Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 9407104, 15 pages
https://doi.org/10.1155/2023/9407104

Research Article

A New Hybrid Forecasting Model Based on Dual Series Decomposition with Long-Term Short-Term Memory

Hao Tang,[1,2] Uzair Aslam Bhatti,[1,2] Jingbing Li,[1,2] Shah Marjan,[3] Mehmood Baryalai,[4] Muhammad Assam,[5] Yazeed Yasin Ghadi,[6] and Heba G. Mohamed[7]

[1]*School of Information and Communication Engineering, Hainan University, Haikou 570100, China*
[2]*State Key Laboratory of Marine Resource Utilization in the South China Sea, Hainan University, Haikou 570100, China*
[3]*Department of Software Engineering, Balochistan University of Information Technology, Engineering, and Management Sciences (BUITEMS), Quetta, Pakistan*
[4]*Department of Information Technology, Balochistan University of Information Technology, Engineering, and Management Sciences (BUITEMS), Quetta, Pakistan*
[5]*Department of Software Engineering, University of Science and Technology Bannu, Bannu, Khyber Pakhtunkhwa, Pakistan*
[6]*Department of Computer Science, Al Ain University, Al Ain, UAE*
[7]*Department of Electrical Engineering, College of Engineering, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia*

Correspondence should be addressed to Uzair Aslam Bhatti; uzairaslambhatti@hotmail.com, Jingbing Li; jingbingli2008@hotmail.com, and Heba G. Mohamed; hegmohamed@pnu.edu.sa

Received 29 October 2022; Revised 14 March 2023; Accepted 29 May 2023; Published 22 June 2023

Academic Editor: Paolo Gastaldo

In recent years, ozone (O_3) has gradually become the primary pollutant plaguing urban air quality. Accurate and efficient ozone prediction is of great significance to the prevention and control of ozone pollution. The air quality monitoring network provides multisource pollutant concentration monitoring data for ozone prediction, but ozone prediction based on multisource monitoring data still faces the challenges of each station's series of data. Aiming at the problems of low prediction accuracy and low computational efficiency in traditional atmospheric ozone concentration prediction, ozone concentration prediction using dual series decomposition was proposed by variational mode decomposition (VMD), ensemble empirical mode decomposition (EEMD), and long short-term memory (LSTM). First, the historical data series of Nanjing air quality monitoring stations is decomposed by VMD, and then the EEMD algorithm is applied to the residual of VMD to obtain several characteristic intrinsic mode function (IMF) components; each characteristic IMF component is trained by LSTM to obtain the prediction result of each component, and then the final result can be obtained by linear superposition. The proposed method achieved the best results with $R^2 = 99\%$, MSE = 5.38, MAE = 4.54, and MAPE = 3.12. Because LSTM has strong adaptive learning ability and good memory function, it has the learning advantage of long-term memory for long-term data, and the prediction results are more accurate. According to the data, the proposed method is superior to the baseline models in terms of statistical metrics. As a result, the proposed hybrid method can serve as a reliable model for ozone forecasting.

1. Introduction

Ozone (O_3) is one of the six major pollutants in the air, and when the ozone concentration in the atmosphere is too high, the ecological environment deteriorates and adversely affects human health [1, 2]. Ozone is a trace gas in the earth's atmosphere. It is formed when oxygen molecules in the atmosphere are decomposed into oxygen atoms by solar radiation, and the oxygen atoms combine with the surrounding oxygen molecules. It contains 3 oxygen atoms, and its chemical formula is O_3. Ozone pollution has special conditions for its formation. Under the conditions of high

temperatures, sufficient sunshine, and dry air, VOCs and NOx in the air "meet" and produce photochemical reactions, which are easy to generate ozone pollution [3, 4]. In recent years, the concentration of ozone and other air pollutants has been changing continuously [5–7]. There are two main reasons for the analysis of the changes in ozone in this study. One is the increase in pollutant emissions caused by frequent human activities, and the other is the weather [8]. The stronger the sunlight, the more the ozone will be produced [9]. As people pay more and more attention to the degree of ozone pollution, it is very important for researchers to forecast the ozone concentration in a timely and effective manner [10].

In recent years, China's ozone pollution problems have become increasingly apparent. Beijing–Tianjin–Hebei and surrounding areas, the Yangtze River Delta region [11], and other regions with ozone concentrations show an upward trend. Especially in the summer and autumn, ozone has become the primary pollutant in some cities. Ozone, nitrogen oxides (NOx), volatile organic compounds (VOCs), and other pollutants in the atmosphere can have a photochemical reaction with secondary pollutants [12], resulting in a strong stimulating effect on the human cardiovascular and respiratory systems, leading to the occurrence of a variety of diseases. In addition, ozone can also cause serious harm to the environment [13, 14]. Advance predictions of ozone pollution notify governments about implementing environmental management decisions.

At present, there are many studies on ozone, and many scholars are also committed to the forecasting of ozone concentrations. The research on ozone at this stage is mainly divided into two aspects:

The first aspect is to study the connection between changes in ozone concentrations and human health, the ecological environment, crops, etc. For example, Jiang et al. [15] studied the effect of ozone concentrations in Fuzhou on the risk of death from circulatory diseases, and the results showed that short-term exposure to ozone increased the risk of death from these diseases. Zhao et al. discussed how excessive ozone concentrations on the ground damage the ecological environment, damage human health, reduce crop yields, and cause certain economic losses [16]. Chen et al. [17] explored the link between short-term exposure to ozone and lung function and airway inflammation. Zhang et al. [18] studied the association between ozone concentrations in Yangzhou and the daily deaths of residents. These studies have shown that ozone concentration exceeding the standard negatively impacts human health, the ecological environment, crops, and so on [18].

The second aspect is the forecast and early warning analysis of ozone, mainly the establishment of statistical prediction models. The first is to establish a regression prediction model by analyzing indicators related to ozone concentration. For example, Shams et al. [19] selected NO_2, SO_2, air temperature, water pressure, and other indicators as forecasting factors to establish a regression model that could better reflect the average daily change in ozone concentration. Gong et al. [20] selected meteorological factors, such as humidity and wind speed, to establish regression models, predict ozone concentration in Xiamen, and establish an evaluation system. Zhang and Ma [21] used meteorological factors, such as wind direction, temperature, humidity, and other meteorological factors as input variables to predict ozone concentration through a back propagation neural network, and the results showed that the model based on meteorological factors helped to improve the prediction performance of the model. The artificial intelligence (AI) method uses machine learning technology to train historical data, has higher prediction accuracy in nonlinear time series data [22], and has been successfully applied to solve nonlinear regression estimation problems. Typical models include artificial neural networks (Masood and Ahmad [23]; Masood and Ahmad [24]), genetic algorithms, support vector machines, random forests, and the AdaBoost model [25, 26].

However, traditional AI technology cannot describe the interdependence between time series data, and its prediction accuracy of time series data is limited [27]. The deep recurrent neural network (RNN) can handle the interdependence between time series data due to its embedded feedback and cyclic structure [28] [29]. Tsai et al. [30] achieved good results in predicting different air pollutants, such as $PM_{2.5}$, PM_{10}, SO_2, and NO_2 based on the RNN model. However, RNN cannot solve the long-term dependency problem. As a variant of the RNN network, long short-term memory (LSTM) can effectively describe time series data by introducing memory units into the network structure [31, 32]. LSTM not only focuses on event-related semantic information but also considers the temporal effects of important events in the past. Xayasouk et al. [33] applied LSTM for the prediction of air quality concentration using the autoencoder. To anticipate $PM_{2.5}$ concentrations, taking into account the impact of wind direction and speed on the variations in spatial-temporal $PM_{2.5}$, Liu et al. [34] presented a novel wind-sensitive attention mechanism with an LSTM neural network model. Compared with other forecasting methods, Liu et al. [25] used LSTM for stock price forecasting of the CSI 300 Index, and the results showed that the LSTM model had a better forecasting effect than the support vector regression and AdaBoost models. Bathla [35] used the LSTM network to predict a data series, and the results showed that the LSTM network performed better than the traditional GARCH model and SVR model in longer range volatility prediction. The above literature shows that the LSTM model has certain advantages in predicting complex time series data. Therefore, this paper selects LSTM as the main component of the model.

In order to overcome the limitations of traditional AI models, another type of forecasting method, which has achieved good development, is to develop hybrid models. In most hybrid models, signal processing methods are used to decompose the time series, and AI methods are used to predict the decomposed components. Typical sequence preprocessing methods include wavelet decomposition and empirical mode decomposition (EMD) [36]. The EMD algorithm does not depend on any basis function and is essentially different from wavelet decomposition. It has significant advantages in dealing with nonstationary and nonlinear complex signals. For example, Jin et al. [37] used the EMD algorithm to decompose a trend and analyze the

periodic fluctuations of the air quality parameters. To a certain extent, it reflects the various cyclical variations in time series data. However, since EMD is prone to mode aliasing, Amanollahi and Ausati [38] used an ensemble empirical mode decomposition (EEMD) algorithm for air quality prediction. EEMD has been widely used in air quality forecasting. Du et al. [39] used the EEMD method to study the tourist impact on air pollution in Zhangjiajie, China. Due to the lack of a mathematical foundation, the inability to separate components with similar frequencies, and the over-envelope and under-envelope problems of the EEMD method, its decomposition effect is limited [40]. As an improved decomposition technology, variational mode decomposition (VMD) can adaptively decompose the effective components corresponding to each center frequency in the frequency domain, and its decomposition accuracy is higher. The VMD decomposition method is more effective for feature selection in prediction models and has been successfully applied to air quality by decomposition of series [41]. Therefore, this paper also adopts the VMD technique as the main decomposition technique for modeling.

Looking at the previous studies, it can be seen that in the prediction of a single AI model, the LSTM model has achieved excellent prediction results; at the same time, the prediction effect of all the combined models is better than that of the single AI method. In some studies using VMD for combined model prediction, the important component information contained in the residual term after the original sequence is decomposed by VMD is ignored. Therefore, this paper considers secondary decomposition of the complex signal contained in the residual term after VMD classification to improve the prediction accuracy of the residual term and proposes a new fusion VMD–EEMD dual decomposition method, combining it as an input to LSTM to develop a VMD–EEMD–LSTM-based ozone prediction model. The objective of this study is to develop a new time series-based machine learning model which is good for prediction than other methods.

The major work in this paper includes the following:

(i) Ozone (O_3) data series have high complexity and much variation with respect to changes in the environment and require secondary decomposition of the highly complex components. VMD decomposes the original complex ozone time series data into multiple subsignal components according to the frequency domain.

(ii) After VMD, original sequence and different variational mode functions (VMF) components and residual items are obtained; the residual series is decomposed by EEMD and combined with the LSTM model for combined prediction analysis.

(iii) The proposed VMD–EEMD–LSTM model is compared with other machine learning models. Beside that results among different stations of the Nanjing city are also compared to verify the effectiveness of the model at different locations. The proposed model performs better than other methods as validated by different indicators, such as mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE).

The structure of the rest of this paper is as follows: Section 2 briefly introduces the hybrid model's construction. Section 3 shares the results and analysis of Nanjing city. Section 4 is a discussion. Section 5 shares conclusions.

2. Method (Proposed Algorithm)

Before constructing the VMD–EEMD–LSTM portfolio model to predict the ozone change, it is necessary to briefly describe the components of the model portfolio: EEMD, VMD technology, and the LSTM neural network.

2.1. EEMD. Wu and Huang [42] added a very small-amplitude white noise sequence to the original time series and extended the EEMD technique. The decomposition algorithm made full use of the frequency-balanced distribution characteristics of the white noise. The obtained intrinsic mode function (IMF) is averaged to cancel the added white noise, thereby improving the mode aliasing problem. The decomposition steps are as follows:

Step 1: This will satisfy the normal distribution of white noise. An equal-length sequence of columns $n_i(t)$ is added to the original time series $x(t)$ multiple times, i.e.,

$$x_i(t) = x(t) + n_i(t). \tag{1}$$

In the formula, $x_i(t)$ is the time series after adding white noise for the ith time.

Step 2: Perform EMD on the time series after adding white noise to obtain the IMF component $C_{i,j}(t)$ and $r_i(t)$ residual term, where $C_{i,j}(t)$ is the jth obtained by EMD after adding white noise for the ith time, an IMF component.

$$x_i(t) = \sum_{j=1}^{J} C_{i,j}(t) + r_i(t). \tag{2}$$

Step 3: Take the average value of each component $C_{i,j}(t)$ by taking advantage of the characteristic of zero mean value between uncorrelated random sequences to cancel the influence of the white noise added multiple times on the real IMF component, and finally, obtain the IMF component decomposition.

$$C_j(t) = \frac{1}{N} \sum_{i=1}^{N} C_{i,j}(t). \tag{3}$$

In the formula, $C_j(t)$ is the jth IMF component obtained after EEMD, and N is the number of white noise sequences.

Step 4, further obtain the final decomposition result of EEMD, namely,

$$x(t) = \sum_{j=1}^{J} C_j(t) + r(t). \tag{4}$$

The IMF component $C_j(t)$ is the information trend of different frequency segments from high to low in the time series, and $r(t)$ is the overall residual term.

2.2. VMD. The core principle of VMD technology is to use an adaptive and quasi-orthogonal decomposition method to decompose the original input signal into k modal components u_k, obey the center frequency and limited bandwidth, and minimize the sum of the bandwidth estimates of all modes [43]. The VMD signal decomposition process is also the solution process for the variational constraint problem. The model expression for the variational constraint problem is shown in the following equation:

$$\left\{ \begin{array}{l} \min\limits_{\langle u_k \rangle, \langle w_k \rangle} \left\{ \sum\limits_k \| \partial_t \left[\left(\delta(t) + \dfrac{J}{\pi t} \right) * u_k(t) \right] e^{-jw_k t} \|_2^2 \right\}, \\ \text{s.t.} \sum\limits_k u_k = f. \end{array} \right. \tag{5}$$

In the formula, $u_k = \{u_1, \ldots, u_k\}$ is the modal component VMF obtained after decomposition and $w_k = \{w_1, \ldots, w_k\}$ are the center frequencies corresponding to the VMF, respectively; $*$ is the convolution symbol; ∂_t is the partial derivative of t, $\delta(t)$ is the shock function; f is the original input signal. The analytic signal of $u_k(t)$ related to it is obtained by Hilbert transform, and then its unilateral spectrum is obtained; the estimated value of the center frequency of each mode is adjusted by multiplying the exponential term $e^{-jw_k t}$, and the spectrum of the mode is adjusted to the fundamental frequency band. In order to obtain the optimal solution to the above constrained variational problem, it needs to be transformed into an unconstrained problem to solve. By introducing the Lagrangian multiplication operator $\lambda(t)$ and the quadratic penalty factor α, the constrained variational problem is transformed into an unconstrained variational problem of the following form:

$$\begin{aligned} L(\{u_k\}, \{w_k\}, \lambda) \coloneqq \alpha \sum_k \| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) u_k(t) \right], \\ e^{-jw_k t} \|_2^2 + \| f(t) - \sum_k u_k(t), \\ \| \sum_2^2 + \left(\lambda(t), f(t) - \sum_k u_k(t) \right). \end{aligned} \tag{6}$$

In the formula, the quadratic penalty factor α can ensure the accuracy of signal reconstruction in the presence of Gaussian noise; the Lagrangian operator can be used to maintain strict constraints. Further, the alternate direction method of multipliers iterative search is used to obtain the saddle point of the above Lagrangian function, that is, to obtain the optimal solution of the constrained variational problem of formula (6), its VMF_{uk} and center frequency. The expressions of w_k are as follows:

$$\begin{aligned} \hat{u}_k^{n+1}(w) &= \frac{\hat{f}(w) - \sum_{i \neq k} \hat{u}_i(w) + (\hat{\lambda}(w)/)}{1 + 2\alpha(w - ww_k)^2}, \\ \hat{w}_k^{n+1} &= \frac{\int_0^\infty w |\hat{u}_k(w)|^2 \mathrm{d}w}{\int_0^\infty \hat{u}_k |w|^2 \mathrm{d}w}. \end{aligned} \tag{7}$$

The specific implementation steps of the VMD method are as follows:

Step 1: Set the initialization values of parameters such as modal components and center frequency $\{u_k^1\}\{w_k^1\}, \lambda^1, n = 0$ and select the appropriate number K of modal components.

Step 2: Update the values of u_k and w_k, respectively, according to formulas (7) and (8).

Step 3: Update the value of λ

$$\hat{\lambda}^{n+1} = \hat{\lambda}^n + \tau \left[\hat{f}(w) - \sum_k \hat{u}_k^{n+1}(w) \right]. \tag{8}$$

Step 4: Given the judgment accuracy, $\varepsilon > 0$ if the following conditions are met:

$$\sum_k \left\| \hat{u}_k^{n+1} - \hat{u}_k^n \right\|_2^2 / \left\| \hat{u}_k^n \right\|_2^2 < \varepsilon. \tag{9}$$

Then, stop the iteration; otherwise, go back to step 2. In the above formula, $\hat{u}_k^n(w)$, $\hat{f}(w)$ and $\hat{\lambda}^n$(w) are Fourier transforms corresponding to n k, $\hat{u}_k^n$, $f(t)$, and λ_n, respectively.

2.3. LSTM Neural Network. The traditional RNN has achieved good results in processing time series because it considers the self-correlation characteristics of time series, but the back propagation algorithm used by RNN results in gradient explosion or gradient disappearance, which cannot describe the long-term dependency problem [22, 44]. A descriptive implementation of the LSTM model is shown in Annexure A.

LSTM models have been successfully applied in sequence generation, machine translation, speech, video analysis, language modeling, handwriting recognition, and other fields. LSTM models more realistically represent or imitate human behavior, logical development, and neural organization with cognitive processes.

2.4. Proposed VMD-EEMD-LSTM Model. Ozone has typical nonstationary, nonlinear, and other complex characteristics, and the accuracy of using a single prediction method is limited. Since VMD technology can decompose a complex signal into several regular modal components with lower complexity, the prediction accuracy will be greatly improved when the common prediction methods are used to predict and model each modal component after VMD. However,

previous studies only modeled the estimated modal components after VMD and directly discarded the complex information contained in the residual terms after modal decomposition. Different from the regular residuals in the EEMD technology, the residuals after VMD are highly complex. If this part of the information is directly discarded, the overall prediction accuracy of the model will be reduced. Therefore, in this paper, a decomposition technique for the residual term of VMD is proposed; that is, the residual term is decomposed by the EEMD technique so as to improve the prediction accuracy of the residual term and then improve the prediction accuracy of the model as a whole. Combined with the excellent performance of the LSTM neural network in characterizing time series' autocorrelation and long memory, the detailed modeling steps are as follows.

Step 1: Use VMD technology to decompose the original sequence to obtain each modal component of VMF and subtract the sum of each VMF data from the original time series data to obtain the remaining residual term of VMD.

Step 2: Normalize the decomposed VMF and select training samples and test samples appropriately. LSTM is used to train each VMF, and the prediction result of each VMF component subsequence is obtained.

Step 3: Use EEMD technology to decompose the remaining residual items after VMD twice, use LSTM to separately predict each IMF subsequence after EEMD, and further superimpose the prediction results of the subsequences to obtain the final prediction result of the residual item.

Step 4: Superimpose the prediction results of each VMF component and residual item after VMD to obtain the final prediction result of the original sequence. The complete flow of the implementation is shown in Figure 1.

3. Study Area

This section explains the study area for data collection and implementation results of the proposed method.

3.1. Monitoring Stations. Nanjing is a subprovincial city and the capital of Jiangsu Province. As of 2019, Nanjing has jurisdiction over 11 municipal districts, including Gulou District, Xuanwu District, Jianye District, Qinhuai District, Qixia District, Yuhuatai District, Pukou District, Liuhe District, Jiangning District, Lishui District, and Gaochun District, with a total of 95 streets. There are six towns with a total area of 6,587 square kilometers. According to the results of the seventh national census, the resident population at the end of 2020 was 9,314,685. Nanjing has a subtropical monsoon climate and abundant rainfall, with an annual precipitation of 1,200 mm and four distinct seasons. Nanjing is sunny in the spring, rainy in the rainy season, hot in the summer, dry and cool in the autumn, and cold and dry in the winter [45].

Nanjing has a short spring and autumn, a long winter and summer, and a significant temperature difference between winter and summer. The four seasons have their own characteristics and are suitable for tourism. There are nine air quality monitoring stations in Nanjing; the details of each station including their coordinates and names, are shown in Table 1. Figure 2 shows the locations of all Nanjing monitoring stations, with the covered areas marked with black dots.

3.2. Ozone Data. This paper primarily takes the daily average data sets of nine stations' ozone. Data has been taken from January 2018 until December 2021 for each station in Nanjing. Concentrations of ozone at all stations were normally distributed; the minimum/maximum average values of each station and the mean, median, and standard deviation were used to describe the concentration of air pollutants. Furthermore, to show regional variation in air pollution levels, graphic maps were developed with a geographic information system using ArcGIS (version 10.5). Statistical description of the data is shown in Figure 3 for each station in each year (i.e., from 2018 to 2021). Data from 2018 to 2020 are used for training, and the remaining data are used for testing and validation. Correlation results among stations are shown in Annexure B.

3.3. Validation Methods and Comparative Algorithms. The evaluation indicators of prediction results are selected as RMSE, MAE, and the mean absolute percentage error (MAPE). Three evaluation indicators are used to test the prediction effect of the model. The calculation formula is as follows:

$$e_{\mathrm{MSE}} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}, \tag{10}$$

$$e_{\mathrm{MAE}} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{11}$$

$$e_{\mathrm{MAPE}} = \frac{1/n \sum_{i=1}^{n} |y_i - \hat{y}_i|}{y_i}, \tag{12}$$

$$R^2 = \frac{\sum_{i=1}^{k} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{k} (y_i - \overline{y})^2}. \tag{13}$$

In the formula, y_i and $\hat{y}_i$ are the actual value and predicted value of the station ozone, respectively; n is the test sample size and i is the serial number of the test sample point. R^2 is measured in percentage while MSE, MAE, and MAPE use the same units as measured values.

To verify the advantages of the proposed model, four direct prediction models of LSTM comparisons are used, such as LSTM, gated recurrent units (GRU), BILSTM, and BIGRU, as well as three time series models—ARIMA, SARIMA, and Prophet models—and one prediction model, which is the ablation study of the proposed approach,

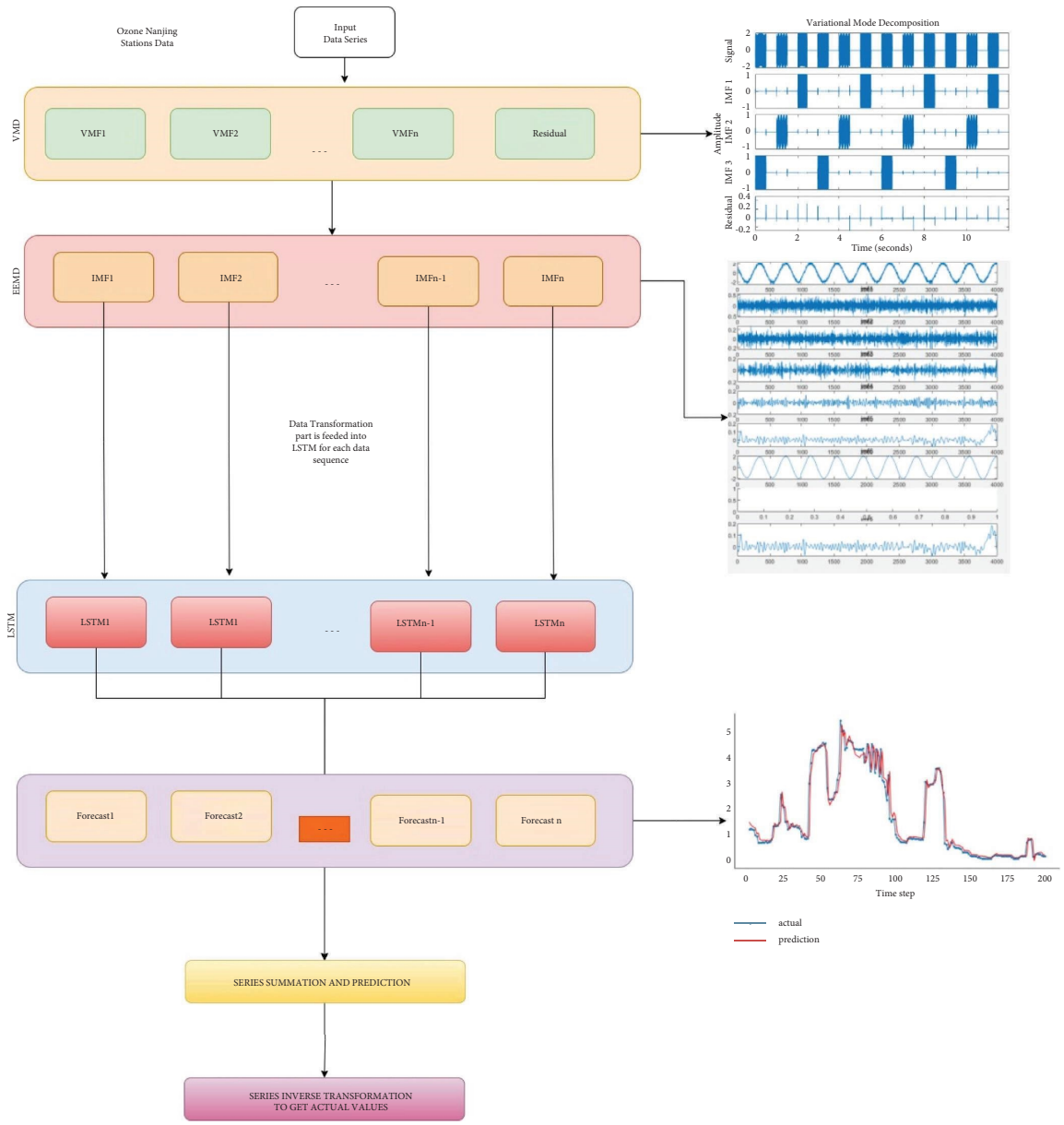

Figure 1: Proposed hybrid forecasting model based on VMD–EEMD–LSTM.

Table 1: Nanjing air quality monitoring station details and geographical locations.

Station codes	Station names	Longitudes	Latitudes
1151A	Maigao bridge	118.8086	32.1065
1152A	Meadow gate	118.7538	32.0551
1153A	Shanxi road	118.7794	32.0745
1154A	Zhonghua gate	118.7752	32.0023
1155A	Ruijin road	118.8109	32.0309
1156A	Xuanwu lake	118.7997	32.0754
1157A	Pukou	118.6414	32.0931
1158A	Olympic sports center	118.7356	31.9996
1159A	Xianlin University town	118.9125	32.1028

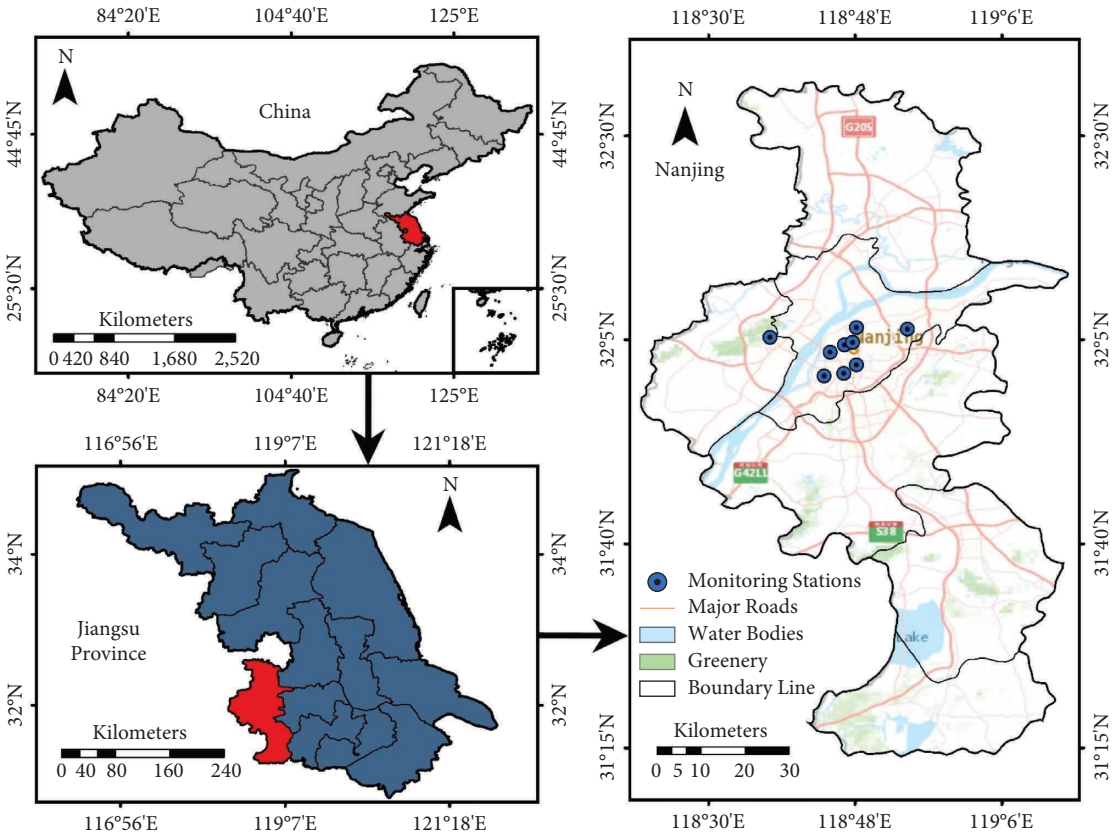

Figure 2: Study area of Nanjing with monitoring stations.

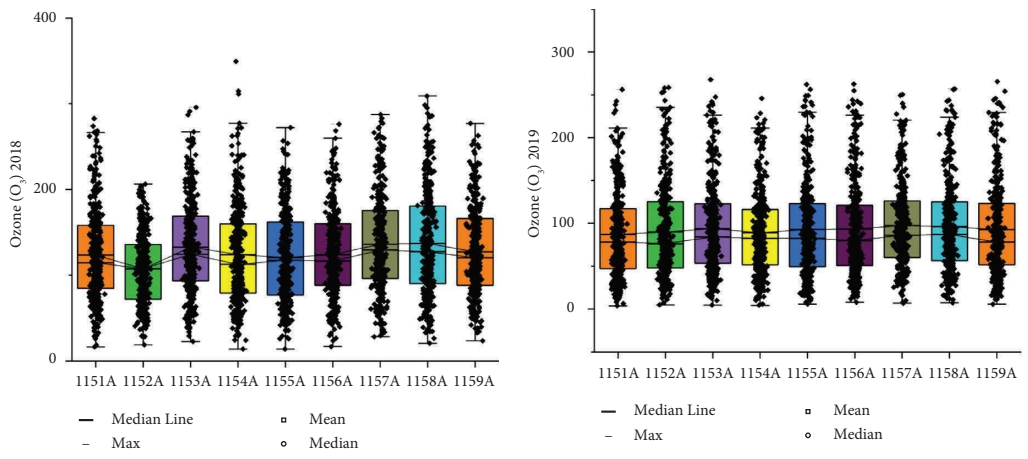

Figure 3: Continued.

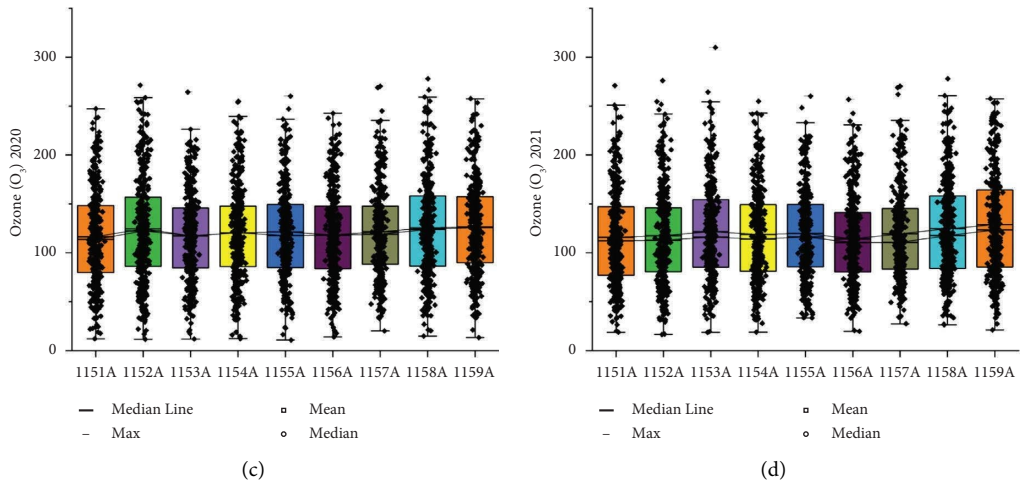

Figure 3: Ozone measurement (in μg m^{-3}) station wise for Nanjing from 2018 to 2021.

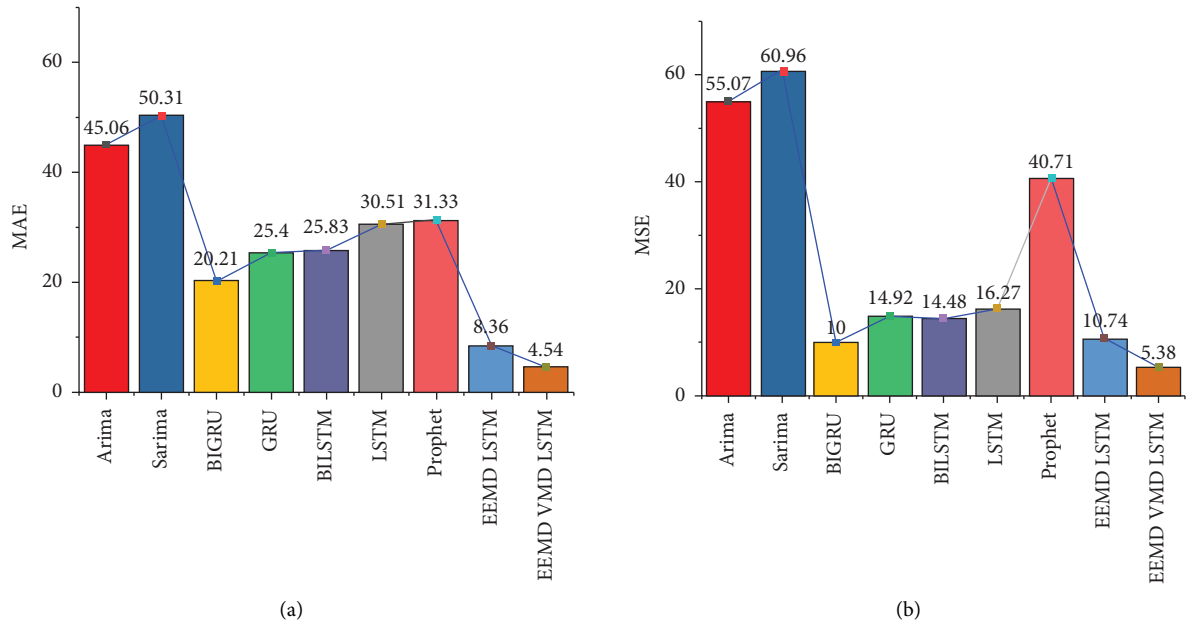

Figure 4: Continued.

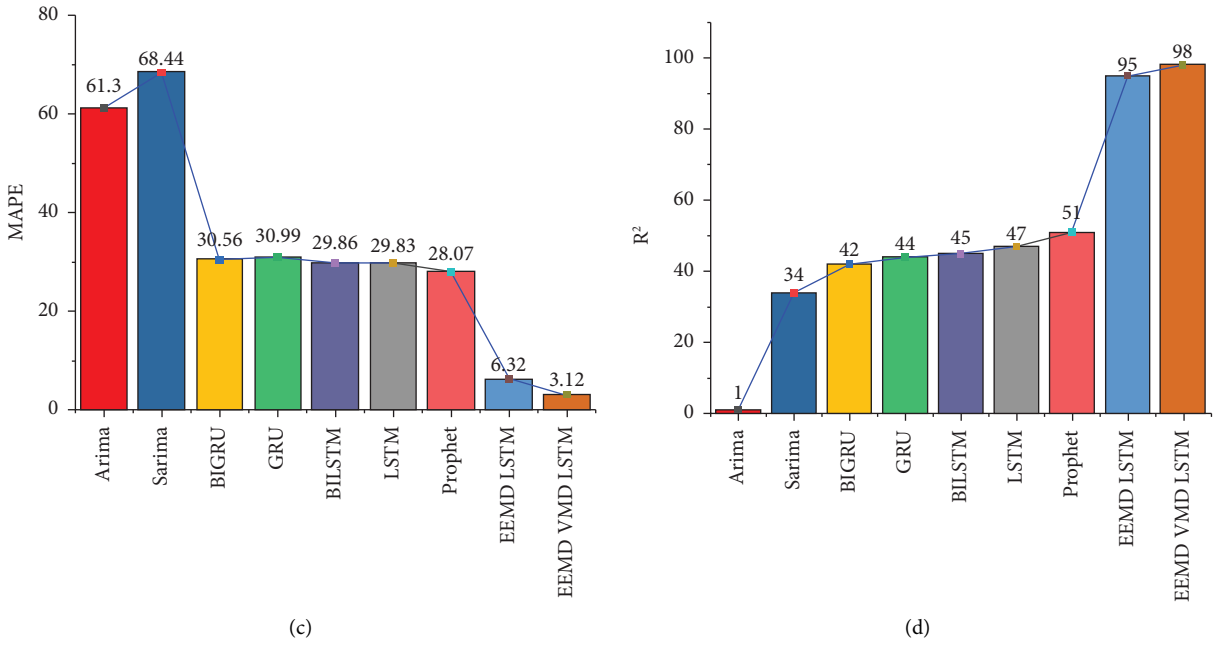

Figure 4: Comparison of different algorithms with a proposed model of Nanjing. (a) MAE. (b) MSE. (c) MAPE. (d) R^2.

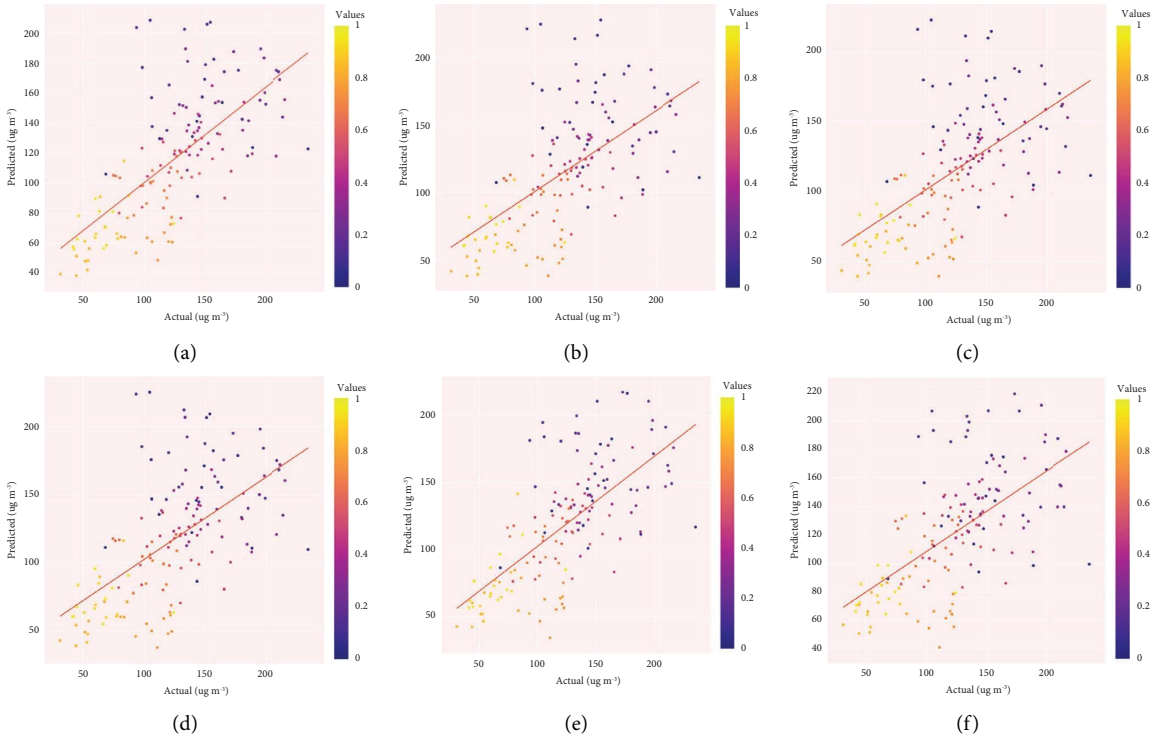

Figure 5: Continued.

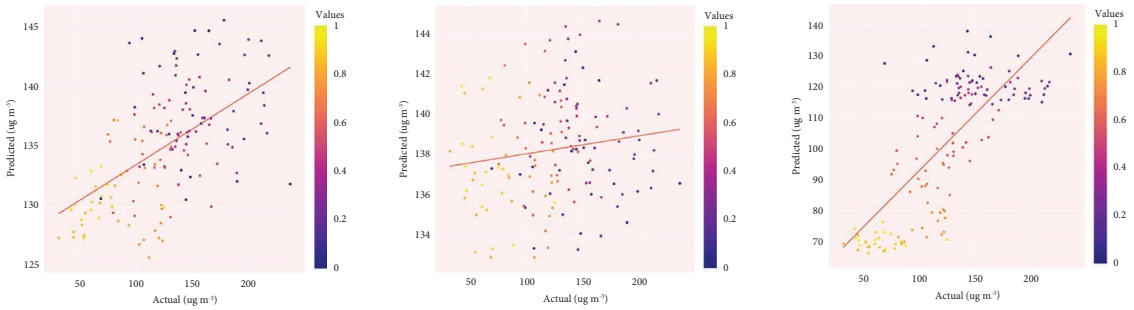

Figure 5: Comparison of actual vs predicted values of all algorithms. (a) ARIMA. (b) SARIMA. (c) Prophet. (d) LSTM. (e) GRU. (f) BILSTM. (g) BIGRU. (h) EEMD–LSTM. (i) VMD–EEMD–LSTM.

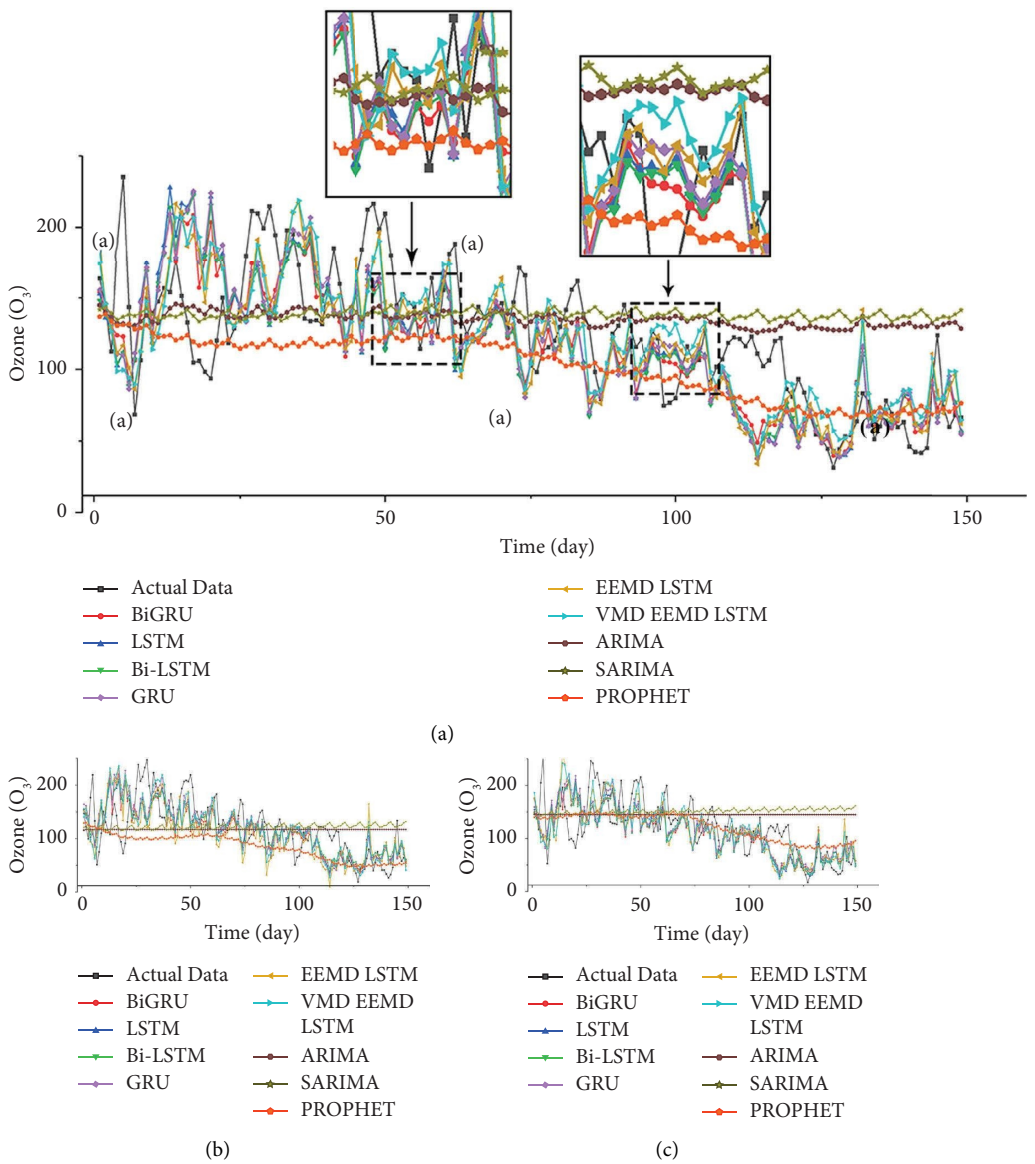

Figure 6: Continued.

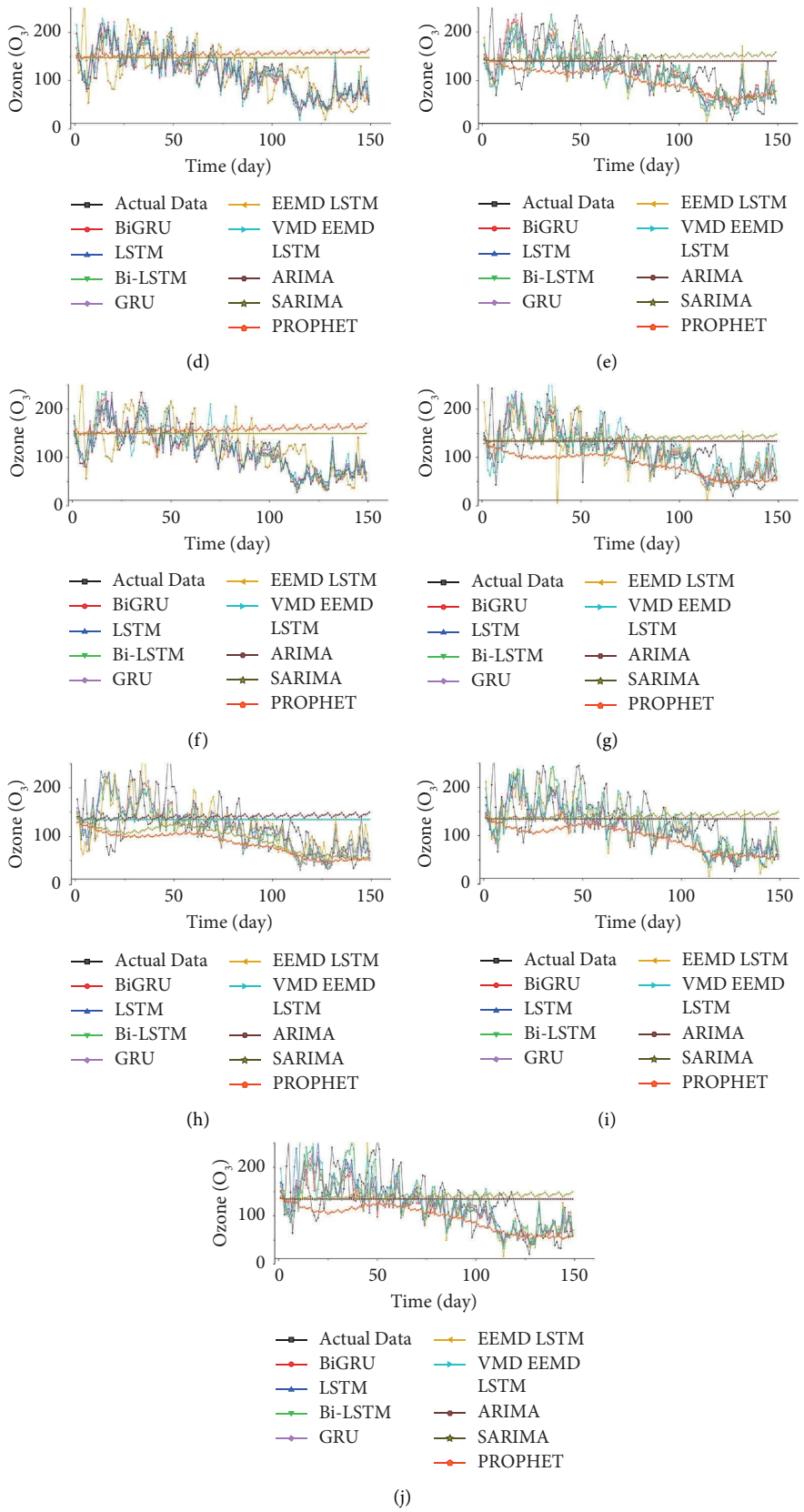

Figure 6: Time series comparison of prediction results from 150 days of observation of all the algorithms. (a) Nanjing. (b) Station 1151A. (c) Station 1152A. (d) Station 1153A. (e) Station 1154A. (f) Station 1155A. (g) Station 1156A. (h) Station 1157A. (i) Station 1158A. (j) Station 1159A.

EEMD–LSTM, developed after excluding VMD. Parameter settings for LSTM models are shown in Annexure A.

4. Results and Discussion

First, the data is decomposed. The results of the EEMD on the ozone time series data are shown in Annexure C. The VMD method is used to decompose the data's original yield sequence in advance, and different VMF components and residual items u are obtained. Then, the residual item u with a series is decomposed by EEMD and combined with the LSTM model for combined prediction analysis. The EEMD–LSTM model is a combined prediction model constructed using EEMD technology as sequence preprocessing to combine with the LSTM method and compare with the proposed method. Next, the prediction effects of different combination models are compared and analyzed.

As shown in Figure 4, the average results of all Nanjing stations are shown and compared with different models. MAE for ARIMA is 45.06, SARIMA is 50.31, BIGRU is 20.21, GRU is 25.4, BILSTM is 25.83, LSTM is 30.51, Prophet is 31.33, EEMD–LSTM is 8.36, and the lowest recorded is for VMD–EEMD–LSTM, 4.54, which shows the accuracy of prediction with low error exists after two series decompositions. The results of each station for all the validation criteria are shown in Annexure D.

Similarly, MSE for ARIMA is 55.07, SARIMA is 60.96, BIGRU is 10, GRU is 14.92, BILSTM is 14.48, LSTM is 16.27, Prophet is 40.71, EEMD–LSTM is 10.74, and the lowest is VMD–EEMD–LSTM, i.e., 5.38. MAPE for ARIMA is 61.3, SARIMA is 68.44, BIGRU is 30.56, GRU is 30.99, BILSTM is 29.86, LSTM is 29.83, Prophet is 28.07, EEMD–LSTM is 3.24, and the lowest is VMD–EEMD–LSTM, 3.1. The result of R^2 for ARIMA is 1%, SARIMA is 34%, BIGRU is 42%, GRU is 44%, BILSTM is 45%, LSTM is 47%, Prophet is 51%, EEMD–LSTM is 95%, and VMD–EEMD–LSTM is 98%. From the comparison of EEMD–LSTM and the proposed method, the change in MAE is a 46% decrease, MSE is decreased by 50%, MAPE is decreased by 4%, and R^2 is increased by 4%.

The value of R^2 is a reliability coefficient between zero and one hundred (or 0 and 1.0). A higher R^2 indicates a more reliable model. Due to the significance of both models, stability and flexibility, optimizing R^2 is not the goal. For the best results when comparing the adjusted R^2 to the original R^2 value, it is ideal for the two numbers to be quite similar. When comparing the R^2 values of all prediction models, it is clear that the VMD–EEMD–LSTM method produced the highest value ($R^2 = 0.98$) (Figure 5).

A visual comparison of the results from 150 days of observation are shown in Figure 6 and highlighted in two different spots where the results of our prediction overlap the actual values. It is also important to observe that, since the data is not linear and is changing constantly, our prediction is approaching. In other aspects, the outcomes showed that LSTM could memorize over long periods of time and had a high degree of accuracy when making predictions. When dealing with complex ozone data, however, a single LSTM model rarely provides optimal results. By breaking complex time series data into time series with different frequencies, EEMD enhanced the prediction accuracy, as seen by an increase in the prediction accuracy across all stations. Similarly, the station comparison experiment revealed that LSTM performed worse than the GRU when incorrect settings were used. To further enhance the model's prediction accuracy, VMD was employed to locate the denoising pattern of the data for LSTM. Specifically, when compared with other models predicting short-term ozone levels, the VMD–EEMD–LSTM model performed better and was useful in other contexts.

Some researchers predict ozone series after one decomposition, and the prediction accuracy is enhanced compared with direct prediction models due to the high complexity of ozone series. EMD, EEMD, and VMD are all decomposition techniques, yet they all suffer from modal aliasing and inefficiency. As a result of this development, VMD is now well-suited for the decomposition of ozone series, a class of problems that had previously plagued the original decomposition method. The ozone series complexity is further reduced by further decomposing the IMF components that have significant complexity after the initial decomposition. The complexity of the IMFs can be efficiently reduced through secondary decomposition; however, it is still unclear how to choose high-complexity IMF components.

After doing a simplex decomposition of the IMF, Wang et al. [46] concluded that the initial component has the most complexity. In this study, we use VMD to quantify the difficulty of each IMF component, and we provide quantitative criteria for selecting complex parts. Modal aliasing and inefficient performance are two issues that VMD can successfully address. However, the decomposition effect will be different if the VMD's decomposition level and penalty factor are not appropriately specified in advance.

It has been shown that the predictive performance of the models given by Wu and Lin [47] which use several series decomposition-integrated frameworks is greatly enhanced. Using pollution data from the city of Anyang as an example, EEMD–LSTM shows improvements of 50.8%, 51.81%, and 52.96% over LSTM in terms of MAE, RMSE, and MAPE. Good prediction performance, early warning accuracy, and prediction stability were also observed using the VMD–SE–LSTM and the EEMD–LSTM across many data sets. These results were similar to our study, which shows that EEMD–LSTM is better than LSTM after series decomposition.

In another approach, noise was removed from air quality data using an EMD model developed by Huang et al. [48] and the resulting data allowed them to extract the IMF components. Each component of the IMF was then modeled using an EMD–IPSO–LSTM air quality prediction model, and values for each were then retrieved. The theoretical and technological support for air pollution prediction and management was supplied by the validation analyses of the algorithm, which revealed that the revised model had higher prediction accuracy and enhanced the model fitting effect compared with LSTM and EMD–LSTM. Compared with this study, a similar approach is proposed in our study by

using EEMD, and our method also produced better results than LSTM and EEMD–LSTM.

The experimental data provided was insufficient because of the experimental settings; however, this strategy was successful in predicting ozone. More work has to be done to refine this study's findings. We did not have information about meteorological factors close to the monitoring stations because of the limits of the air quality monitoring station data. It is conceivable that including information about these aspects in future studies would significantly improve the performance of our model. The spread of air contaminants, for instance, could be influenced by factors such as temperature and wind speed. Better air quality forecasts could be achieved with additional study of climatic conditions, automobile emissions, and interactions between citywide monitoring stations. Furthermore, cubic spline interpolation in EEMD could be swapped out for more modern data-interpolation technology to increase the quality of signal decomposition by minimizing the error introduced by fitting the envelope of each extreme point of the signal.

5. Conclusion

In order to improve the prediction accuracy, various prediction models based on soft computing have been proposed. However, some existing models only emphasize the classifier of the model and pay little attention to data preprocessing. Due to the presence of noise and redundant information in high-dimensional raw data, data preprocessing is a crucial step in predictive models. In this study, a decomposition algorithm is introduced as a preprocessing tool to reduce the dimensionality and extract the intrinsic features of the input raw data. Decomposition algorithms and deep learning latest approaches based on graph as well as transformer based methods [49–51] have many achievements in natural language processing, computer vision, and other fields. In the environment, however, and especially in air quality time series forecasting, there has been little progress recently.

The aim of this paper was to present a new type of prediction model that combines the strengths of EEMD, VMD, and LSTM. The following findings are based on a study of ozone data from nine stations in Nanjing:

(i) The accuracy of ozone prediction was significantly boosted by decomposing the data using VMD with EEMD into many components of different frequencies and then putting these components into the LSTM model.

(ii) In many cases, LSTM's hidden layer neural units were chosen automatically based on past data. LSTM helps to predict more results for short-term and long-term data.

(iii) The VMD–EEMD–LSTM hybrid model described here was found to have the best prediction performance based on experimental comparisons, with a high degree of fitting between the true and predicted values. These results demonstrated the efficacy of the hybrid prediction strategy suggested here for making accurate forecasts in the future. Since this is an approach with real-world implications,

The fluctuations of ozone data are irregular and complex, and the sequence of time series data is affected by multi-dimensional and complex factors. For example, the level of humidity and weather factors impact air pollutants. It is very difficult to predict the future trend of ozone values based on several factors. Therefore, in future research, the model proposed in this paper can be combined with multidimensional complex influencing factors to further improve the overall forecasting effect.

In the future, optimal ensemble models for the decomposed modes can be explored rather than a simple addition approach, and an intelligent forecasting system and smart decision system for ozone monitoring can be developed so that appropriate policies for management can be formulated in light of forecasting results. Future work will be focused on exploring the relationship among pollutants and ozone and using some data fusion approaches. In addition, the suggested method can be applied to other areas of energy forecasting, such as crude oil price forecasting and wind speed forecasting.

Data Availability

All data are available in tabular format.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Hao Tang and Uzair Aslam Bhatti contributed equally as co-first authors.

Acknowledgments

This study was supported by the Princess Nourah Bint Abdulrahman University Researchers Supporting Project number PNURSP2023TR140, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia.

Supplementary Materials

Annexure A LSTM gives description about LSTM implementation. Annexure B Correlation Maps highlights the correlation maps of data variables. Annexure C: EEMD Decomposition Results. Annexure D: Station wise results of prediction. (*Supplementary Materials*)

References

[1] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: a review," *Frontiers in Public Health*, vol. 8, p. 14, 2020.

[2] M. K. Mostafa, G. Gamal, and A. Wafiq, "The impact of COVID 19 on air pollution levels and other environmental

indicators-A case study of Egypt," *Journal of Environmental Management*, vol. 277, Article ID 111496, 2021.

[3] U. A. Bhatti, H. Tang, G. Wu, S. Marjan, and A. Hussain, "Deep learning with graph convolutional networks: an overview and latest applications in computational intelligence," *International Journal of Intelligent Systems*, vol. 2023, Article ID 8342104, 28 pages, 2023.

[4] A. Hasnain, Y. Sheng, M. Z. Hashmi, U. A. Bhatti, Z. Ahmed, and Y. Zha, "Assessing the ambient air quality patterns associated to the COVID-19 outbreak in the Yangtze River Delta: a random forest approach," *Chemosphere*, vol. 314, Article ID 137638, 2023.

[5] U. A. Bhatti, Z. Zeeshan, M. M. Nizamani, S. Bazai, Z. Yu, and L. Yuan, "Assessing the change of ambient air quality patterns in Jiangsu Province of China pre-to post-COVID-19," *Chemosphere*, vol. 288, Article ID 132569, 2022.

[6] L. P. C. Galvan, U. A. Bhatti, C. C. Campo, and R. A. S. Trujillo, "The nexus between CO2 emission, economic growth, trade openness: evidences from middle-income trap countries," *Frontiers in Environmental Science*, vol. 10, 2022.

[7] A. Hasnain, M. Z. Hashmi, U. A. Bhatti et al., "Assessment of air pollution before, during and after the COVID-19 pandemic lockdown in nanjing, China," *Atmosphere*, vol. 12, no. 6, p. 743, 2021.

[8] L. Li, Q. Li, L. Huang et al., "Air quality changes during the COVID-19 lockdown over the Yangtze River Delta Region: an insight into the impact of human activity pattern changes on air pollution variation," *Science of the Total Environment*, vol. 732, Article ID 139282, 2020.

[9] D. L. Hartmann, J. M. Wallace, V. Limpasuvan, D. W. Thompson, and J. R. Holton, "Can ozone depletion and global warming interact to produce rapid climate change?" *Proceedings of the National Academy of Sciences*, vol. 97, no. 4, pp. 1412–1417, 2000.

[10] Y. Bai, Y. Li, X. Wang, J. Xie, and C. Li, "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions," *Atmospheric Pollution Research*, vol. 7, no. 3, pp. 557–566, 2016.

[11] C. Liu, H. Dai, L. Zhang, and C. Feng, "The impacts of economic restructuring and technology upgrade on air quality and human health in Beijing-Tianjin-Hebei region in China," *Frontiers of Environmental Science & Engineering*, vol. 13, no. 5, pp. 70–18, 2019.

[12] J. Wang, W. Xu, J. Dong, and Y. Zhang, "Two-stage deep learning hybrid framework based on multi-factor multi-scale and intelligent optimization for air pollutant prediction and early warning," *Stochastic Environmental Research and Risk Assessment*, vol. 36, no. 10, pp. 3417–3437, 2022.

[13] C. M. Eckhardt, A. A. Baccarelli, and H. Wu, "Environmental exposures and extracellular vesicles: indicators of systemic effects and human disease," *Current Environmental Health Reports*, vol. 9, no. 3, pp. 465–476, 2022.

[14] E. Habeeb, S. Aldosari, S. A. Saghir et al., "Role of environmental toxicants in the development of hypertensive and cardiovascular diseases," *Toxicology Reports*, vol. 9, pp. 521–533, 2022.

[15] Y. Jiang, J. Chen, C. Wu et al., "Temporal cross-correlations between air pollutants and outpatient visits for respiratory and circulatory system diseases in Fuzhou, China," *BMC Public Health*, vol. 20, no. 1, pp. 1131–1213, 2020.

[16] H. Zhao, Y. Zhang, Q. Qi, and H. Zhang, "Evaluating the impacts of ground-level O3 on crops in China," *Current Pollution Reports*, vol. 7, no. 4, pp. 565–578, 2021.

[17] J. H. Chen, D. T. Hu, X. Jia, W. Niu, F. R. Deng, and X. B. Guo, "Monitoring metrics for short-term exposure to ambient ozone and pulmonary function and airway inflammation in healthy young adults," *Beijing da xue xue bao*, vol. 52, no. 3, pp. 492–499, 2020.

[18] J. Zhang, Q. Chen, Q. Wang, Z. Ding, H. Sun, and Y. Xu, "The acute health effects of ozone and PM2. 5 on daily cardiovascular disease mortality: a multi-center time series study in China," *Ecotoxicology and Environmental Safety*, vol. 174, pp. 218–223, 2019.

[19] S. R. Shams, A. Jahani, S. Kalantary, M. Moeinaddini, and N. Khorasani, "The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO2 concentration," *Urban Climate*, vol. 37, Article ID 100837, 2021.

[20] X. Gong, S. Hong, and D. A. Jaffe, "Ozone in China: spatial distribution and leading meteorological factors controlling O3 in 16 Chinese cities," *Aerosol and Air Quality Research*, vol. 18, no. 9, pp. 2287–2300, 2018.

[21] Z. Zhang and N. Ma, "Research on air quality prediction method based on GA-BP model," in *Proceedings of the International Conference on Computer Application and Information Security (ICCAIS 2021)*, pp. 345–359, Wuhan, China, December 2022.

[22] U. A. Bhatti, Y. Yan, M. Zhou et al., "Time series analysis and forecasting of air pollution particulate matter (PM 2.5): an SARIMA and factor analysis approach," *IEEE Access*, vol. 9, pp. 41019–41031, 2021.

[23] A. Masood and K. Ahmad, "A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: fundamentals, application and performance," *Journal of Cleaner Production*, vol. 322, Article ID 129072, 2021.

[24] A. Masood and K. Ahmad, "Prediction of pm2. 5 concentrations using soft computing techniques for the megacity Delhi, India," *Stochastic Environmental Research and Risk Assessment*, vol. 37, pp. 1–14, 2023.

[25] D. R. Liu, S. J. Lee, Y. Huang, and C. J. Chiu, "Air pollution forecasting based on attention-based LSTM neural network and ensemble learning," *Expert Systems*, vol. 37, no. 3, Article ID e12511, 2020.

[26] A. Masih, "Application of ensemble learning techniques to model the atmospheric concentration of SO2," *Global Journal of Environmental Science and Management*, vol. 5, no. 3, pp. 309–318, 2019.

[27] Y. Xiao, H. Yin, Y. Zhang, H. Qi, Y. Zhang, and Z. Liu, "A dual-stage attention-based Conv-LSTM network for spatio-temporal correlation and multivariate time series prediction," *International Journal of Intelligent Systems*, vol. 36, no. 5, pp. 2036–2057, 2021.

[28] K. Shang, Z. Chen, Z. Liu et al., "Haze prediction model using deep recurrent neural network," *Atmosphere*, vol. 12, no. 12, p. 1625, 2021.

[29] A. Masood and K. Ahmad, "Data-driven predictive modeling of PM2. 5 concentrations using machine learning and deep learning techniques: a case study of Delhi, India," *Environmental Monitoring and Assessment*, vol. 195, no. 1, pp. 60–21, 2023.

[30] Y. T. Tsai, Y. R. Zeng, and Y. S. Chang, "Air pollution forecasting using RNN with LSTM," in *Proceedings of the 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress*

(DASC/PiCom/DataCom/CyberSciTech), pp. 1074–1079, IEEE, Athens, Greece, 2018, August.

[31] N. Jin, Y. Zeng, K. Yan, and Z. Ji, "Multivariate air quality forecasting with nested long short term memory neural network," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8514–8522, 2021.

[32] J. Ma, Z. Li, J. C. Cheng, Y. Ding, C. Lin, and Z. Xu, "Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network," *Science of the Total Environment*, vol. 705, Article ID 135771, 2020.

[33] T. Xayasouk, H. Lee, and G. Lee, "Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models," *Sustainability*, vol. 12, no. 6, p. 2570, 2020.

[34] H. Liu, S. Yin, C. Chen, and Z. Duan, "Data multi-scale decomposition strategies for air pollution forecasting: a comprehensive review," *Journal of Cleaner Production*, vol. 277, Article ID 124023, 2020b.

[35] G. Bathla, "Stock Price prediction using LSTM and SVR," in *Proceedings of the 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pp. 211–214, IEEE, Waknaghat, India, 2020, November.

[36] M. Özger, E. E. Başakın, Ö. Ekmekcioğlu, and V. Hacısüleyman, "Comparison of wavelet and empirical mode decomposition hybrid models in drought prediction," *Computers and Electronics in Agriculture*, vol. 179, Article ID 105851, 2020.

[37] X. B. Jin, N. X. Yang, X. Y. Wang, Y. T. Bai, T. L. Su, and J. L. Kong, "Deep hybrid model based on EMD with classification by frequency characteristics for long-term air quality prediction," *Mathematics*, vol. 8, no. 2, p. 214, 2020.

[38] J. Amanollahi and S. Ausati, "PM2. 5 concentration forecasting using ANFIS, EEMD-GRNN, MLP, and MLR models: a case study of Tehran, Iran," *Air Quality, Atmosphere & Health*, vol. 13, no. 2, pp. 161–171, 2020.

[39] J. Du, C. Liu, B. Wu, J. Zhang, Y. Huang, and K. Shi, "Response of air quality to short-duration high-strength human tourism activities at a natural scenic spot: a case study in Zhangjiajie, China," *Environmental Monitoring and Assessment*, vol. 193, no. 11, pp. 697–714, 2021.

[40] H. Pan, J. Zheng, Y. Yang, and J. Cheng, "Nonlinear sparse mode decomposition and its application in planetary gearbox fault diagnosis," *Mechanism and Machine Theory*, vol. 155, Article ID 104082, 2021.

[41] A. Rahimpour, J. Amanollahi, and C. G. Tzanis, "Air quality data series estimation based on machine learning approaches for urban environments," *Air Quality, Atmosphere & Health*, vol. 14, no. 2, pp. 191–201, 2021.

[42] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 01, no. 01, pp. 1–41, 2009.

[43] C. Kaur, A. Bisht, P. Singh, and G. Joshi, "EEG Signal denoising using hybrid approach of Variational Mode Decomposition and wavelets for depression," *Biomedical Signal Processing and Control*, vol. 65, Article ID 102337, 2021.

[44] A. Hasnain, Y. Sheng, M. Z. Hashmi et al., "Time series analysis and forecasting of air pollutants based on Prophet forecasting model in Jiangsu province, China," *Frontiers in Environmental Science*, vol. 10, Article ID 945628, 2022.

[45] Z. Wu and Y. Qian, "An integration method to predict the impact of urban land use change on green space connectivity under different development scenarios using a case study of Nanjing, China," *Environmental Science and Pollution Research*, vol. 29, no. 56, pp. 85243–85256, 2022.

[46] P. Wang, S. Zhu, M. Vrekoussis, G. P. Brasseur, S. Wang, and H. Zhang, "Is atmospheric oxidation capacity better in indicating tropospheric O3 formation?" *Frontiers of Environmental Science & Engineering*, vol. 16, no. 5, pp. 65–67, 2022b.

[47] Q. Wu and H. Lin, "Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network," *Sustainable Cities and Society*, vol. 50, Article ID 101657, 2019.

[48] Y. Huang, J. Yu, X. Dai, Z. Huang, and Y. Li, "Air-quality prediction based on the EMD– IPSO–LSTM combination model," *Sustainability*, vol. 14, no. 9, p. 4889, 2022.

[49] M. A. Al-qaness, A. Dahou, A. A. Ewees et al., "ResInformer: residual transformer-based artificial time-series forecasting model for PM2. 5 concentration in three major Chinese cities," *Mathematics*, vol. 11, no. 2, p. 476, 2023.

[50] U. A. Bhatti, M. Huang, H. Neira-Molina et al., "MFFCG–Multi feature fusion for hyperspectral image classification using graph attention network," *Expert Systems with Applications*, vol. 229, Article ID 120496, 2023b.

[51] Y. Chen, X. Chen, A. Xu, Q. Sun, and X. Peng, "A hybrid CNN-Transformer model for ozone concentration prediction," *Air Quality, Atmosphere & Health*, vol. 15, no. 9, pp. 1533–1546, 2022.