deep learning hybrid framework is constructed to model the prediction and nonlinear integration of the reconstructed components using a long short-term memory artificial neural network optimized by the gray wolf optimization algorithm. Finally, based on the proposed hybrid prediction framework, effective prediction and early warning of air pollutants are achieved. In an empirical study in three cities in China, the prediction accuracy, warning accuracy and prediction stability of the proposed hybrid framework outperformed the other comparative models. The analysis results indicate that the developed hybrid framework can be used as an effective tool for air pollutant prediction and early warning.

Keywords Feature extraction · Feature selection · Grey wolf optimization algorithm · Two-stage intelligent optimization · Air pollutant prediction and early warning

Abbreviations

AE	Approximate entropy	IMF	Intrinsic mode function
ARIMA	Autoregressive integrated moving average	GWO	Grey Wolf Optimizer
ANN	Artificial neural networks	GRNN	Generalized regression neural networks
CEEMDAN	Complete ensemble empirical mode decomposition with adaptive noise	GRU	Generalized regression neural networks
CTM	Chemical transport models	SE	Sample entropy
EMD	Empirical mode decomposition	LSTM	Long short-term memory neural network
EEMD	Ensemble empirical mode decomposition	MAE	Mean absolute error
FE	Fuzzy entropy	MAPE	Mean absolute percentage error
		mRMR	Max-relevance and min-redundancy
		PSO	Particle swarm optimization
		R^2	Coefficient of determination
		RMSE	Root mean square error
		RNN	Recurrent neural networks
		WNN	Wavelet neural networks

✉ Jujie Wang
jujiawang@126.com

- ¹ School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
- ² Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China

1 Introduction

In recent years, China's rapid urbanization and industrialization have brought about rapid economic development, while air pollution in most Chinese cities has become increasingly serious (Huang and Hu 2018). Air pollution directly affects the quality of the environment and people's physical and mental health seriously, among which $PM_{2.5}$ is one of the main components of air pollutants, mainly composed of highly reactive toxic and harmful substances. A large number of clinical cases and related studies have proved that there is a correlation between the occurrence of various respiratory and cardiovascular diseases and high concentrations of $PM_{2.5}$. $PM_{2.5}$ comes not only from natural sources such as wind, sand, dust and forest fires, but also mainly from human energy combustion and industrial production (Samal et al. 2021). During the continuous development of China's economy, air pollution is in some sense unavoidable. However, this does not mean that pollution emissions cannot be effectively prevented and controlled.

In order to strengthen air pollution control and improve air quality, China has amended the Air Pollution Control Law (Tao et al. 2013). In addition, accurate prediction of $PM_{2.5}$ concentrations is listed as one of the key objectives of air pollution prevention in China's Action Plan for the Prevention and Control of Air Pollution, which was proposed in 2018. Accurate $PM_{2.5}$ concentration prediction can help people understand future changes in air quality, so that they can prepare protective measures in advance to protect their health, such as wearing anti-haze masks (Zhu et al. 2018). It can also help researchers to develop response strategies in advance to prevent further deterioration of air quality (Liu et al. 2018). Therefore, the accurate prediction and early warning of $PM_{2.5}$ concentration has become a hot issue in the field of air pollution management research (Wu et al. 2016).

In a review of related studies, prediction models for air pollutant concentrations can be broadly classified into four types: chemical transport models (CMT), statistical models, artificial intelligence (AI) techniques, and hybrid models. CMT is a deterministic prediction based on the sources and transport of chemical substances (Xu et al. 2021; Shin et al. 2021). However, the prediction accuracy of CMT depends on the accurate description of the physical–chemical processes of pollutants and the quality of emission data (Konovalov et al. 2009). Therefore, CMT is more time consuming and complex than statistical models, while the accuracy is not stable (Han et al. 2008). Common statistical models are multiple linear regression and autoregressive integrated moving average (ARIMA). Donnelly et al. (2015) constructed a real-time air quality

prediction model using multiple linear regression. García et al. (2018) constructed ARIMA to predict daily PM_{10} concentrations in northern Spain with good prediction accuracy. Zhang et al. (2018) used ARIMA to analyze the trend of $PM_{2.5}$ concentrations and found a significant positive correlation with the changes in PM_{10} , SO_2 and NO_2 concentrations. Although a statistical model can obtain valid prediction results, it is based on a set of statistical assumptions. This makes statistical models not capable enough to capture nonlinear features from time series (Li et al. 2021a, b, c). To overcome the limitations of statistical models, AI models started to be applied to time series forecasting.

Data-driven AI techniques have excellent nonlinear fitting ability and robustness, so researchers have applied them widely in air pollutant prediction (Ren et al. 2021). Common AI models include artificial neural networks (ANN) (Ogliari et al. 2021; Zhang et al. 1998), generalized regression neural networks (GRNN) (Li et al. 2013), and recursive neural networks (Biancofiore et al. 2017). Feng et al. (2015) used air mass trajectory analysis to improve the accuracy of ANN prediction for daily average $PM_{2.5}$ concentrations. Combining ANN with effective training algorithms can extract potential nonlinear relationships between variables. It is demonstrated that a fast and economical air pollution warning system can be constructed using neural networks (Bo et al. 2021). Biancofiore et al. (2017) used the measured meteorological parameters as input variables to the recursive neural networks and predicted PM_{10} concentrations for the next one to three days. Yan et al. (2021) used GRNN to predict $PM_{2.5}$ concentration levels in three urban clusters in China. The results showed that GRNN could accurately predict $PM_{2.5}$ concentration levels in these clusters. Time series prediction is a prediction relative to data over a period of time. Using only the latest $PM_{2.5}$ concentration data for prediction, information from past data will be lost. Unlike traditional neural networks that ignore the long-term dependence of time series, recurrent neural networks (RNN) are able to maintain the memory of recent information. This gives it excellent performance in processing time series data (Wang et al. 2021). Long short-term memory neural network (LSTM), as a variant of RNN, has long-time memory capability and improves the problems of long-term dependence and gradient explosion that exist in RNN (Ahmed et al. 2021). Bai et al. (2019a, b) used LSTM to forecast $PM_{2.5}$ concentrations from two Beijing meteorological stations. The results demonstrate that LSTM can effectively capture complex features in nonlinear sequences. Although AI models have some advantages in air pollution prediction, single AI models still have problems such as unstable prediction results and easy over-fitting.

In order to further improve the performance and stability of prediction models, researchers have developed various hybrid models by effectively integrating different techniques and methods. And among various types of hybrid models, the hybrid model based on the idea of decomposition and integration can effectively deal with nonlinear and nonstationary time series and has excellent prediction performance, which becomes one of the hot spots in time series forecasting nowadays. Based on the idea of decomposition followed by integration, the nonlinear time series is first decomposed into several smoother subseries. Then, a suitable prediction model is constructed based on the characteristics of the decomposed time series, and finally the obtained results are integrated. The decomposition and integration method can effectively improve the prediction accuracy and prediction stability of nonlinear and non-smooth time series.

As an important module for decomposing integrated models, time series decomposition methods can extract more meaningful information and reduce the difficulty of prediction (Sun et al. 2022). Wavelet analysis is considered to be an effective algorithm for decomposing time series (Kisi and Alizamir, 2018). Huang and Wang (2018) used db6 wavelets and wavelet neural networks (WNN) to forecast four energy market price forecasts and experimentally prove that the hybrid model has higher accuracy. Nourani and Farboudfamm (2019) combined sym3 wavelet decomposition with LSSVM and ANN models for decomposing rainfall time series. However, the researcher needs to select a suitable wavelet basis function subjectively and without a specific theoretical basis when performing wavelet analysis. Empirical mode decomposition (EMD) is a data-driven adaptive method capable of decomposing nonlinear and nonstationary signals. Huang et al. (2012) constructed a decomposition integration framework based on EMD and gated recurrent unit neural network (GRU) for $PM_{2.5}$ prediction. Due to the lack of a complete theoretical basis itself, EMD algorithms suffer from problems such as modal mixing and endpoint effects (Li et al. 2021a, b, c). To solve the drawbacks of EMD, Wu and Huang (2009) proposed ensemble empirical mode decomposition (EEMD). Bai et al. (2019a, b) applied EEMD to $PM_{2.5}$ concentration prediction and improved the prediction accuracy. Although increasing the number of EEMD integrations can minimize the reconstruction error, the reconstructed components still contain residual noise of some magnitude. To extract more efficient features, Guo et al. (2020) applied a combination of complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and LSTM to chaotic sequence prediction. Lin et al. (2021) used the CEEMDAN-LSTM model to forecast the Chinese stock index, which proved to be the best among developed and emerging stock markets.

CEEMDAN has almost zero reconstruction error by adaptively increasing and weakening the white noise, which allows it to extract more effective information. However, multi-scale decomposition of nonlinear time series using signal decomposition algorithms often results in a new set of subseries. Previous studies usually predict each subsequence, but this also leads to an increase in the computational complexity of the model and may bring problems such as the accumulation of errors in prediction. The existence of similar trends and complexity among different subsequences is often overlooked, and the effective treatment of decomposed subsequences still needs further research.

Although the hybrid model based on the decomposition integration framework has been successfully applied to the prediction of various pollutant concentrations. However, $PM_{2.5}$ concentration variation is affected by complex factors such as meteorological and environmental factors (Hu et al. 2013), such as topography, vegetation, wind speed and temperature (Zhu et al. 2017). The training data consisting of only historical time series of $PM_{2.5}$ concentrations cannot provide enough valid information, which will undoubtedly hinder the prediction accuracy and generalization performance of the model (Wang et al. 2015). Related studies have shown that meteorological factors and pollutants have a strong influence on $PM_{2.5}$ fluctuations. Yoo et al. (2014) analyzed Korean $PM_{2.5}$ data between 2000 and 2012 and found a significant negative correlation between atmospheric precipitation and $PM_{2.5}$. Ma et al. (2021) found that factors such as precipitation, temperature, and wind direction can affect the concentration and dispersion range of $PM_{2.5}$. Bai et al. (2019a, b) concluded that meteorological data had a significant seasonal effect on $PM_{2.5}$ and used Kendall correlation analysis to extract the relationship between meteorological factors and $PM_{2.5}$. Gu et al. (2020) collected meteorological and pollutant data and divided $PM_{2.5}$ concentration data into environmental factors, temporal factors, and selected samples to construct a new superposition selective integration support predictor to achieve effective prediction of $PM_{2.5}$. However, there may be redundancy and similarity among different influencing factors, and if all factors are directly introduced into the prediction model, it may bring the problem of error accumulation and reduce the accuracy of the prediction model (Feng et al. 2021). Therefore, it is still challenging and needs further research to select the appropriate influencing factors for prediction.

From the above literature review, it can be found that although the hybrid model has excellent predictive performance and robustness, it still has some drawbacks. First, the similar complexity between decomposed subsequences is often ignored by researchers. Modeling each subseries separately not only increases the computational complexity

of the model, but also may lead to the accumulation of errors making the prediction accuracy lower. Second, many previous studies often use historical time series data of $PM_{2.5}$ for modeling, ignoring the influence of complex influencing factors on $PM_{2.5}$ fluctuation trends, which limits the prediction performance of the models. Third, deep learning models, as the main prediction models, are very sensitive to the selection of their hyperparameters. Different subsequences have different data characteristics, and choosing appropriate hyperparameters to model them can effectively improve the accuracy and stability of prediction. However, in previous studies, the selection of hyperparameters relied on empirical selection or repeated debugging, which made it difficult to determine the optimal hyperparameters. Fourth, after obtaining the prediction results for each subsequence, the present integration methods are mainly limited to linear integration, i.e., the predicted values are accumulated to obtain the final prediction results. Due to the errors in the prediction process, the linear integration method is not applicable to all cases and may lead to a decrease in prediction accuracy and stability. The nonlinear integration method can explore the intrinsic features among subseries to further improve the prediction accuracy.

Based on the above considerations, this study proposes a multi-factor multi-scale and intelligent optimization based two-stage deep learning hybrid framework for air pollutant forecasting and early warning, including CEEMDAN, fuzzy entropy (FE), the max-relevance and min-redundancy (mRMR), Gray Wolf Optimization algorithm (GWO) and LSTM. First, the $PM_{2.5}$ concentration sequence is decomposed into several subseries using CEEMDAN to reduce the complexity of the sequence and make it smoother. Then, each subsequence is reconstructed into several new components based on its FE value, which reduces the complexity of the model and improves the computational efficiency. Then, the mRMR algorithm is used to select several exogenous variables for each reconstructed component that have a large impact on it for prediction. Next, a two-stage intelligent optimization prediction model based on GWO algorithm and LSTM is developed to predict and nonlinearly integrate the reconstructed components to obtain the final $PM_{2.5}$ concentration prediction results. Finally, based on the accurate $PM_{2.5}$ concentration prediction results, an effective air pollution warning is achieved. In this paper, historical data of $PM_{2.5}$

concentrations in three Chinese cities are selected to validate the proposed hybrid framework. Compared with other benchmark models, the proposed model has good performance and prediction accuracy.

As shown above, the main contributions and innovations of this paper are as follows:

- (1) Considering the similar trend and complexity between different decomposition patterns, this paper develops a novel feature extraction method combining CEEMDAN and FE to effectively decompose $PM_{2.5}$ concentration sequences and extract different types of components from them, which improves the computational efficiency and accuracy of the prediction model.
- (2) Most previous studies have focused on prediction models based on historical $PM_{2.5}$ concentration time series. This paper develops an mRMR-based feature selection method that uses $PM_{2.5}$ data and multi-influence factor data as input features in the modeling process to construct a multi-influence factor-based hybrid prediction framework.
- (3) In order to further improve the prediction performance and stability of the neural network, this paper uses GWO to intelligently seek the optimal hyperparameters of the LSTM. Based on GWO-LSTM, a two-stage intelligent optimization model is developed to model and predict each reconstruction component separately and integrate the predictions nonlinearly.
- (4) In this paper, a two-stage deep learning hybrid prediction framework based on multi-factor multi-scale and intelligent optimization are proposed for the first time. The hybrid framework outperforms all comparative models and has good prediction accuracy and stability. Based on this hybrid prediction framework, an air pollution prediction and early warning system is established to achieve effective forecasting and warning of future air pollutant concentrations and air pollution conditions.

The rest of the paper is structured as follows: Sect. 2 outlines the methods used in this paper. Section 3 details the structure of the hybrid prediction framework and the evaluation metrics of prediction performance. Section 4 describes the data preprocessing, forecasting process and comparative experiments. Finally, Sect. 5 shows conclusions and outlook.

2 Methodology

2.1 Complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN)

Huang et al. (1998) proposed an adaptive signal processing method for handling nonlinear nonstationary data, namely EMD. EMD does not require assumptions on the data and can decompose complex nonstationary signals and preserve the time scale of the data. However, in practical applications, EMD often encounters the problem of modal mixing (Sun et al. 2022). Therefore, Wu et al. (2009) proposed EEMD based on EMD. EEMD decomposes the data by repeatedly adding varied white noise to the original signal, based on the condition that white noise’s average value is zero. This can effectively improve the modal mixing problem, but it also has the problems of large upper reconstruction error and long computation time. For this reason, the EEMD-based CEEMDAN is proposed to solve the above problems. CEEMDAN not only effectively overcomes the problem of modal mixing by adding adaptive white noise, but also removes the reconstruction error and reduces the computing cost (Li et al. 2021a, b, c). Therefore, CEEMDAN can handle nonsmooth and nonlinear data more effectively.

The CEEMDAN algorithm is implemented in the following steps.

Step 1: The Gaussian white noise with mean zero is first added to the original signal $s(t)$ to obtain the preprocessed signal $s_i(t)$ for k experiments.

$$s_i(t) = \varepsilon\omega_i(t) + s(t), \quad i = 1, 2, \dots, k \tag{1}$$

where $\omega_i(t)$ is the Gaussian white noise of the i th processing, ε is the noise ratio bar. Then the first intrinsic mode function (IMF) component $IMF_1^i(t)$ is obtained by decomposing $s_i(t)$ using EMD, and its mean value is found as the first IMF component $IMF_1(t)$ obtained by CEEMDAN decomposition.

$$IMF_1(t) = \frac{1}{k} \sum_{i=1}^k IMF_1^i(t) \tag{2}$$

Step 2: To calculate the first residual $r_1(t)$, subtract the first IMF from the initial sequence.

$$r_1(t) = s(t) - IMF_1(t) \tag{3}$$

Step 3: Gaussian white noise is added into the residual signal of the j th stage obtained from the decomposition, and the EMD decomposition is continued.

$$IMF_j(t) = \frac{1}{k} \sum_{i=1}^k E_1(r_{j-1}(t) - \varepsilon_{j-1}E_{j-1}(\delta_i(t))) \tag{4}$$

$$r_j(t) = r_{j-1}(t) - IMF_j(t) \tag{5}$$

where $IMF_j(t)$ is the j th IMF obtained from the CEEMDAN decomposition, E_{j-1} is the j th IMF component obtained by performing the EMD decomposition, and ε_{j-1} is the noise factor added to the residual component of the stage $j - 1$. Finally, $r_j(t)$ is the residual component of the i -th stage.

Step 4: Repeat steps 1 to 3 until the number of extreme value points of the residual components is less than or equal to 2, and the decomposition process of CEEMDAN is finished when the decomposition cannot continue. At this time, the $PM_{2.5}$ concentration sequence is decomposed into several IMF components and one residual component.

2.2 Fuzzy entropy (FE)

In order to strike a balance between the computational efficiency of the model and the accuracy of the prediction, a method to measure the complexity of time series is adopted, which is called fuzzy entropy (FE). The FE algorithm is an improvement of the sample entropy (SE) and approximate entropy (AE) methods, which retains the advantages of sample entropy and approximate entropy and addresses the shortcomings of imprecise analysis in the presence of small fluctuations and baseline drift. In general, the larger the value of FE, the lower the serial autocorrelation. Therefore, the features can be recombined according to the values calculated by the FE algorithm to balance the computational efficiency and prediction accuracy. The calculation process of FE are as follows:

Step 1: Defining the phase space dimension as m , the phase space reconstruction is performed on the N -dimensional time series $\{x(1), x(2), \dots, x(N)\}$ to obtain X_i^m . Where $w_0(i)$ is the mean value.

$$X_i^m = \{x(i), x(i + 1), \dots, x(i + m - 1)\} - w_0(i) \tag{6}$$

$$w_0(i) = \frac{1}{m} \sum_{j=0}^{m-1} x(i + j) \tag{7}$$

Step 2: Define the absolute distance $d_{a,b}^m$ between the vectors X_i^m and X_j^m as the maximum value of the difference of their corresponding elements. Where $j = 1, 2, \dots, N - m + 1$, and $j \neq i$.

$$d_{ij}^m = d[x_a^m, x_b^m] = \max_{k=0,1,2,\dots,m} (|x(a+k-1) - w_0(i)| - |x(a+k-1) - w_0(j)|) \tag{8}$$

Step 3: Next, a fuzzy function $F(d_{ij}^m, n, r)$ is introduced to define the correlation D_{ij}^m between X_i^m and X_j^m . In Eq. (9), r denotes the boundary width and n denotes the boundary gradient.

$$D_{ij}^m = F(d_{ij}^m, n, r) = \exp\left(-\left(\frac{d_{ij}^m}{r}\right)^n\right) \tag{9}$$

Step 4: Define the fuzzy degree similarity function as:

$$\delta^m(n, r) = \frac{1}{N - m} \sum_{i=1}^{N-m} \left(\frac{1}{N - m - 1} \sum_{j=1, j \neq i}^{N-m} D_{ij}^m\right) \tag{10}$$

Step 5: Change the phase space dimension to $m + 1$ and repeat the above calculation steps to obtain the function.

$$\delta^{m+1}(n, r) = \frac{1}{N - m} \sum_{i=1}^{N-m} \left(\frac{1}{N - m - 1} \sum_{j=1, j \neq i}^{N-m} D_{ij}^{m+1}\right) \tag{11}$$

Step 6: Ultimately, the fuzzy entropy of this time series is defined as:

$$FuzzyEn(N, m, n, r) = \ln \delta^m(n, r) - \ln \delta^{m+1}(n, r) \tag{12}$$

The results of FE algorithm are mainly determined by the parameters m, n and r . In general, m is often taken as 1 or 2. r is usually set to $0.1\sigma_{SD}$ to $0.25\sigma_{SD}$, and the σ_{SD} is the standard deviation of the original sequence. n is generally taken as a smaller integer value, such as 1 or 2.

2.3 Max-relevance and min-redundancy (mRMR)

The mRMR method is a typical spatial search-based filtering method proposed by Peng et al. in 2005, which uses mutual information to measure the relevance and redundancy of features (González-Enrique et al. 2021). Let $W_n = \{z_1, z_2, \dots, z_n\}$ be the set of influencing factor features, we need to select m meteorological features with high correlation with $PM_{2.5}$ from n influencing factors. Firstly, the

mutual information $MI(s(t), z_i)$ between $PM_{2.5}$ concentration $s(t)$ and all influencing factors is calculated as:

$$MI(s(t), z_i) = \int \int p(s(t), z_i) \log \frac{p(s(t), z_i)}{p(s(t))p(z_i)} ds(t) dz_i \tag{13}$$

The mutual information between the influencing factors is:

$$I(z_i, z_j) = \int \int p(z_i, z_j) \log \frac{p(z_i, z_j)}{p(z_i)p(z_j)} dz_i dz_j \tag{14}$$

where p is the probability density function, $z_i, z_j \in W_n$, $i \neq j$, and $s(t)$ is the $PM_{2.5}$ concentration sequence. Then find the feature subset S_m containing m features, where $m \leq n, S_m \subseteq W_n$. The formulae for the maximum relevance calculation principle and the minimum redundancy calculation principle are as follows:

$$D(S_m, s(t)) = \frac{1}{|S_m|} \sum_{z_i \in S_m} MI(s(t), z_i) \tag{15}$$

$$N(S_m) = \frac{1}{|S_m|^2} \sum_{z_i, z_j \in S_m} MI(s(t), z_i) \tag{16}$$

where $|S_m|$ is the number of features in the set S_m . The formula for integrating the maximum relevance and minimum redundancy is as follows.

$$\max \phi(D, N), \phi = D - N \tag{17}$$

Suppose that the factor z_k with the largest mutual information with $s(t)$ among the influencing factors is extracted as the first characteristic factor within S_m , and the remaining influencing factors are $W_n = W_n - z_k$. The mutual information of the influencing factors within W_n with $s(t)$ is calculated separately, and ϕ is maximized by selecting the characteristics. ϕ is calculated as:

$$\max \Delta\phi = MI(s(t), z_i) - \frac{1}{|W_n| - 1} \sum_{z_i, z_j \in W_n} MI(z_i, z_j) \tag{18}$$

In the above equation, $\Delta\phi$ is the operator increment, which is the difference between the mutual information of influences z_i within W_n and $s(t)$ and the mutual information of z_i and other influences within W_n . The magnitude of $\Delta\phi$ can be used as a basis for evaluating the importance of features. In addition, $|W_n|$ is the number of feature values in the set W_n .

2.4 Long short-term memory (LSTM)

RNN as a new type of neural network with memory function is suitable for time series problems. However, it also suffers from problems such as gradient disappearance and gradient explosion. LSTM as a modified RNN consists of a memory cell, input gate, forgetting gate and output gate. The input and forgetting gates are used to determine whether to add new input information and whether to forget past states. The output gate, on the other hand, determines whether the long-term state is propagated to the final state. LSTM effectively avoids the gradient disappearance problem of RNN and has long-time memory capability at the same time (Barzegar et al. 2020). In this study, LSTM was used for prediction and nonlinear integration of $PM_{2.5}$ concentration.

The actual $PM_{2.5}$ concentration value at time t is assumed to be x_t , and \hat{x}_t is the predicted value corresponding to the $PM_{2.5}$ concentration. Moreover, f_t , i_t and o_t represent the forgetting gate, the input gate and the output gate, respectively. The main formulae for each component of the LSTM are shown below:

$$f_t = \sigma(U_f x_t + V_f \hat{x}_t + b_f) \quad (19)$$

$$i_t = \sigma(U_i x_t + V_i \hat{x}_t + b_i) \quad (20)$$

$$o_t = \sigma(U_o x_t + V_o \hat{x}_t + b_o) \quad (21)$$

$$\tilde{c}_t = \tanh(U_{\tilde{c}} x_t + V_{\tilde{c}} \hat{x}_t + b_{\tilde{c}}) \quad (22)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (23)$$

$$h_t = \tanh(c_t) * o_t \quad (24)$$

In the above equation, U and V are the weight matrices, b is the bias term, and $\tanh(\cdot)$ and $\sigma(\cdot)$ are the activation functions, ' $*$ ' denotes the scalar product. The LSTM consists of these memory blocks and is learned by a temporal algorithm using back propagation. More, LSTM is prone to overfitting or gradient explosion when dealing with long sequences. Adding Dropout layers and adjusting the appropriate Dropout layer rate can improve the generalization ability of the model and avoid overfitting.

2.5 Grey wolf optimizer (GWO)

Mirjalili et al. (2014) proposed the GWO algorithm, which is a new swarm intelligence method. GWO is considered to possess stronger performance than many existing superior algorithms such as particle swarm optimization algorithms (PSO) (Sulaiman et al. 2015). GWO simulates the hunting ability and social hierarchy of gray wolves. In general, GWO divides the social level of wolves into four levels. The first level of the pyramid is the leader of the wolf pack, which is called α . The second level is β , which has

dominance only after α . The third level is δ , which obeys the decisions of α and β . The bottom level is ω , which is responsible for the balance within the pack. The GWO optimization process involves hierarchical hierarchy, tracking, encircling and attacking prey and finding prey. GWO keeps the best three wolves (α , β , δ) in each iteration and updates them according to the three best solutions.

2.6 GWO-LSTM model

The training process of LSTM is mainly based on the update of weights and bias of hyperparameters, and the choice of hyperparameters can significantly affect the prediction performance of LSTM. Related studies show that the number of neural units directly affects the fitting ability of the model, and increasing the number of LSTM neural units can increase the fitting ability of the prediction model, but too many neural units may also lead to overfitting. However, there is no clear method to select the number of neural units. In addition, batch size is closely related to the weight update of the model. The size of batch size affects the convergence speed and prediction performance of the prediction model. Traditional prediction model research often relies on empirical selection when adjusting hyperparameters, by repeatedly experimenting and adjusting hyperparameters until the training set prediction error is minimized. This approach is time-consuming and difficult to obtain the best hyperparameters for prediction models.

To balance the complexity of prediction and prediction accuracy, the hyperparameters of the LSTM network are optimized using the GWO algorithm. When training the data, the number of iterations, the number of gray wolves, and the dimensionality of GWO are determined by first determining the historical data step lookback of the input layer in the LSTM. The fitness function is set as:

$$fitness = \frac{1}{N} \sum_{i=1}^N |y_i - y'_i| \quad (25)$$

where y_i is the training set data, y'_i is the predicted data, N is the length of the data. The hidden layer neurons, batch size and Dropout rate are selected as the target hyperparameters for optimization. The target hyperparameters are corresponding to the wolf positions of GWO in different dimensions, thus transforming the learning process of the neural network into the process of searching for the best position of wolves in the multidimensional space. The hyperparameters are substituted into the LSTM to calculate the corresponding prediction value y and the fitness value of each individual is calculated according to Eq. (25). Continuously iterating, the hyperparameter optimal solution of the LSTM network is finally returned. The pseudo code of the GWO-LSTM algorithm is shown below.

Algorithm 1. *GWO-LSTM***Parameters:**

G_{max} - maximum number of iterations t - current iteration number
 n - population number F_i - the fitness function of x_i
 Lb - lower bounds of the search domain Ub - upper bounds of the search domain
1: /* Initialize GWO population $X = \{x_i, i = 1, 2, \dots, n\}$ in random position. */
2: /* Initialize the parameters X_α , X_β and X_δ . Determine the topology of the LSTM network. */
3: **WHILE** $t < G_{max}$ **DO**
4. /* For search space. */
5. /* Update the current agent's location. */
6: **FOR EACH** $i = 1: n$ **DO**
7: /* Update the parameters X_α , X_β and X_δ . */
8. /* Pass wolfpacks' values to the given LSTM network as units, batch size and learning rate. */
9: /* Calculate the corresponding fitness function F_i^* . */
10. /* Find the best nest location x_i^0 and the minimum fitness f_{min} . */
11. /* Adjust the parameters of the LSTM and minimize the error. */
12: **END FOR**
13: **IF** $f_{best} < f^*$ **THEN**
14: /* Find the best nest position x_i^i . */
15: **END IF**
16: $iter = iter + 1$
17: **END WHILE**
18: **RETURN** X_α

3 Structure of the proposed hybrid framework

3.1 Proposed hybrid framework

This study proposes a deep learning hybrid framework for air pollutant prediction and early warning based on multi-factor multi-scale and two-stage intelligent optimization, which combines CEEMDAN, FE, mRMR, GWO and LSTM. As shown in Fig. 1, the framework is mainly divided into four stages.

Stage 1: Feature extraction

CEEMDAN can decompose the $PM_{2.5}$ concentration sequence adaptively into several patterns of different amplitudes and frequencies. The patterns obtained from the decomposition are arranged by frequency from high to low

frequencies. Compared with the original $PM_{2.5}$ sequence, these patterns have simpler structure, more stable fluctuations and more regularity, which can be predicted more easily. However, there are similar trends and complexity between these patterns, and fuzzy entropy can effectively calculate the complexity of different sequences, the higher the entropy value the higher the complexity of the sequence, and the lower the entropy value the lower the complexity of the entropy value. According to the similar trend and complexity of different patterns, the decomposed patterns can be reconstructed into several new components. Each reconstructed component has unique characteristics and contains different intrinsic features of $PM_{2.5}$ concentration.

Stage 2: Feature Selection

The fluctuation of $PM_{2.5}$ concentration is affected by complex factors such as environmental factors, human factors and meteorological factors, which leads to complex characteristics such as nonlinearity and nonstationarity of $PM_{2.5}$. In this study, meteorological factors and pollutant factors are taken into account in the prediction of $PM_{2.5}$ concentrations to improve the prediction accuracy and generalization ability of the model. There is redundancy between different influencing factors. Directly using all influencing factors for prediction research may lead to problems such as the cumulative error of prediction model. At the same time, different reconstructed components contain different intrinsic characteristics, and the same influencing factors have different degrees of influence on different reconstructed components. Therefore, in this paper, the mRMR algorithm is used to select the features of different components obtained after the decomposition and reconstruction of $PM_{2.5}$. Several exogenous variables that have a strong influence on different reconstructed components are selected as input variables to improve the prediction accuracy and generalization performance of the model.

Stage 3: Two-stage intelligent optimization model

$PM_{2.5}$ concentrations are influenced by its own historical concentration data and related factors and change gradually over time. LSTM can effectively capture the nonlinear relationship in the sequence and has the ability of long-term memory, which can combine the historical and current information in the long-term memory to make effective prediction for the future. Therefore, LSTM is used to predict $PM_{2.5}$ future concentrations.

To balance the computational efficiency and prediction accuracy of the prediction model, this paper uses the GWO algorithm to optimize the hyperparameters of the LSTM. Based on the GWO-LSTM, a two-stage intelligent optimization model is developed to model the prediction for each subset of sequences, and all predictions are nonlinearly integrated to obtain the final $PM_{2.5}$ concentration prediction results.

Stage 4: Air Pollution Prediction and Warning

China's Ambient Air Quality Standards, released in 2012 and implemented in 2016, regulate air environmental quality standards to further prevent and control air pollution and protect people's physical and mental health. The standard divides the ambient air functional areas into two categories, such as areas requiring special protection, such as nature reserves, and areas such as residential and industrial zones, and sets standards for pollutant concentration limits, which provide scientific support for the monitoring and management of environmental quality nationwide. To make it easier for people to pay attention to air pollution, Chinese private individuals have organized themselves to set up various environmental monitoring

websites for the release of air pollution information in major cities. These air pollution monitoring websites set more refined criteria for assessing air pollution levels, making it easier for people to understand air pollution levels more intuitively.

Based on the proposed hybrid prediction framework, this paper makes effective predictions of $PM_{2.5}$ future concentrations. As shown in Table 1, this paper also makes reference to the air pollution level criteria from the $PM_{2.5}$ real-time monitoring network (<http://www.pm25china.net/>) to provide early warnings of future air pollution levels, helping people to prepare coping strategies in advance and the government to take pollution prevention and control measures in advance in a targeted manner.

3.2 Evaluation criteria

To assess the predictive performance of various models, we must choose appropriate evaluation metrics. In this paper, four popular evaluation metrics are chosen to measure the performance of models, including mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and mean absolute percentage error (MAPE). These metrics have been widely used in pollution prediction studies (Sun and Li 2020; Wu et al. 2020), and the details of each metric are described as follows:

$$MAE = \frac{1}{k} \sum_{i=1}^k |y_i - \hat{y}_i| \quad (26)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^k (\hat{y}_i - y_i)^2}{k}} \quad (27)$$

$$R^2 = 1 - \frac{\sum_{i=1}^k (y_i - \hat{y}_i)^2}{\sum_{i=1}^k (y_i - \bar{y}_i)^2} \quad (28)$$

$$MAPE = \frac{1}{k} \sum_{i=1}^k \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100\% \quad (29)$$

where k represents the number of test sets, y represents the true value, \bar{y} represents the mean value, and \hat{y} represents the predicted outcome.

4 Case analysis

4.1 Data collection

4.1.1 $PM_{2.5}$ concentration data

With the rapid development of the economy, air pollution has become an urgent problem in China. Researchers have focused their $PM_{2.5}$ prediction studies on various economically developed cities in China. Examples include cities such as Beijing (Luo et al. 2018), Shanghai (Xu et al.

Table 1 PM_{2.5} air pollution standards ($\mu\text{g}/\text{m}^3$)

Ranges	0–50	51–100	101–150	151–200	201–300	> 300
Level	Excellent	Good	Light pollution	Medium pollution	Highly pollution	Severe pollution

2017), and Wuhan (Wang et al. 2017). However, other industrial cities in China, where air pollution is more severe, are often neglected. In this paper, we fill this gap in the literature by selecting Xingtai and Anyang, two of the most polluted Chinese cities in terms of air pollution, as the study samples by referring to the 2019 China Ecological Environment Status Bulletin published by the Chinese Ministry of Ecology and Environment in 2020. In addition, Beijing, the capital city of China, is added as a research sample to verify the validity of the hybrid framework proposed in this paper.

In 2020 and 2021, some meteorological observations were stopped by COVID-19, and the relevant data were missing. Therefore, the data set for this study is the PM_{2.5} daily average concentration data from three cities from January 1, 2018 to December 31, 2019. The data were obtained from the Ministry of Ecology and Environment of China (<http://www.mee.gov.cn>). As shown in Figs. 1 and 2, the sample data are divided into training and test sets by 8:2.

4.1.2 Influencing factors of PM_{2.5}

The causes of PM_{2.5} pollution are complex. For example, pollutants such as nitrogen oxides and sulfur dioxide in the atmosphere are easy to produce secondary fine-grained pollutants through chemical reactions. In addition, human industrial production and living activities also bring a large number of fine particle pollutants. In addition, relevant studies show that meteorological factors have stronger influence on PM_{2.5} than other factors (Chen et al. 2017). For example, wind speed and direction can affect the diffusion range and speed of pollutants. PM_{2.5} is easily adsorbed to water vapor, so when the humidity is high, the concentration of PM_{2.5} is high. In addition, when the temperature increases, the concentration of PM_{2.5} decreases continuously, and when the temperature decreases, the concentration of PM_{2.5} increases significantly. As shown in Fig. 2, draw the daily concentration change curve of PM_{2.5}. PM_{2.5} concentration changes obviously in different seasons, showing a “double peak” distribution mode. The concentration of fine particles in winter and spring is significantly higher than that in summer and autumn, which is related to the temperature difference in different seasons. Based on referring to relevant literature and considering the availability of data, this paper introduces 11 influencing factors of PM_{2.5}, including average wind speed, maximum sustainable wind speed, average air temperature, average

dew point, maximum temperature, minimum temperature, PM₁₀, SO₂, CO, NO₂ and O₃.

In artificial intelligence algorithms, dimensionless quantization of data can accelerate convergence and avoid the influence of singular sample data on calculation results. The typical normalization method is adopted in this study, and the formula is as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (39)$$

where the original value is x and the normalized result is x' .

4.2 Decomposition of original PM_{2.5} series by CEEMDAN

In the proposed framework, the original PM_{2.5} concentration sequence is decomposed by CEEMDAN. Before that, two parameters, k and ε , need to be set for CEEMDAN. Referring to relevant literature and several attempts, the values of k and ε are set to 100 and 0.005, respectively. The Xingtai PM_{2.5} concentration sequence is split into seven subsequences, as illustrated in Fig. 3A. Each decomposition pattern is named *IMFi* ($i = 1, 2, \dots, 6$) and Residual, respectively. Meanwhile, the original PM_{2.5} sequence in Anyang and Beijing is decomposed into 8 and 7 subsequences, respectively. In addition, the Pearson correlation coefficients of the original data and each IMF are calculated in this paper to facilitate the exploration of the relationship between the decomposed subsequences and the original sequences, and are presented as bar charts in Figs. 3, 4 and 5B.

4.3 Subsequence reconstruction by FE

Fuzzy entropy can measure the complexity of different sequences, and Figs. 3C, 4C and 5C show the results of fuzzy entropy calculation for each subsequence. In this paper, according to the similar trends and fuzzy entropy values between different subsequences after decomposition, they can be reconstructed into three new components. (1) IMF1 is the high frequency component S-IMF1, which can respond to the random fluctuation of PM_{2.5} concentration caused by various complex factors. Although the high-frequency component may lead to short-term drastic changes in PM_{2.5} concentration, it does not cause long-term effects on PM_{2.5} concentration fluctuations (Tai et al. 2010). (2) IMF2 and IMF3 are reconstructed as the

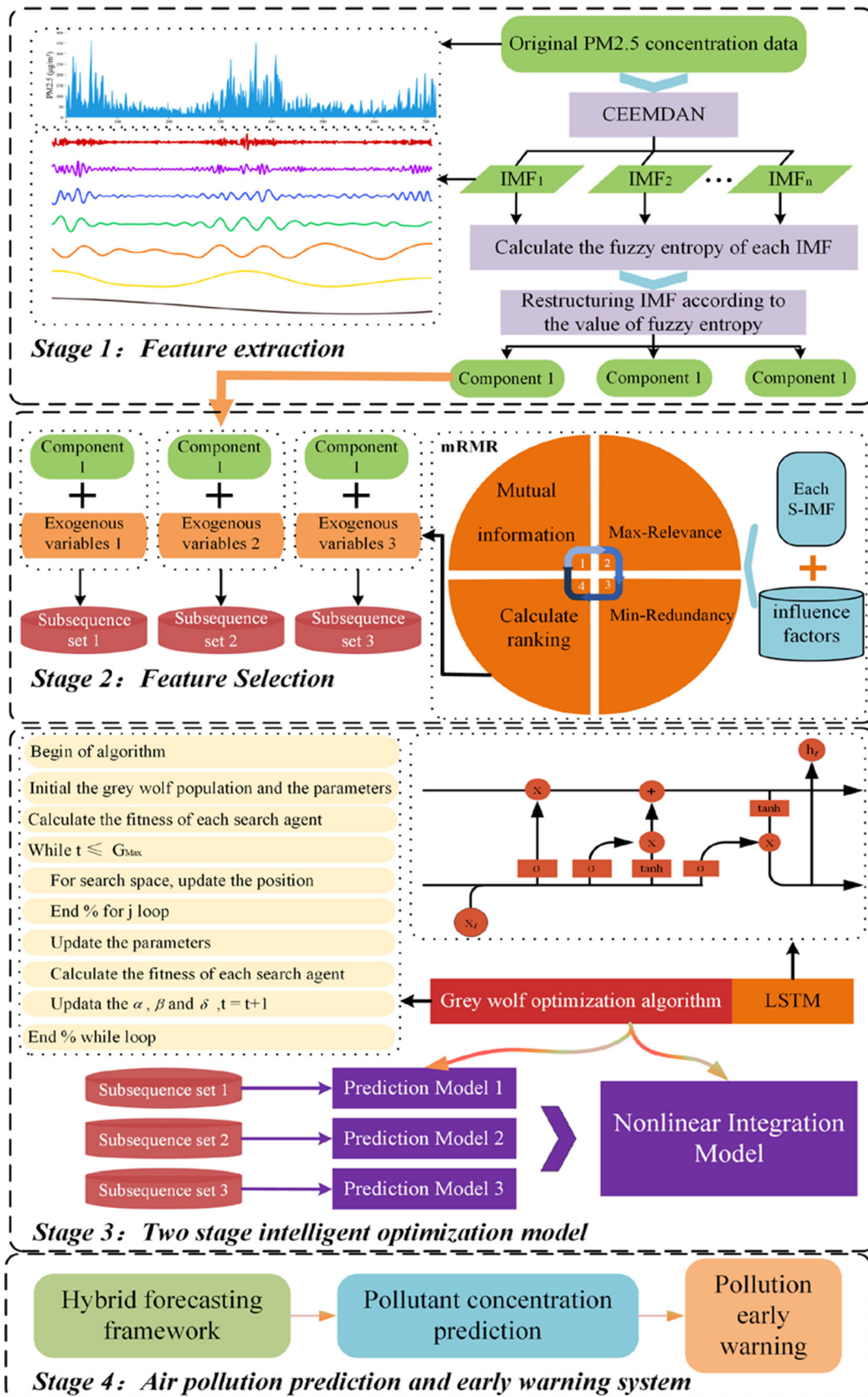


Fig. 1 Flow chart of the framework

intermediate frequency component S-IMF2, which responds to the periodic variation of $PM_{2.5}$ concentration caused by atmospheric quasi-biennial oscillations, weather-scale system cycles, or human periodic activities (Kim et al. 2010; You et al. 2009; Zhang et al. 2015). (3) IMF4, IMF5, IMF6 and residuals are reconstructed as low frequency component S-IMF3, and such components are more stable and can effectively characterize the trend of $PM_{2.5}$ concentrations during seasonal change (Wang et al. 2006). Taking Xingtai as an example, the three components obtained from the reconstruction are shown in Figs. 3D, 4D and 5D. Each component has unique characteristics, and the accuracy and stability of its prediction will be improved by selecting appropriate influencing factors and constructing prediction models according to the data characteristics of different components.

4.4 Influencing factors selection by mRMR

After obtaining the three reconstructed components, this paper uses the mRMR method for each component to explore the main influencing factors of different components. Using Xingtai as an example, Table 2 demonstrates the mRMR results for the three components.

According to the results in Table 2, for the irregular fluctuation components of $PM_{2.5}$, O_3 , PM_{10} , mean dew point, maximum temperature and CO lead the ranking order, indicating a greater influence on them. O_3 , CO, mean dew point, mean temperature and mean wind speed have a greater influence on the short-term fluctuation of $PM_{2.5}$. The short-term fluctuation component of $PM_{2.5}$ is strongly influenced by O_3 , CO, average dew point, average temperature and average wind speed. While the low frequency component is strongly affected by O_3 , CO, average temperature, average dew point and maximum sustainable wind speed. Due to the correlation between influencing factors, all as input features may reduce the prediction performance and accuracy. Therefore, this paper selects the top 5 exogenous variables for each component as the input variables for prediction.

4.5 Two-stage intelligent optimization model

LSTM is well-suited to processing and forecasting time series data, and the selection of hyperparameters is critical in LSTM training. Increasing the number of layers and neurons of the neural network can effectively improve the fitting ability of the model, but also increases the risk of

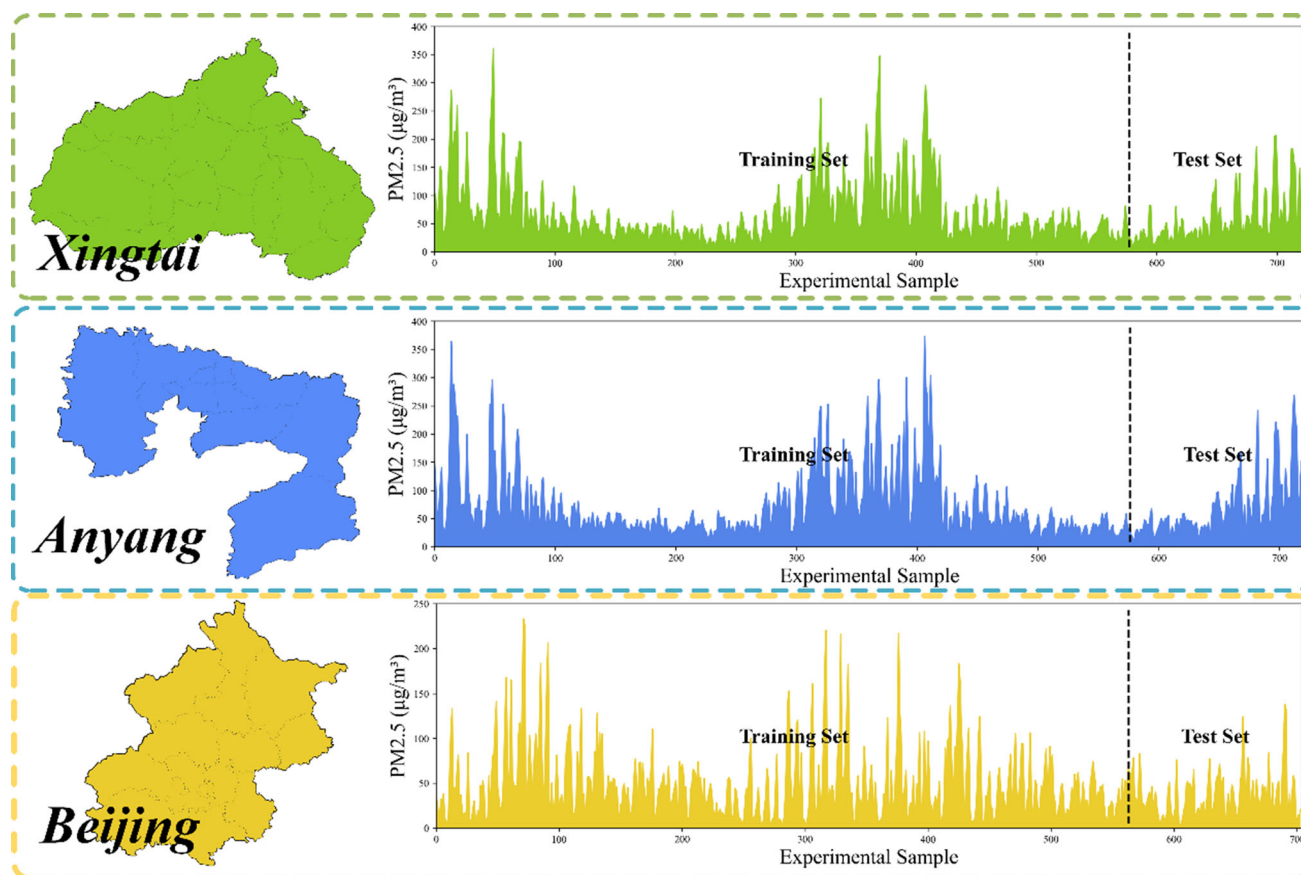


Fig. 2 Sample data of Xingtai, Anyang and Beijing

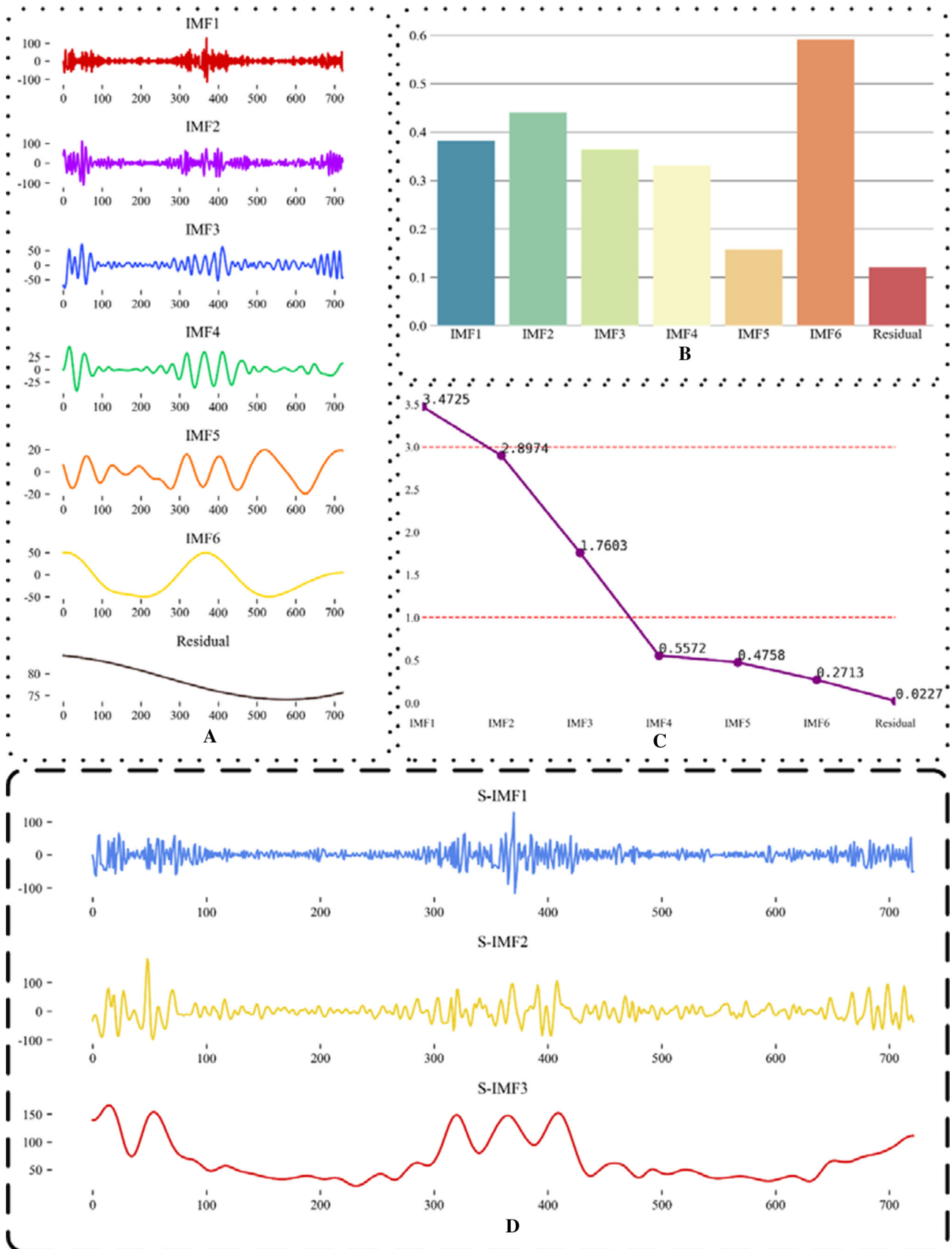


Fig. 3 Decomposition and reconstruction results of Xingtai dataset

overfitting. And by introducing the Dropout mechanism, given the Dropout ratio, so that the model randomly discards the corresponding number of neurons in the training process, it can effectively prevent overfitting. After a large number of experiments and parameter adjustment, it is found that LSTM with two hidden layers has excellent prediction performance and robustness on different data sets, and the look back is set to 4, and the upper limit of epoch for each experiment is set to 1000. Based on the consideration of balancing prediction efficiency and prediction accuracy, the GWO algorithm is used to optimize the three hyperparameters of the number of hidden layer neurons, batch size and Dropout ratio. The gray wolf population size is set to 25, and the number of iterations is 100. Finally, a suitable prediction model is established for each reconstructed component. Table 3 shows the prediction accuracy of the hybrid framework on three data sets.

The LSTM can effectively capture the information in nonlinear data, and the nonlinear integration can obtain higher prediction accuracy and prediction stability. Therefore, after obtaining the prediction results for each

reconstructed component, the GWO-LSTM is used to nonlinearly integrate all the predictions to obtain the final prediction results. The performance of the hybrid framework proposed in this paper on three datasets is illustrated in Fig. 6.

4.6 Air pollutant forecasting and warning

The hybrid framework proposed in this paper obtains accurate $PM_{2.5}$ concentration prediction results on all three data sets, and effective forecasting of air pollutant concentrations can be achieved based on the prediction results. More, based on the air quality criteria in Table 1, the future air quality levels are warned based on the prediction results, and the accuracy of the warning results is shown in Table 5. As two of the most polluted cities in China, Xingtai and Anyang have large fluctuations in pollutant concentrations. In the test set of 141 days, Xingtai has 14 days of light pollution, 6 days of medium pollution and 2 days of highly pollution, and the warning accuracy of the hybrid framework proposed in this paper reaches 87%. In

Table 2 Order of exogenous variables in Xingtai dataset

Ranking order	S-IMF1	S-IMF2	S-IMF3
1	O ₃	O ₃	O ₃
2	PM ₁₀	CO	CO
3	Average dew point	Average dew point	Average air temperature
4	Maximum temperature	Average air temperature	Average dew point
5	CO	Average wind speed	Maximum sustainable wind speed
6	NO ₂	SO ₂	SO ₂
7	SO ₂	NO ₂	Minimum temperature
8	Minimum temperature	Minimum temperature	NO ₂
9	Average wind speed	Maximum continuous wind speed	Maximum temperature
10	Average air temperature	Maximum temperature	Maximum continuous wind speed
11	Maximum continuous wind speed	PM ₁₀	PM ₁₀

Table 3 Model prediction accuracy of each reconstructed sub-sequences

		Subsequence set 1	Subsequence set 2	Subsequence set 3
Xingtai	MAE	0.0237	0.0167	0.0098
	RMSE	0.0281	0.0211	0.0111
	MAPE	5.2616	5.4754	4.5586
Anyang	MAE	0.0216	0.0161	0.0079
	RMSE	0.0244	0.0234	0.0091
	MAPE	4.2446	4.1260	7.6934
Beijing	MAE	0.0235	0.0145	0.0048
	RMSE	0.0279	0.0194	0.0056
	MAPE	5.0212	3.3774	12.6619

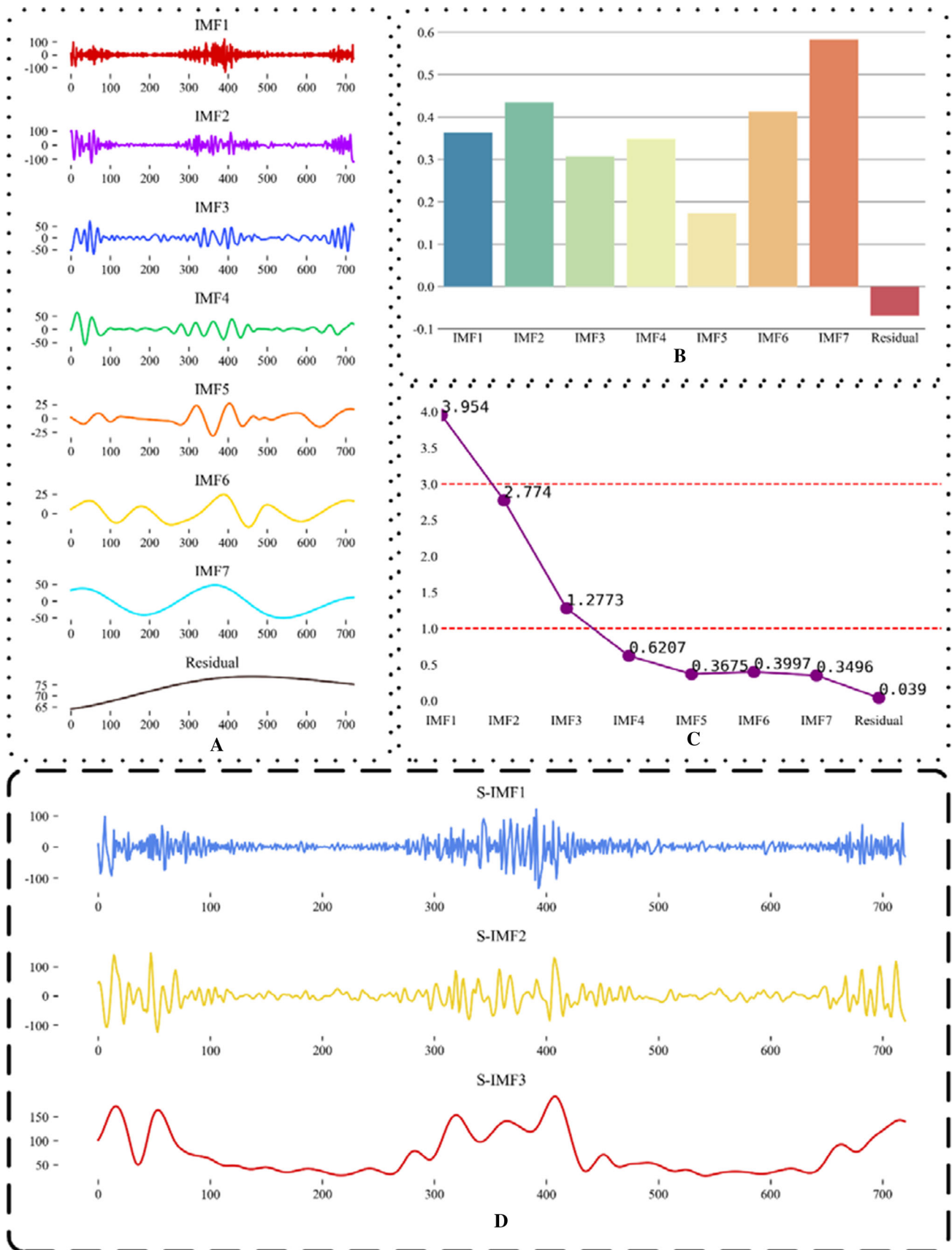


Fig. 4 Decomposition and reconstruction results of Anyang dataset

Anyang, there were 9 days of light pollution, 7 days of medium pollution and 7 days of highly pollution, and the accuracy of early warning reached 90%. In Beijing, which has a developed economy and vigorously combats air

pollution, there are only 3 days of light pollution, and the rest of the time are good or excellent, and the warning accuracy of the hybrid framework out of this paper reaches 93%. Therefore, the hybrid framework proposed in this paper can be used as an effective tool for air pollution forecasting and early warning.

Table 4 Comparison results of optimization algorithms

Model	Iteration time (minutes)	MAE	RMSE	R ²	MAPE
GA-LSTM	86	19.8431	23.3231	0.5511	41.3411
PSO-LSTM	85	17.9315	21.1001	0.6022	37.9812
SA-LSTM	92	19.1767	22.8110	0.5721	38.1251
GWO-LSTM	79	17.0412	20.8713	0.6231	35.2032

4.7 Comparative experiments

In order to verify the effectiveness and superiority of the hybrid model proposed in this paper, we designed two sets of comparison experiments. The first set of experiments uses three commonly used optimization algorithms, Genetic Algorithm (GA), Particle Swarm Algorithm (PSO) and Simulated Annealing Algorithm (SA), to perform hyperparametric optimization of the LSTM to verify the

Table 5 Results of eight comparative models based on relevant literature

Model	Cases	MAE	RMSE	R ²	MAPE	Accuracy (%)
RF-SVR Ishak (2016)	Xingtai	20.8076	28.0947	0.5717	60.0995	49
	Anyang	25.0864	36.1994	0.5351	55.1897	53
	Beijing	17.1767	22.1765	0.1979	103.2237	42
ANN Goudarzi et al. (2021)	Xingtai	21.3315	30.7800	0.4858	50.1228	33
	Anyang	24.0341	36.3902	0.5302	46.9429	56
	Beijing	15.8040	21.2815	0.2612	87.1735	40
LSTM Yang et al. (2021)	Xingtai	20.9384	29.9231	0.5142	45.6219	67
	Anyang	23.4964	35.7736	0.5460	48.3475	55
	Beijing	15.8368	21.2492	0.2635	83.9802	41
EEMD-LSTM Bai et al. (2019a, b)	Xingtai	9.8000	13.8968	0.8952	24.3531	75
	Anyang	11.5542	17.2382	0.8950	22.7417	77
	Beijing	7.1375	10.1139	0.8337	36.0034	82
CEEMD-RF Liu and Sun (2019)	Xingtai	13.0316	17.7063	0.8299	31.9986	52
	Anyang	14.6151	21.1709	0.8410	30.8094	64
	Beijing	12.7847	16.9727	0.5301	72.8125	80
EMD-GRU Huang et al. (2021)	Xingtai	12.6310	16.4351	0.8534	29.6622	50
	Anyang	16.5346	22.1601	0.8258	33.6153	62
	Beijing	13.4090	17.5617	0.4969	69.5675	67
CEEMD-GWO-SVR Niu et al. (2016)	Xingtai	13.0920	17.1708	0.8400	35.1119	50
	Anyang	14.1738	19.9151	0.8717	31.1315	67
	Beijing	12.6228	15.8322	0.59116	78.4423	88
VMD-SE-LSTM Wu and Lin (2019)	Xingtai	8.9445	13.2124	0.9042	23.4516	75
	Anyang	10.3212	12.2382	0.9134	20.1415	82
	Beijing	7.5357	9.2145	0.8628	33.1348	84
CEEMDAN-FE-mRMR-GWO-ANN	Xingtai	15.1254	17.3254	0.8012	29.4110	62
	Anyang	15.8183	17.8753	0.8311	25.1539	68
	Beijing	12.1156	14.7865	0.8458	27.6423	80
CEEMDAN-FE-mRMR-GWO-LSTM	Xingtai	7.6823	9.6855	0.9492	19.3643	87
	Anyang	6.9228	8.8539	0.9722	14.0953	90
	Beijing	5.1911	6.2419	0.9364	25.8494	93

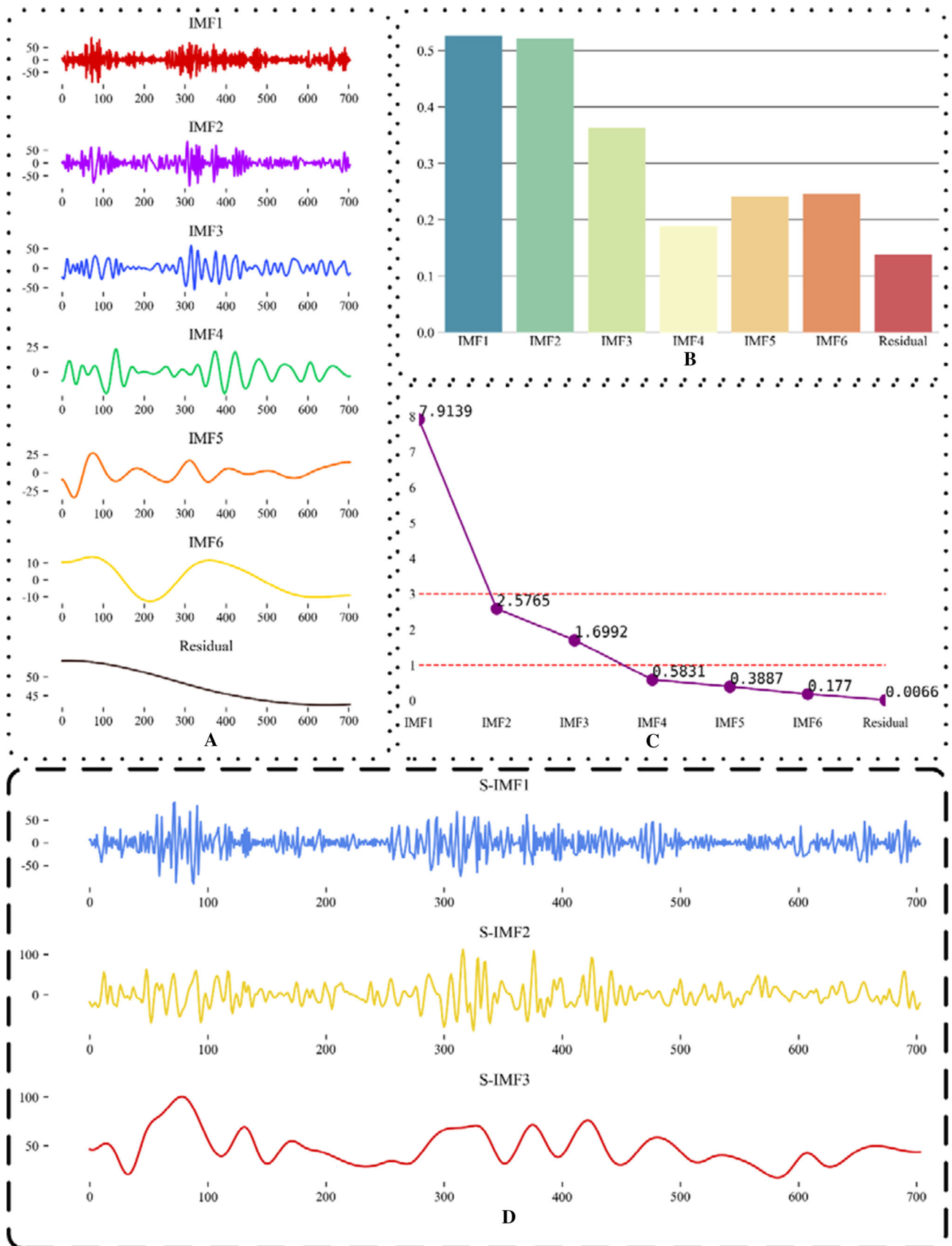


Fig. 5 Decomposition and reconstruction results of Beijing dataset

superiority and effectiveness of the GWO optimized LSTM. The second set of experiments introduces nine comparison models, eight of which are the results from eight excellent papers in the same research area in recent years. As shown in Tables 4 and 5, the prediction accuracy of the hybrid model proposed in this paper outperforms all the comparative models.

In the first set of comparison experiments, the number of hidden layers of LSTM is set to 1, and the number of hidden layer neurons, batch size and Dropout rate of hidden layers are determined by using the optimization algorithm. The number of hidden layer neurons is [2,128], the batch size is [2,256], and the Dropout rate is [0,0.6]. Taking the data of Xingtai as an example, the number of iterations of the optimization algorithm is 50, and the iteration time and prediction performance are shown in the Table 4. GWO can find the optimal hyperparameters of LSTM faster and more effectively than other optimization algorithms, which can effectively improve the prediction of the model.

In the second set of comparison experiments, the SVR model based on random forest (RF) for feature selection has the worst prediction performance. Although the introduction of exogenous variables can improve the robustness and prediction accuracy of the model, this can be achieved only on the basis of a reasonable treatment of PM_{2.5} concentration series. In addition, the introduction of exogenous variables can easily lead to problems such as error accumulation, which in turn affects the prediction accuracy. The advantages of ANN and LSTM in handling nonlinear sequences make their prediction accuracy better than RF-SVR. However, the advantage of LSTM in temporal patterns does not make its prediction performance significantly better than ANN. This is because the original PM_{2.5} concentration sequence is more complex and more volatile, which makes it more difficult for the LSTM to learn valid information from it. Therefore, we additionally constructed CEEMDAN-FE-mRMR-GWO-ANN for comparison. The results show that the hybrid model proposed in this paper has higher prediction performance. The effective

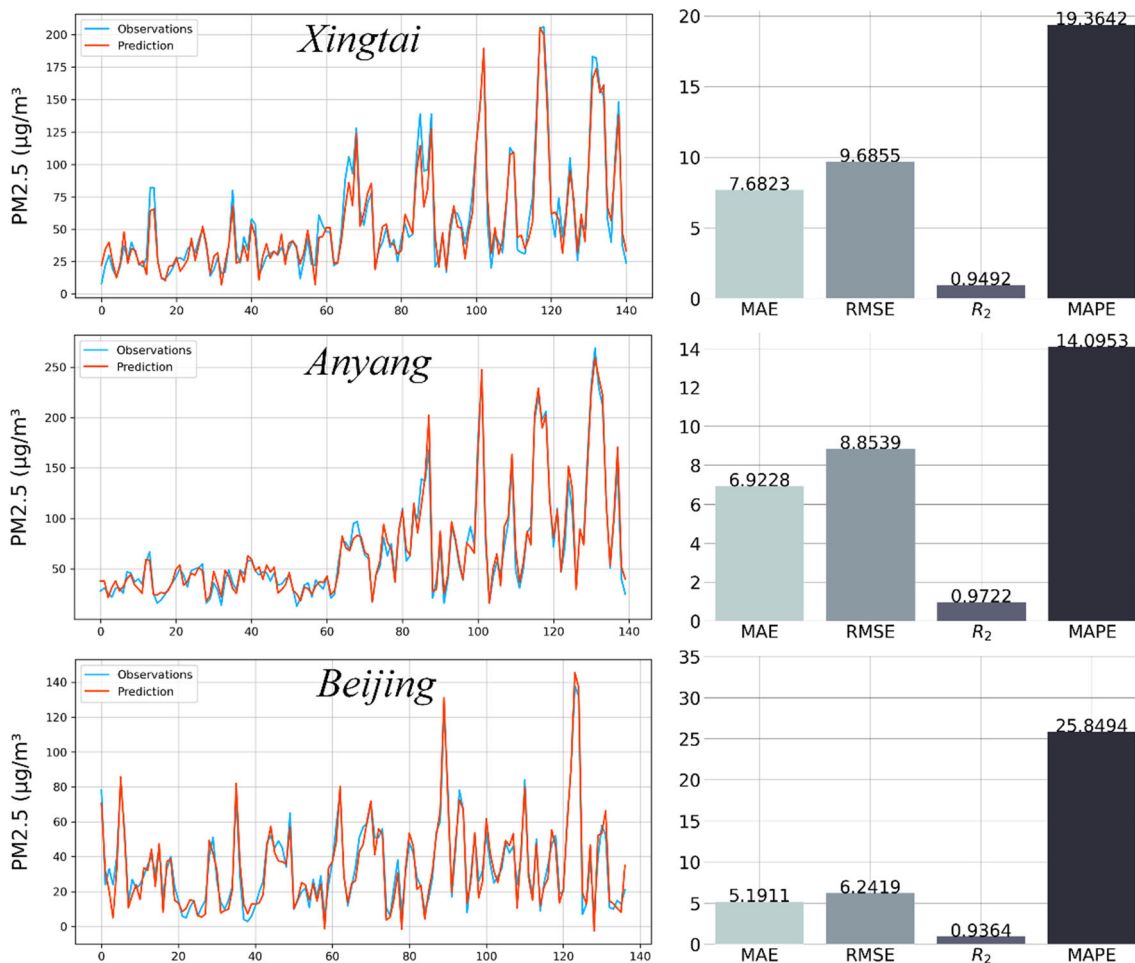


Fig. 6 Prediction accuracy of the hybrid model on three data sets

processing of $PM_{2.5}$ sequences and the inclusion of $PM_{2.5}$ influencing factors make it easier for the LSTM to capture the long-term dependence in the data, which further improves the prediction performance.

As shown in Table 5, models 4 through 8 are decomposed integrated frameworks, and these models show a significant improvement in predictive performance compared to models 1 through 3. Taking Anyang as an example, the MAE, RMSE and MAPE of EEMD-LSTM are improved by 50.82%, 51.81% and 52.96%, respectively, compared with LSTM. The warning accuracy of EEMD-LSTM model reached 77%, while the warning accuracy of LSTM model was only 55%. Among these five decomposition integration frameworks, CEEMD-GWO-SVR has the worst prediction performance, which is because SVR is less capable of handling nonlinear time series than RF and LSTM. Although the prediction performance of CEEMD-RF and EMD-GRU is good in two cities, Xingtai and Anyang, the warning accuracy is low. In addition, the prediction performance of both models in the Beijing dataset shows a substantial decrease, which indicates that the model cannot be effectively applied to datasets in different cities and the stability of prediction is poor. Among these five decomposition integration frameworks, the VMD-SE-LSTM and EEMD-LSTM showed good prediction performance, early warning accuracy and prediction stability in different datasets. And the hybrid prediction framework proposed in this paper, with the Anyang dataset, improves the MAE, RMSE and MAPE by 32.92%, 27.65% and 30.02%, respectively, compared to the EEMD-LSTM. And compared with MAE, RMSE and MAPE of VMD-SE-LSTM, the improvement is 40.08%, 48.64% and 38.02%, respectively. In addition, the hybrid prediction model proposed in this paper outperforms VMD-SE-LSTM and EEMD-LSTM in terms of warning accuracy and prediction stability on different data sets.

In summary, the hybrid prediction framework proposed in this paper outperforms all comparative models in terms of prediction accuracy, warning accuracy and prediction stability. This proves that the hybrid prediction framework is suitable for air pollution forecasting and warning.

5 Conclusion

In order to prevent air pollution and protect human health, this paper proposes a multi-factor, multi-scale, and intelligent optimization based two-stage deep learning hybrid framework for air pollution forecasting and warning. First, feature extraction is performed using CEEMDAN and FE to decompose and reconstruct the original sequence into three components. Then, the mRMR algorithm is used for feature selection of the influencing factors to filter out the

influencing factors that have a greater impact on each reconstructed component. Then, a two-stage deep learning hybrid framework is proposed in this paper to predict and nonlinearly integrate each reconstructed component. Finally, based on the proposed hybrid model, air pollution prediction and early warning are achieved. The results show that: (1) the feature extraction methods based on CEEMDAN and FE can effectively discover the multiscale relationships in $PM_{2.5}$ sequences, reduce the complexity of prediction; (2) the mRMR-based influence factor selection method can not only reduce the complexity of data, but also improve the performance of the model; (3) A two-stage GWO-LSTM can effectively improve the prediction accuracy; (4) the model has good practical significance and application value, and can realize effective forecasting and early warning of air pollution.

Using $PM_{2.5}$ concentration data from Xingtai, Anyang and Beijing as the study samples, the empirical results statistically support the effectiveness of the proposed hybrid model in terms of prediction accuracy and robustness, and the model outperforms all comparative models.

In conclusion, the hybrid framework has advantages in prediction stability, prediction accuracy and accuracy of air pollution warning. Not limited to air pollution prediction studies, the framework can be extended to other complex systems to verify its generality and versatility. However, no technique is perfect and flawless. As the theory matures and research progresses, more advanced and effective algorithms will be proposed. In the future, on top of the hybrid framework proposed in this paper, more novel and effective algorithms can be added to further improve the prediction performance of the model. In addition, only daily $PM_{2.5}$ concentration data were considered in this study, so prediction of air pollutant concentrations on other time scales is also an option for future research.

Acknowledgements This research was supported by the National Natural Science Foundation of China (Grant Nos. 71971122 and 71501101)

Funding Funding was provided by National Natural Science Foundation of China (71971122, 71501101).

Declarations

Conflict of 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.

References

- Ahmed AAM, Deo RC, Ghahramani A, Raj N, Feng Q, Yin Z et al (2021) LSTM integrated with Boruta-random forest optimiser for soil moisture estimation under RCP4.5 and RCP8.5 global

- warming scenarios. *Stoch Environ Res Risk Assess* 35:1851–1881
- Bai Y, Zeng B, Li C, Zhang J (2019a) An ensemble long short-term memory neural network for hourly PM_{2.5} concentration forecasting. *Chemosphere* 222:286–294
- Bai Y, Li Y, Zeng B, Li C, Zhang J (2019b) Hourly PM_{2.5} concentration forecast using stacked autoencoder model with emphasis on seasonality. *J Clean Prod* 224:739–750
- Barzegar R, Aalami MT, 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
- Biancofiore F, Busilacchio M, Verdecchia M, Tomassetti B, Aruffo E, Bianco S et al (2017) Recursive neural network model for analysis and forecast of PM₁₀ and PM_{2.5}. *Atmos Pollut Res* 8:652–659
- Bo L, Shuo Y, Jianqiang L, Yong L, Jianlei L, Guangzhi Q (2021) A spatiotemporal recurrent neural network for prediction of atmospheric PM_{2.5}: a case study of Beijing. *IEEE Trans Comput Soc Syst* 8:578–588
- Chen Z, Cai J, Gao B, Xu B, Dai S, He B et al (2017) Detecting the causality influence of individual meteorological factors on local PM_{2.5} concentration in the Jing-Jin-Ji region. *Sci Rep* 7:40735
- Donnelly A, Misstear B, Broderick B (2015) Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmos Environ* 103:53–65
- Feng X, Li Q, Zhu Y, Hou J, Jin L, 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
- Feng Q, Sun X, Hao J, Li J (2021) Predictability dynamics of multifactor-influenced installed capacity: a perspective of country clustering. *Energy* 214:118831
- García Nieto PJ, Sánchez Lasheras F, García-Gonzalo E, de Cos Juez FJ (2018) PM₁₀ concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: a case study. *Sci Total Environ* 621:753–761
- González-Enrique J, Ruiz-Aguilar JJ, Moscoso-López JA, Urda D, Turias IJ (2021) A comparison of ranking filter methods applied to the estimation of NO₂ concentrations in the Bay of Algeciras (Spain). *Stoch Environ Res Risk Assess* 35:1999–2019
- Goudarzi G, Hopke PK, Yazdani M (2021) Forecasting PM_{2.5} concentration using artificial neural network and its health effects in Ahvaz Iran. *Chemosphere* 283:131285
- Gu K, Xia Z, Qiao J (2020) Stacked selective ensemble for PM_{2.5} forecast. *IEEE Trans Instrum Meas* 69:660–671
- Guo Y, Cao X, Liu B, Peng K (2020) Chaotic time series prediction using LSTM with CEEMDAN. *J Phys Conf Ser* 1617:012094
- Han Z, Ueda H, An J (2008) Evaluation and intercomparison of meteorological predictions by five MM5-PBL parameterizations in combination with three land-surface models. *Atmos Environ* 42:233–249
- Hu X, Waller LA, Al-Hamdan MZ, Crosson WL, Estes MG, Estes SM et al (2013) Estimating ground-level PM_{2.5} concentrations in the southeastern U.S. using geographically weighted regression. *Environ Res* 121:1–10
- Huang W, Hu M (2018) Estimation of the Impact of Traveler Information Apps on Urban Air Quality Improvement. *Engineering* 4:224–229
- Huang L, Wang J (2018) Forecasting energy fluctuation model by wavelet decomposition and stochastic recurrent wavelet neural network. *Neurocomputing* 309:70–82
- Huang N, Shen Z, Long S, Wu MLC, Shih 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 Ser A* 454:903–995
- Huang G, Li X, Zhang B, Ren J (2021) PM_{2.5} concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. *Sci Total Environ* 768:144516
- Ishak A (2016) Variable selection based on statistical learning approaches to improve PM₁₀ concentration forecasting. *J Environ Inf* 30:79–94
- Kim K, Park RJ, Kim K, Na H (2010) Weekend effect: anthropogenic or natural? *Geophys Res Lett* 37:L09808
- Kisi O, Alizamir M (2018) Modelling reference evapotranspiration using a new wavelet conjunction heuristic method: wavelet extreme learning machine vs wavelet neural networks. *Agric For Meteorol* 263:41–48
- Konovalov IB, Beekmann M, Meleux F, Dutot A, Foret G (2009) Combining deterministic and statistical approaches for PM₁₀ forecasting in Europe. *Atmos Environ* 43:6425–6434
- Li H, Guo S, Li C, Sun J (2013) A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. *Knowl-Based Syst* 37:378–387
- Li L, Cen Z, Tseng M, Shen Q, Ali MH (2021a) Improving short-term wind power prediction using hybrid improved cuckoo search arithmetic-Support vector regression machine. *J Clean Prod* 279:123739
- Li J, Hao J, Sun X, Feng Q (2021b) Forecasting China's sovereign CDS with a decomposition reconstruction strategy. *Appl Soft Comput* 105:107291
- Li J, Hao J, Feng Q, Sun X, Liu M (2021c) Optimal selection of heterogeneous ensemble strategies of time series forecasting with multi-objective programming. *Expert Syst Appl* 166:114091
- Lin Y, Yan Y, Xu J, Liao Y, Ma F (2021) Forecasting stock index price using the CEEMDAN-LSTM model. *N Am J Econ Finance* 57:101421
- Liu D, Sun K (2019) Short-term PM_{2.5} forecasting based on CEEMD-RF in five cities of China. *Environ Sci Pollut Res* 26:32790–32803
- Liu T, Lau AKH, Sandbrink K, Fung JCH (2018) Time series forecasting of air quality based on regional numerical modeling in Hong Kong. *J Geophys Res Atmos* 123:4175–4196
- Luo H, Wang D, Yue C, Liu Y, Guo H (2018) Research and application of a novel hybrid decomposition-ensemble learning paradigm with error correction for daily PM₁₀ forecasting. *Atmos Res* 201:34–45
- Ma J, Cao Y, Xu J, Qu Y, Yu Z (2021) PM_{2.5} concentration distribution patterns and influencing meteorological factors in the central and eastern China during 1980–2018. *J Clean Prod* 311:127565
- Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. *Adv Eng Softw* 69:46–61
- Niu M, Wang Y, Sun S, Li Y (2016) A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM_{2.5} concentration forecasting. *Atmos Environ* 134:168–180
- Nourani V, Farboudfam N (2019) Rainfall time series disaggregation in mountainous regions using hybrid wavelet-artificial intelligence methods. *Environ Res* 168:306–318
- Ogliari E, Guilizzoni M, Giglio A, Pretto S (2021) Wind power 24-h ahead forecast by an artificial neural network and an hybrid model: comparison of the predictive performance. *Renew Energy* 178:1466–1474
- Ren M, Sun W, Chen S (2021) Combining machine learning models through multiple data division methods for PM_{2.5} forecasting in Northern Xinjiang, China. *Environ Monit Assess* 193:476
- Samal KKR, Babu KS, Das SK (2021) Multi-directional temporal convolutional artificial neural network for PM_{2.5} forecasting

- with missing values: a deep learning approach. *Urban Clim* 36:100800
- Shin U, Park S, Park J, Koo J, Yoo C, Kim S, Lee J (2021) Predictability of PM_{2.5} in Seoul based on atmospheric blocking forecasts using the NCEP global forecast system. *Atmos Environ* 246:118141
- Sulaiman MH, Mustaffa Z, Mohamed MR, Aliman O (2015) Using the gray wolf optimizer for solving optimal reactive power dispatch problem. *Appl Soft Comput* 32:286–292
- Sun W, Li Z (2020) Hourly PM_{2.5} concentration forecasting based on mode decomposition-recombination technique and ensemble learning approach in severe haze episodes of China. *J Clean Prod* 263:121442
- Sun X, Hao J, Li J (2022) Multi-objective optimization of crude oil-supply portfolio based on interval prediction data. *Ann Oper Res* 309:611–639
- Tai APK, Mickley LJ, Jacob DJ (2010) Correlations between fine particulate matter (PM_{2.5}) and meteorological variables in the United States: implications for the sensitivity of PM_{2.5} to climate change. *Atmos Environ* 44:3976–3984
- Tao J, Zhang L, Engling G, Zhang R, Yang Y, Cao J et al (2013) Chemical composition of PM_{2.5} in an urban environment in Chengdu, China: importance of springtime dust storms and biomass burning. *Atmos Res* 122:270–283
- Wang Y, Zhuang G, Sun Y, An Z (2006) The variation of characteristics and formation mechanisms of aerosols in dust, haze, and clear days in Beijing. *Atmos Environ* 40:6579–6591
- Wang P, Liu Y, Qin Z, Zhang G (2015) A novel hybrid forecasting model for PM₁₀ and SO₂ daily concentrations. *Sci Total Environ* 505:1202–1212
- Wang D, Liu Y, Luo H, Yue C, Cheng S (2017) Day-ahead PM_{2.5} concentration forecasting using WT-vmd based decomposition method and back propagation neural network improved by differential evolution. *Int J Environ Res Public Health* 14:764
- Wang J, Sun X, Cheng Q, Cui Q (2021) An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting. *Sci Total Environ* 762:143099
- Wu Z, Huang N (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv Adapt Data Anal* 1:1–41
- Wu Q, Lin H (2019) Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. *Sustain Cities Soc* 50:2210–6707
- Wu Z, Huang NE, Chen X (2009) The multi-dimensional ensemble empirical mode decomposition method. *Adv Adapt Data Anal* 01:339–372
- Wu J, Zhang P, Yi H, Qin Z (2016) What causes haze pollution? An empirical study of PM_{2.5} concentrations in Chinese cities. *Sustainability* 8:132
- Wu H, Liu H, Duan Z (2020) PM_{2.5} concentrations forecasting using a new multi-objective feature selection and ensemble framework. *Atmos Pollut Res* 11:1187–1198
- Xu Y, Du P, 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
- Xu Y, Huang Y, Guo Z (2021) Influence of AOD remotely sensed products, meteorological parameters, and AOD–PM_{2.5} models on the PM_{2.5} estimation. *Stoch Environ Res Risk Assess* 35:893–908
- Yan D, Kong Y, Bin Y, Xiang H (2021) Spatio-temporal variation and daily prediction of PM_{2.5} concentration in world-class urban agglomerations of China. *Environ Geochem Health* 43:301–316
- Yang J, Yan R, Nong M, Liao J, Li F, Sun W (2021) PM_{2.5} concentrations forecasting in Beijing through deep learning with different inputs, model structures and forecast time. *Atmos Pollut Res* 12:101168
- Yoo J, Lee Y, Kim D, Jeong M, Stockwell WR, Kundu PK et al (2014) New indices for wet scavenging of air pollutants (O₃, CO, NO₂, SO₂, and PM₁₀) by summertime rain. *Atmos Environ* 82:226–237
- You Q, Kang S, Flügel W-A, Sanchez-Lorenzo A, Yan Y, Xu Y et al (2009) Does a weekend effect in diurnal temperature range exist in the eastern and central Tibetan Plateau? *Environ Res Lett* 4:045202
- Zhang G, Eddy Patuwo BY, Hu M (1998) Forecasting with artificial neural networks: the state of the art. *Int J Forecast* 14:35–62
- Zhang Z, Zhang X, Gong D, Quan W, Zhao X, Ma Z et al (2015) Evolution of surface O₃ and PM_{2.5} concentrations and their relationships with meteorological conditions over the last decade in Beijing. *Atmos Environ* 108:67–75
- Zhang L, Lin J, Qiu R, Hu X, Zhang H, Chen Q et al (2018) Trend analysis and forecast of PM_{2.5} in Fuzhou, China using the ARIMA model. *Ecol Ind* 95:702–710
- Zhu S, Lian X, Liu H, Hu J, Wang Y, Che J (2017) Daily air quality index forecasting with hybrid models: a case in China. *Environ Pollut* 231:1232–1244
- Zhu S, Lian X, Wei L, Che J, Shen X, Yang L et al (2018) PM_{2.5} forecasting using SVR with PSO-GSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. *Atmos Environ* 183:20–32

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.