Research Article

Air Pollutant Concentration Forecasting Using Long Short-Term Memory Based on Wavelet Transform and Information Gain: A Case Study of Beijing

Bingchun Liu [ID],[1] **Xiaoling Guo,**[1] **Mingzhao Lai,**[1] **and Qingshan Wang**[2]

[1]*School of Management, Tianjin University of Technology, Tianjin 300384, China*
[2]*School of Humanities, Tianjin Agricultural University, Tianjin 300384, China*

Correspondence should be addressed to Bingchun Liu; tjutlbc@163.com

Air pollutant concentration forecasting is an effective way which protects health of the public by the warning of the harmful air contaminants. In this study, a hybrid prediction model has been established by using information gain, wavelet decomposition transform technique, and LSTM neural network, and applied to the daily concentration prediction of atmospheric pollutants ($PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, and CO) in Beijing. First, the collected raw data are selected by feature selection by information gain, and a set of factors having a strong correlation with the prediction is obtained. Then, the historical time series of the daily air pollutant concentration is decomposed into different frequencies by using a wavelet decomposition transform and recombined into a high-dimensional training data set. Finally, the LSTM prediction model is trained with high-dimensional data sets, and the parameters are adjusted by repeated tests to obtain the optimal prediction model. The data used in this study were derived from six air pollution concentration data in Beijing from 1/1/2014 to 31/12/2016, and the atmospheric pollutant concentration data of Beijing between 1/1/2017 and 31/12/2017 were used to test the predictive ability of the data set test model. The results show that the evaluation index MAPE of the model prediction is 7.45%. Therefore, the hybrid prediction model has a higher value of application for atmospheric pollutant concentration prediction, because this model has higher prediction accuracy and stability for future air pollutant concentration prediction.

1. Introduction

The rapid development of urbanization and industrialization has brought enormous economic results but also caused pressure on resources, energy, and the environment. The air pollution caused by the rapid development of urbanization and industrialization has become an important issue that restricts social and economic development and affects human health [1]. The six major atmospheric pollutants in the air ($PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, and CO) are harmful to human health. Moreover, when the concentration of these pollutants exceeds the standard, it will destroy the human respiratory system and may cause a headache, dyspnea, and heart attack, which will seriously affect the health of human beings and thus restrict social development [2]. Therefore, it is vital to monitor and predict the concentration of pollutants, to avoid the health threats caused by excessive pollutant concentrations effectively. At the same time, the prediction of significant air pollution concentrations can be used as a policy tool for the environmental protection department to regulate social and economic activities such as transportation, industry, and urban construction under extreme air pollution conditions [3]. Therefore, in order to support the decision of environmental management and avoid serious accidents caused by air pollution, it is urgent to establish a precise and stable pollutant concentration prediction model, which can predict the concentration of air pollutants in the future, helping the government to publish control measures for air pollutants and public health protection work.

At present, the research on the prediction of atmospheric pollutant concentration in the world mainly focuses on the application of two methods: the deterministic model and the computational model. The deterministic model research does not require a large amount of historical data, but it needs to have complete knowledge of the source of pollution, the number of timely emissions, and the main chemical reactions and space-time physical transformation processes of the exhaust gases [4]. Computational models usually require a large amount of historical measurement data under various meteorological conditions, and the relationship between historical pollutant data and predicted variables is established by regression and neural network methods [5]. Sánchez et al. proposed a combination of three different methods, including Elman neural network, autoregressive integrated moving average, and a combination of these two methods, applied to predict a certain control near the coal-fired power plant station's SO_2 concentration [6]. The results show that the hybrid method can still obtain excellent prediction results when the concentration of the particles is high. Wu et al. took Beijing as an example, using three-layer FFNN and regression Elman network, and the air quality prediction model was designed to predict the change of PM_{10} concentration of air pollutants on the next day [7]. The optimal model is then selected based on the test performance metrics and learning time. It is proved that the three-layer FFNN of the first-order line cut (OSS) training algorithm is superior to the Elman network of the gradient adaptive learning rate (GDX) training algorithm. As researchers continue to explore neural network algorithms, Das and Padhy constructed a new ANN model to predict PM_{10} concentrations in 23 EU countries [8]. This model reduces the average error of the results to less than 13% during the test. Subsequently, Elangasinghe et al. [9] extracted key information from daily available meteorological parameters and seasonal emission patterns and established a physics-based ANN air pollution prediction tool. The neural network model predicts better results than the linear regression model based on the same input parameters, and it can fully capture the temporal variation of air pollutant concentration in a specific scene. However, these models usually have a common defect; that is, the ability to predict the concentration of particulate matter such as $PM_{2.5}$ and PM_{10} decreases at very high concentrations [5]. This flaw will mislead environmental control decisions and seriously affect human health. Therefore, more adequate experiments and sophisticated modeling techniques are needed to capture sudden changes in particle concentration. Taking this into account, Zhou et al. considered a low data dimension and used a hybrid EMMD-GRNN model based on data preprocessing and analysis to predict $PM_{2.5}$ concentration one day ahead of time [10]. The model can be used to quickly and accurately predict the $PM_{2.5}$ concentration for the next day. Wang et al. proposed an air pollutant prediction model based on a hybrid artificial neural network and hybrid support vector machine [11]. By modifying the error term of traditional methods, artificial neural network and support vector machine effectively improved the prediction accuracy. It has been noted that previous studies have often been used for single contaminant concentration predictions, ignoring the possible nonlinear correlations between different atmospheric contaminants [12]. Lv et al. established an empirical regression model for the prediction of $PM_{2.5}$ and O_3 air pollutant concentrations in three large Chinese cities in Beijing, Nanjing, and Guangzhou in 2016 [13]. The predictive model is an empirical nonlinear regression model designed for automated data retrieval and prediction platforms. The traditional neural network model cannot meet the requirements of high-precision, multioutput air quality prediction, and then the researchers improve the prediction accuracy by improving the input variable structure. Ni et al.'s results show that the selection of historical data such as $PM_{2.5}$, PM_{10}, temperature, wind direction, and wind speed on the previous day to train the model is crucial for the improvement of prediction accuracy [14]. On the other hand, Liu et al. proposed a new collaborative prediction model, using the SVR method of support vector regression to predict the Chinese urban air quality index AQI [15]. The experimental results show that when the air quality characteristic attribute and the air quality index when there are strong interaction and correlation between them, the MAPE value of the multicity multidimensional regression model is reduced.

Therefore, this study will use six major atmospheric pollutants as output variables, and the remaining five as input variables to explore the interactive prediction ability of various pollutants. Relative to the dynamic characteristics of the air pollution index, recurrent neural network (RNN) can effectively solve the adverse effects of the spatial and temporal evolution of the air pollution index. RNN is a deep learning method that can use any memory unit between networks to process any sequence in the input so that it can learn time series [16]. RNN technology has been proposed to solve the problem of time-series prediction, but studies have shown that the typical RNN model cannot solve the long-term dependence of the input sequence. In order to solve this problem, this paper uses a special long short-term memory artificial neural network (LSTM NN) of RNN structure. LSTM can learn time series of long spans and automatically determine the optimal time lag in the prediction. In recent years, LSTM has been successfully applied to image classification, natural language processing, human motion recognition, robot intelligence development, and oil price forecasting [17–19]. Therefore, based on the ability of LSTM to analyze and predict spatiotemporal data, this study applies it to the prediction of air pollution and can obtain good performance.

This research focuses on two aspects: (1) developing an LSTM atmospheric pollutant concentration prediction model based on deep learning; (2) optimizing the input indicators by selecting methods and data dimension processing to improve the prediction accuracy of the LSTM model. Taking Beijing as an example, the prediction of six major atmospheric pollutants ($PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, and CO) was taken as the research object, and the stability and accuracy of the model were analyzed.

2. Methodology

2.1. Long Short-Term Memory. As the current popular recurrent neural network algorithm, the LSTM neural network was first proposed by Hochreiter and Schmidhuber, which improves the memory ability of long (static) and short (cyclic) dynamic features of time series [20]. Similar to the traditional cyclic neural network model, this model models the temporal data by mining the cyclical connections between neurons and mining the intrinsic connections between time-series data. However, unlike the traditional cyclic neural network model, it has a unique neuron structure called a "memory unit." The hidden layer of the LSTM network constructed by the structure can store information of any length of time and obtain a more accurate time-series model [21].

The memory cell structure of the LSTM network is shown in Figure 1. The memory unit module is composed of three "door" structures of input gate, forgetting gate, and output gate and one loop unit. The core idea is to control the switching of each "gate" through a nonlinear function to protect and control the state of the memory unit, thereby controlling the increase and decrease of information [22]. Therefore, the key to the LSTM network is the long-term storage of data information through the state of the memory unit. In general, the three "gates" output value of 0-1 through the sigmoid function to determine how much information can be input to the memory unit.

Assuming that at time t, the input of a memory unit module is x_t, the output is h_t, and the unit state is C_t, and then the formulas of input gate, forgetting gate, output gate, input conversion, unit state update, and output of the hidden layer of the memory unit module are shown in the following equations:

$$i_t = \sigma\left(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i\right), \quad (1)$$

$$f_t = \sigma\left(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f\right), \quad (2)$$

$$o_t = \sigma\left(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o\right), \quad (3)$$

$$C'_t = \tanh\left(W_{cx}x_t + W_{cm}m_{t-1} + b_c\right), \quad (4)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes C'_t, \quad (5)$$

$$y_t = w_{ym}m_t + b_y. \quad (6)$$

In this formula, σ is sigmoid function; tanh is hyperbolic tangent function; i_t, f_t, o_t, and C'_t are input of input gate, forgetting gate, output gate, and input conversion pair unit, respectively; W_{ix}, W_{fx}, W_{ox}, and W_{cx}, and W_{im}, W_{fm}, W_{om}, and W_{cm} are the weight matrix of the input gate, forgetting gate, output gate, and input conversion corresponding to x_t and h_{t-1}, respectively; b_i, b_f, b_o, and b_c are the offset vectors of input gate, forgetting gate, output gate, and input conversion, respectively [22].
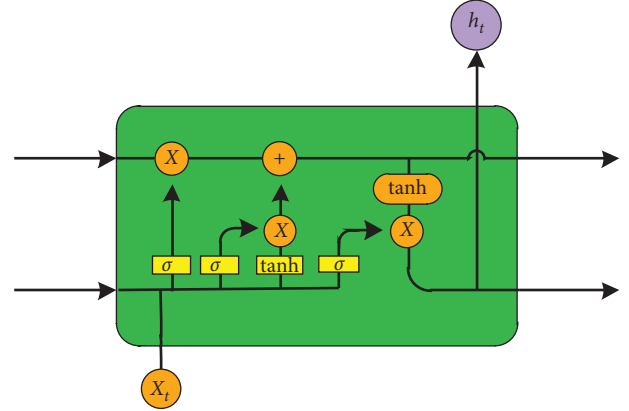

FIGURE 1: Schematic diagram of neurons.

2.2. Wavelet Transform. The wavelet transform is locally adjusted by the window adjustment, and the input signal is decomposed into a low-frequency signal capable of reflecting the true change trend of the signal data and a random disturbance high-frequency signal [23]. The contaminant concentration data are decomposed into a sequence group composed of different components by wavelet transformation. These subsequences have a more stable variance and fewer singular value points than the original data. Therefore, when using the LSTM model to predict the temporal data of atmospheric pollutants, in order to more accurately and accurately express the original signal information and improve the prediction accuracy, the input vector can be structurally transformed to increase the one-dimensional data to a higher level [24]. Air pollutant concentration time-series data $\{y_1, y_2, \ldots, y_n\}$ can be considered as a set of signal sequences. Since wavelet analysis applies to such nonlinear and nonstationary time-series data, this method can be applied to analyze and extract the information characteristics of time-series data of atmospheric pollutant concentration at different frequencies [25]. The wavelet transform adjusts the window through the window to achieve the purpose of localized analysis and decomposes the input signal into a low-frequency signal that can reflect the actual change trend of the signal data and a random-disturbed high-frequency signal. The contaminant concentration data are decomposed into a sequence group composed of different components by wavelet transformation. Compared with the original data, these subsequences have more stable variance and fewer singular value points, which can express the original signal information more effectively and accurately, so the prediction accuracy is better. If the scaling function of the wavelet transform is $\varphi(t)$, the parent wavelet function is $\psi(t)$.

So,

$$\varphi_{j,k}(t) = 2^{j/2}\varphi\left(2^j t - k\right), \quad (7)$$

$$\psi_{j,k}(t) = 2^{j/2}\psi\left(2^j t - k\right), \quad (8)$$

where j and k are scale parameters and translation parameters respectively, and signal $y(t)$ can be expressed by formulas (7) and (8) as follows:

$$y(t) = \sum_k c_{j_o}(k) 2^{j_0/2} \psi\left(2^{j_0} t - k\right) + \sum_k \sum_{j=j_0}^{\infty} d_j(k) 2^{j/2} \psi\left(2^j t - k\right),$$

(9)

where $c_{j_0}(k)$ and $d_j(k)$ are estimated coefficients and detail coefficients, respectively, and the pollutant concentration data can be decomposed into m steps by wavelet transform:

$$y(t) = A_{mt} + D_{1t} + \cdots + D_{mt}.$$

(10)

In formula (10), a is imperfect information representing the original information feature, and b is a piece of high-frequency information indicating a little signal fluctuation, that is, a noise portion of the original information. The low-frequency approximation information and the high-frequency information obtained by the wavelet transform decomposition constitute a new set of input vectors.

2.3. Information Gain.

Usually, feature selection is always selected after quantifying the importance of the feature, and how to quantify the importance of the feature becomes the most significant difference between the various methods [26]. The correlation between features and categories is used in the square test to quantify. The stronger the association, the higher the feature score and the feature is more likely to be retained. In information gain, the measure of importance is to see how much information a feature can bring to a classification system [27]. The more information it brings, the more critical it is.

There is a variable X, which has more than n kinds of values, which are $x_1, x_2, \ldots, x_n$, respectively. The probability of each is $P_1, P_2, \ldots, P_n$, and then the entropy of X is defined as

$$H(X) = -\sum_{i=1}^{n} P_i \cdot \log_2 P_i.$$

(11)

For the classification system, category C is a variable, and its possible values are $C_1, C_2, \ldots, C_n$ and the probability of occurrence of each category is $P(C_1), P(C_2), \ldots, P(C_n)$, So, n is the total number of categories. At this point, the entropy of the system can be expressed as

$$H(C) = -\sum_{i=1}^{n} P(C_i) \cdot \log_2 P(C_i).$$

(12)

In order to distinguish the symbol of t from the symbol of feature t itself, this paper uses T to represent the feature; then,

$$H(C \mid T) = P(t)H(C \mid t) + P(\bar{t})H(C \mid \bar{t}),$$

$$H(C \mid t) = -\sum_{i=1}^{n} P(C_i \mid t) \cdot \log_2 P(C_i \mid t).$$

(13)

The other side can be expanded to

$$H(C \mid \bar{t}) = -\sum_{i=1}^{n} P(C_i \mid \bar{t}) \cdot \log_2 P(C_i \mid \bar{t}).$$

(14)

Therefore, the information gain that the feature T brings to the system can be written as the difference between the original entropy of the system and the conditional entropy after the fixed feature T:

$$\begin{aligned}
IG(T) &= H(C) - H(C \mid T) \\
&= -\sum_{i=1}^{n} P(C_i) \cdot \log_2 P(C_i) + P(t) \sum_{i=1}^{n} P(C_i \mid t) \\
&\quad \cdot \log_2 P(C_i \mid t) \\
&\quad + P(\bar{t}) \sum_{i=1}^{n} P(C_i \mid \bar{t}) \cdot \log_2 P(C_i \mid \bar{t}).
\end{aligned}$$

(15)

2.4. LSTM Forecasting Model.

The flow of the atmospheric pollutant concentration prediction model based on LSTM is shown in the following figure. The input variable consists of three parts, including the eigenvector group obtained by information gain, the high-frequency and low-frequency information vector group, and the historical data group, after wavelet decomposition transformation. In the design of the network structure, after repeated experimental debugging, the complete LSTM model was finally determined, as shown in Figure 2. The entire LSTM neural network contains N LSTM hidden layers, and each layer contains 256 nodes.

The detailed steps for running the model are as follows:

(i) Step 1: form the time-series data set $\{AP_1, AP_2, \ldots, AP_n\}$ from the air pollutant concentration data, use information gain to select the input index characteristics of different prediction targets in the air pollutant concentration prediction, respectively, and obtain the significant factor data set $\{I'_1, I'_2, \ldots, I'_t\}$.

(ii) Step 2: using the wavelet decomposition transform, the significant factor data set $\{I'_1, I'_2, \ldots, I'_t\}$ is decomposed into m layers to obtain a high-dimensional input information set $\{X'_1, X'_2, \ldots, X'_t\}$, where $X'_i = (A_{mi-1}, D_{1i-1}, \ldots, D_{mi-1}), i = 1, \ldots, t$, and $t+1$ decompose the result $X'_{t+1} = (A_{mt}, D_{1t}, \ldots, D_{mt})$ at the moment.

(iii) Step 3: construct a new training data set $\{(X'_i Y'_i)\}_{i=1}^{t}$ based on the high-dimensional input information $\{X'_1, X'_2, \ldots, X'_t\}$ obtained in the second step.

(iv) Step 4: use the new group dataset $\{(X'_i Y'_i)\}_{i=1}^{t}$ to train the LSTM model, and through trial and error, adjust the parameters to get the prediction mode $f(X_i)$.

(v) Step 5: using the prediction model obtained from the above training and the $t+1$ stage input vector X_{t+1} obtained in Step 1, the predicted value $f(X'_{t+1})$ of the $t+1$ stage atmospheric pollutant concentration can be measured.
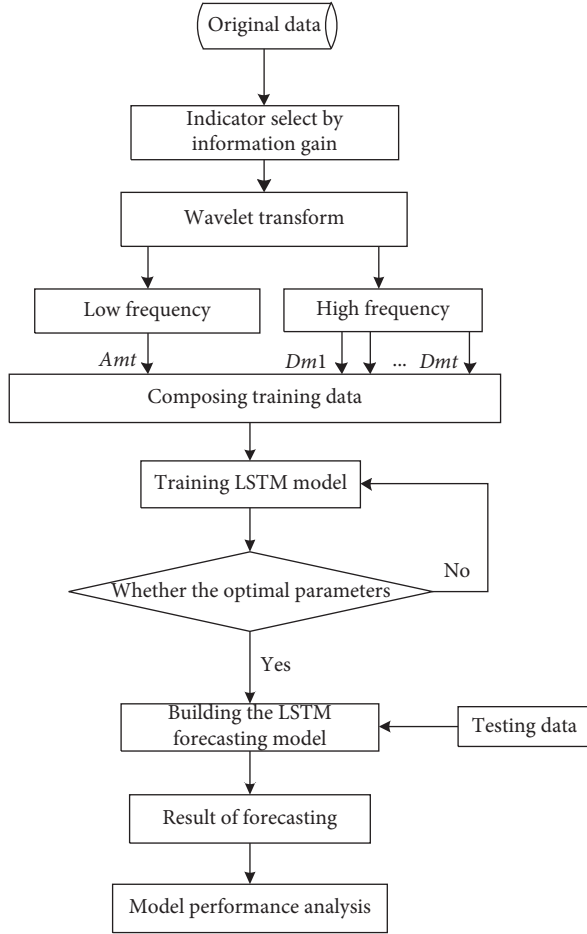
Figure 2: LSTM forecasting model.

Table 1: Statistical descriptions of the main indicators.

	PM$_{2.5}$	PM$_{10}$	NO$_2$	CO	SO$_2$	O$_3$
Unit	μg/m^3	μg/m^3	μg/m^3	μg/m^3	μg/m^3	μg/m^3
Mean	73.63	99.24	12.76	1.17	49.45	98.49
Std.	65.73	73.89	15.84	0.96	23.68	64.68
Min	5	7	2	0.2	10	3
Max	477	550	133	8	155	308
Median	55	84	7	0.9	43.5	84.5
Skewness	2.02	1.79	2.91	2.83	1.19	0.68
Kurtosis	5.49	4.79	11.21	11.52	1.56	−0.35

fluctuations in the raw data of pollutant concentration with seasonal changes, as shown in Figure 3. Due to the Beijing Huilongguan fire accident on June 5, 2015, the CO concentration suddenly increased to 6.8 μg/m^3, and such abnormal fluctuations will seriously affect the predictive ability of the model. Therefore, this study introduces the wavelet decomposition transform; the original data are transformed to obtain low-frequency data and high-frequency data subsequences. These subsequences have more stable variance and fewer singular value points than the original data, which makes the input vector smooth. The concentration of O$_3$ has increased significantly in the same month in the same period of four years and has become the main factor of air pollution in Beijing. Under the influence of environmental management policies such as Beijing's industrial migration, the concentrations of PM$_{10}$ and PM$_{2.5}$ have decreased year by year. However, due to the increasing number of private cars in Beijing, the concentration of nitrogen oxides has increased significantly, which has become a new major air pollution factor.

In order to study the mutual prediction ability between the six major atmospheric pollutants in Beijing, this study will use a total of 1095 data from January 1, 2014, to December 31, 2016, as the training data set for the prediction model. Others include 2017. Three hundred sixty-five data from January 1 to December 31, 2017, were used as test data sets.

4. Results and Discussion

4.1. Result of Information Gain. This study uses nonpredictive targets as input vectors. The purpose is to cope with the specific pollution indicators that can be obtained in the prediction environment with a strong influence of complex, uncertain factors, use the clear indicators to predict each other, and obtain the prediction model with higher applicability. In order to improve the accuracy of model prediction, this study will further select the input variables, select the input variables by information gain, determine the correlation degree with each pollutant concentration, and screen the significant indicators. Using the six atmospheric pollutants (PM$_{10}$, PM$_{2.5}$, NO$_2$, SO$_2$, O$_3$, and CO) in the original data set as targets, respectively, through information gain to complete correlation exploration and feature selection, we can obtain the entropy of each input variable for the prediction target. The value is sorted as shown in Table 2.

The predicted value $f(X_1'), \ldots, f(X_{t+1}')$ will be obtained by repeating Step 1–Step 5.

3. Research Object and Exploratory Data Analysis

The object of this study is Beijing, which is located in the northern part of the North China Plain. It consists of 16 functional areas in 6 districts and 10 suburbs, with a total area of 16,410.51 km^2. Beijing belongs to 39.4°N to 41.6°N and 115.7°E. To 117.4°E.

Collect historical data on the concentration of atmospheric pollutants in Beijing from January 1, 2013, to December 31, 2017. The ground measures six major pollutant concentrations per hour (PM$_{10}$, PM$_{2.5}$, NO$_2$, SO$_2$, O$_3$, and CO). The data collected are from the China National Environmental Monitoring Centre (http://www.cnemc.cn/) and the Ministry of Ecology and Environment of the People's Republic of China (http://www.zhb.gov.cn/). The statistical description of the main indicators is shown in Table 1.

When using the deep learning method to build the model, the selection and normalization of features are essential for the performance of the model. Through exploratory analysis, it is found that there are apparent abnormal

Figure 3: Original data (2014/1/1–2017/12/31).

Table 2: Feature selection results.

	PM_{10}	$PM_{2.5}$	NO_2	SO_2	O_3	CO
PM_{10}	*	3.4629	2.7605	1.3870	3.6033	1.8323
$PM_{2.5}$	1.8323	*	2.7416	2.7416	3.4725	1.8708
NO_2	3.6033	1.8708	*	2.7604	1.9318	1.3870
SO_2	3.4629	3.4725	2.6294	*	1.7792	0.9957
O_3	2.0010	1.9008	1.4521	1.4521	*	1.7791
CO	2.7416	2.6294	1.3870	2.6294	2.7605	*

Table 3: The results of feature selection.

	PM_{10}	$PM_{2.5}$	NO_2	SO_2	O_3	CO
1	NO_2	SO_2	$PM_{2.5}$	$PM_{2.5}$	PM_{10}	$PM_{2.5}$
2	SO_2	PM_{10}	PM_{10}	NO_2	$PM_{2.5}$	PM_{10}
3	CO	CO	SO_2	CO	CO	O_3

Table 3 lists the three items that are most important in predicting pollutants in order. It can be known from the table that PM_{10} and $PM_{2.5}$ have a dangerous influence on the concentration prediction of SO_2, NO_2, O_3, and CO when they are used as input variables, and there are the two pollutants with the most substantial influence on the results during the prediction of atmospheric pollutant concentration. It may be because its source has a specific correlation, which is derived from the burning of fossil fuels such as automobile exhaust. The research model does not consider the complex and uncertain prediction environment. Therefore, in order to improve the accuracy of the prediction model, only the feature vector acquired by the IG is used as an input variable for model training.

4.2. Result of Wavelet Decomposition. Although LSTM is applied to time-series prediction, it can show good prediction performance. However, the LSTM model efficiently represents the high-dimensional nonlinear relationship between the input vectors and the predicted targets through the kernel function. The appropriate high-dimensional input vectors can be used to describe the information features more effectively and accurately and express the meaning of the data. Therefore, model prediction ability depends on no small extent on the choice of input vectors in the model design. In this study, when using the LSTM model to predict the concentration of pollutants, in order to make the prediction results more accurate and stable, the input variables can be structurally transformed to obtain a new set of input variables. The data are upgraded from one-dimensional to high-dimensional data by wavelet decomposition, which more fully and adequately represents the trend of data changes, thereby improving the prediction accuracy. In this study, wavelet decomposition is based on Daubechies (DB) wavelet basis function [4]. Daubechies has low-pass and high-pass filtering properties, which is suitable for feature selection. Because of its inherent orthogonality, Daubechies wavelet can be widely used and shows good performance in time-series data of analysis applications.

Matlab tool was used to form a new prediction data set of six atmospheric pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3, and CO) using the low-frequency approximate information and high-frequency information obtained by wavelet decomposition transformation, respectively, as another new input vector group of the LSTM model. The result of the transformation is shown in Figure 4, which is a high-frequency information group and is a low-frequency information group. The wavelet decomposition set generates high-dimensional input vectors by wavelet decomposition transformation from the density time-series data of three input feature variables, which can effectively increase the data representing information, and the prediction stability of the prediction model is significantly improved.

4.3. Determination of the Best Parameters of LSTM Model. In order to ensure that the hybrid model obtains the best experimental results, the best parameters of the LSTM model
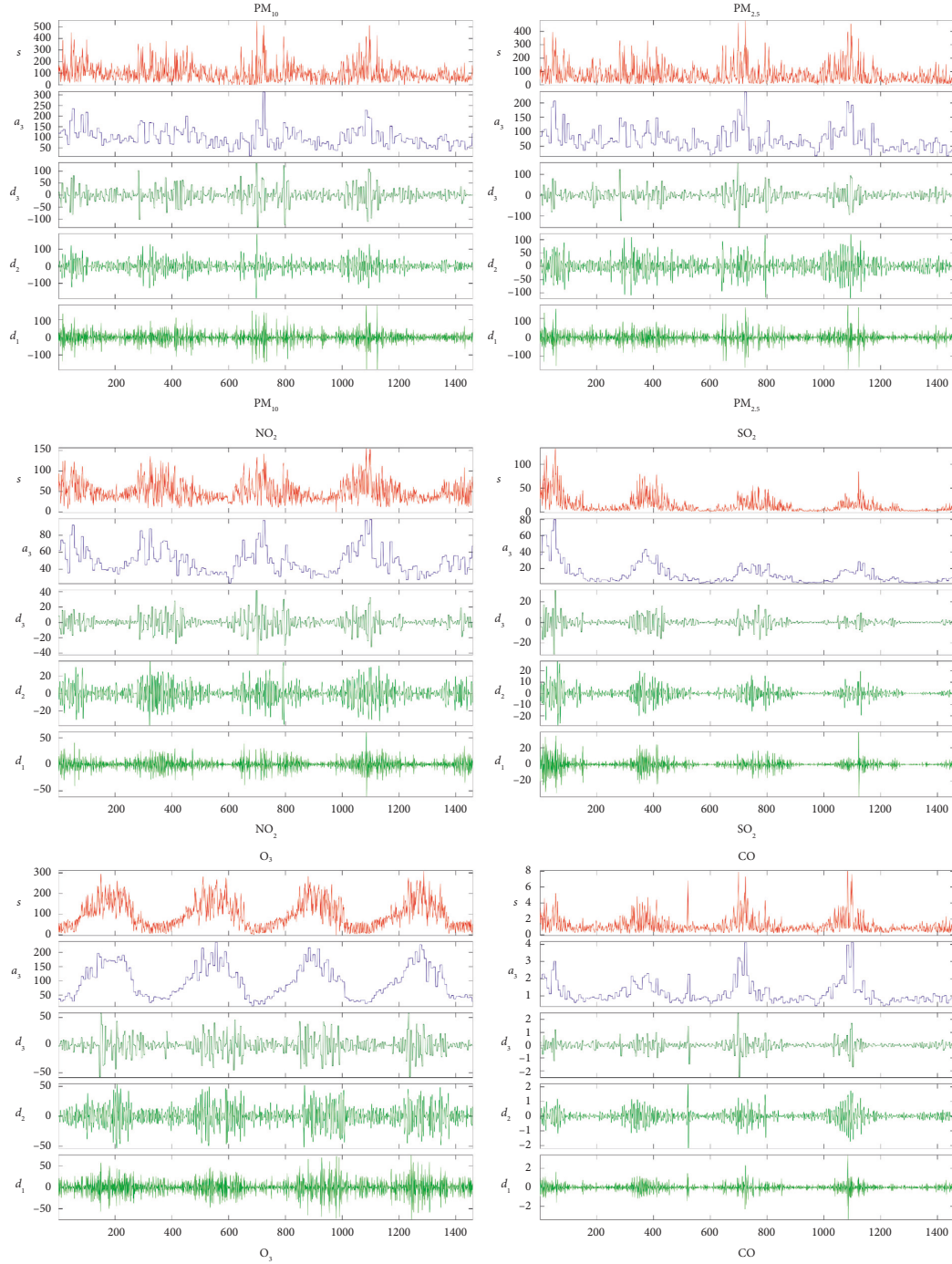
Figure 4: The result of wavelet transform.

should be determined before the experiment starts, so as to reduce the influence of parameter factors on the experimental results. There are three major parameters in the LSTM model, namely, the number of time steps L of each layer in the LSTM, the size of the hidden unit (the same hidden unit is used for each layer in the LSTM), and the batch size during training. Also it includes learning rate (Lr) and max epochs. When selecting the experiment for predicting the best model in the experiment, the number of frames in each sample is set to L. When one parameter is different, the other parameters are fixed, and finally, the best prediction model is found. The model parameters are shown in Table 4.

4.4. *Result of Hybrid Model.* In order to investigate the performance of the LSTM atmospheric pollutant concentration prediction model, this study provided four sets of input variables for the training prediction model, namely, Beijing atmospheric pollutant concentration original data set, characteristic variable set, high-dimensional data set, and high-dimensional characteristic variable set. Verifying

TABLE 4: Parameters for the LSTM network.

Parameter t	Time_steps	Hidden_layers	Batch_size	Lr	Epoch
LSTM	2	64	2	0.001	5000

the high prediction accuracy of the hybrid LSTM model established in this study, six predictions of atmospheric pollution concentration using different methods for feature selection were compared in this study. This prediction model in the experiment was designed by Matlab 2017a and Ubuntu system using Python 2.7 programming. The minimum MAPE was selected as the target for the selection of relevant parameters in the model. Mean absolute percentage error (MAPE) is an important indicator to measure the accuracy of prediction in the statistical field [28].

In this paper, the MAPE index is also used to measure the error of the load prediction algorithm and compare it with other algorithms. MAPE not only considers the error between the predicted value and the real value but also considers the proportion between the error and the real value. The following formula gives the calculation expression of MAPE:

$$e^{\mathrm{MAPE}} = \frac{100\%}{N} \sum_{t=1}^{N} \left| \frac{L_t^a - L_t^f}{L_t^a} \right|, \tag{16}$$

where e^{MAPE} represents the load prediction error measured by the MAPE index; N represents the total number of load prediction time points; L_t^a represents the actual load value; and L_t^f represents the predicted load.

At the beginning of the experiment, the LSTM model with four hidden layers was selected. The original data of six atmospheric pollutants were used as the prediction targets, and the remaining five were used as independent variables. When PM_{10} is used as the dependent variable, MAPE can be as low as 7.54%, and when $PM_{2.5}$ is used as the dependent variable, MAPE reaches 17.25%. The predicted results of the MAPE standard deviation are substantial. The above results show that the prediction stability of the model cannot be guaranteed unless when the independent variables are different using the LSTM model. Therefore, in order to improve the stability and accuracy of the LSTM prediction, the learning efficiency of the auxiliary vector enhancement model is added to the prediction model. Through the feature selection of information gain, only the feature variable set consisting of three input variables with higher correlation is used to train the LSTM model. The average MAPE of the prediction result is reduced from 12.62% to 10.75%, and the prediction accuracy of the model is improved. However, the CO prediction result is improved. The MAPE is 4.35%, and the $PM_{2.5}$ prediction result is 17.85%. The stability of the prediction model is still not guaranteed. In order to achieve the effect of improving the data dimension, the three sets of high-frequency information and a set of low-frequency information obtained by wavelet decomposition are used to form a high-dimensional data set. The subsequence data transformed by the wavelet decomposition have a more stable variance and fewer singular value points than the original data and can express the original signal information more effectively and

accurately. After the LSTM model was fitted using only G2 as training data, the average MAPE was 10.65%, which was 0.05% lower than the G2-assisted prediction alone. However, the variance of the model for different predicted target MAPE is significantly smaller, and the stability can be significantly improved. Therefore, it can be concluded that the time-series data after wavelet decomposition has more stable prediction performance in prediction.

On this basis, this study attempts to set up a new training dataset after the original data of atmospheric pollutants are selected by information gain features and dimensionality enhanced by wavelet transform, aiming at smoothing the training data set of the deep learning model and enhancing the learning ability of the LSTM model. The average MAPE of the prediction result is reduced to 7.45% when the LSTM model is trained with the high-dimensional characteristic data set. Figure 5 shows that when the input vector is processed using wavelet transform and IG, the prediction accuracy is higher and the stability is excellent. The evaluation of the prediction results of each type of air pollution prediction under different models is shown in Table 5 and Figure 6. The actual and predicted concentrations of various atmospheric pollutants in 2017 are compared as shown in Figure 5.

LSTM has achieved excellent results as a high-dimensional nonlinear learning algorithm for the prediction of atmospheric pollutant concentration's time-series data. However, due to the incomplete representation of one-dimensional time-series data, the generalization ability of the prediction model is restricted to some extent. The hybrid prediction model proposed in this study uses wavelet to decompose the time-series data of various pollutant concentrations, constructs new high-dimensional feature vectors to express the relevant information of different pollutants at different frequencies, and better displays the data characteristics.

In order to evaluate the forecasting performance of the hybrid model more comprehensively, the experiment compares the model with some state-of-the-art time-series prediction models, including machine learning methods and deep learning methods. In the experimentation, the SVR, RNN, and GRU models were selected as the comparison models. Furthermore, all the control models were combined with the IG and wavelet model. The training set and the test set were the same as for the hybrid model. A total of three groups of comparative experiments were performed. The final results are shown in Table 6. From the results, It can be seen that the prediction model proposed in this paper improves the accuracy of prediction and is more effective than other prediction models.

4.5. The Stability of the Hybrid Models. In order to measure and evaluate the stability of the hybrid models, the experiment collected data from different regions for verification,
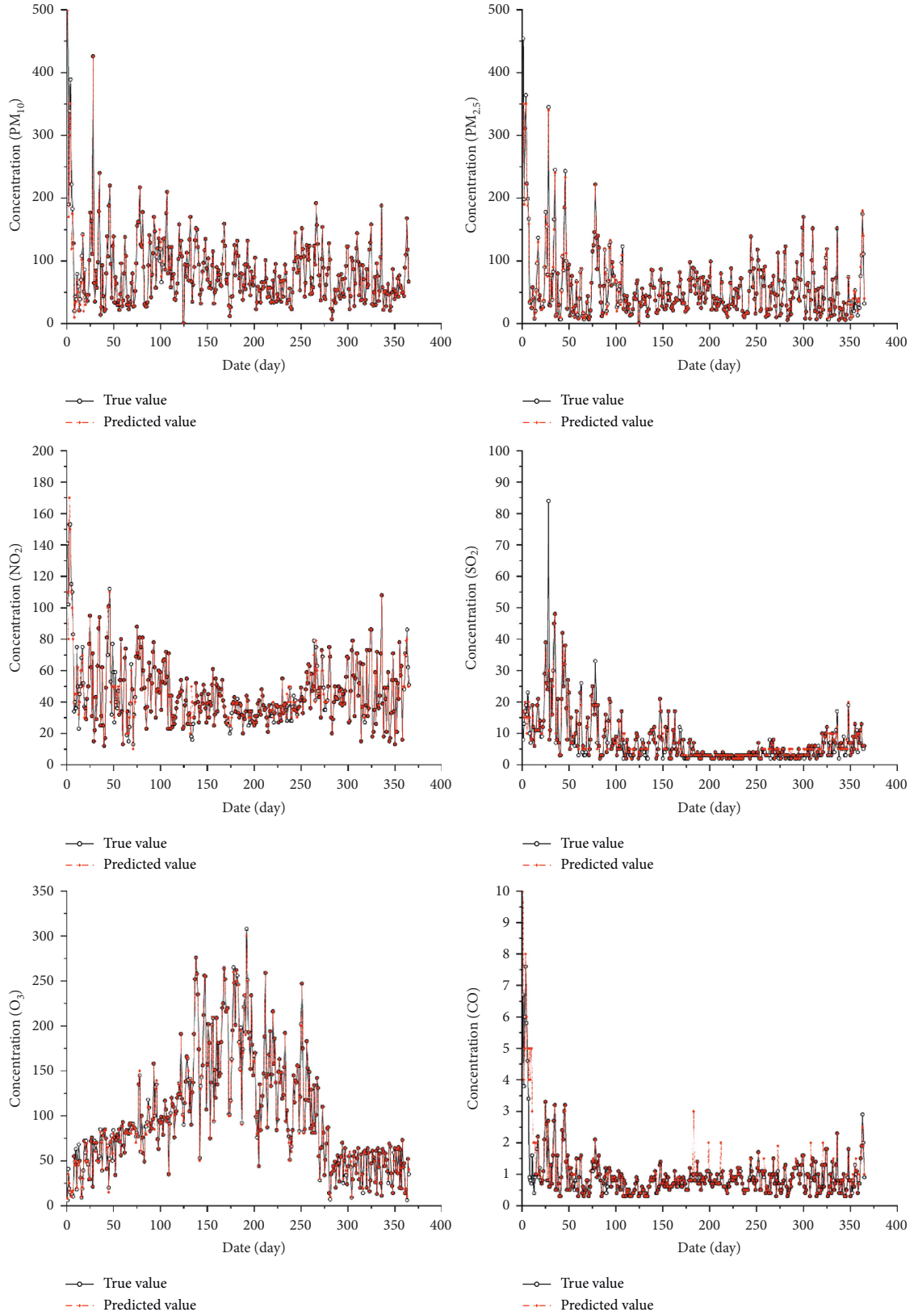
Figure 5: The concentration forecasting results of the hybrid LSTM model.

including Tianjin (TJ), Shanghai (SH), and Shijiazhuang (SJZ). The characteristics of the data collected from different regions are also different. Tianjin is also located in the northern part of the North China Plain, and the level of air pollution is similar to Beijing (BJ). Shanghai is situated in the Yangtze River delta, and the level of air pollution is milder to

TABLE 5: MAPE of models (%).

	PM_{10}	$PM_{2.5}$	NO_2	SO_2	O_3	CO	Average
LSTM	**7.54**	17.25	12.63	15.73	9.05	13.53	12.62
IG-LSTM	8.30	17.85	5.31	11.17	17.51	**4.35**	10.75
Wavelet-LSTM	7.78	12.59	10.73	11.02	11.77	10.37	10.71
Hybrid model	8.84	**7.59**	**4.83**	**9.69**	**7.80**	5.96	**7.45**
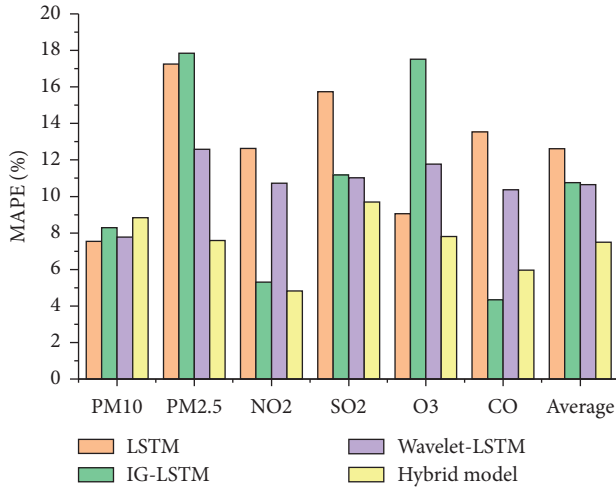
FIGURE 6: MAPE of models.

TABLE 6: Comparison of prediction accuracy of different models (%).

	PM_{10}	$PM_{2.5}$	NO_2	SO_2	O_3	CO
Hybrid model	**8.84**	**7.59**	**4.83**	**9.69**	**7.80**	**5.96**
IG-wavelet-SVR	16.39	17.23	21.29	22.36	20.75	21.55
IG-wavelet-RNN	12.60	12.28	13.51	16.96	12.21	9.48
IG-wavelet-GRU	10.09	9.92	13.50	10.10	9.78	7.75

TABLE 7: Stability measure and evaluation of the hybrid model (%).

	PM_{10}	$PM_{2.5}$	NO_2	SO_2	O_3	CO	Average
Hybrid model (TJ)	10.07	7.09	6.74	4.22	6.85	8.34	7.21
Hybrid model (SH)	3.21	6.73	5.46	6.72	4.98	11.63	6.45
Hybrid model (SJZ)	10.26	9.35	13.21	7.61	8.75	11.68	10.14
Hybrid model (BJ)	**8.84**	**7.59**	**4.83**	**9.69**	**7.80**	**5.96**	**7.45**

Beijing. Shijiazhuang is located in the middle of the North China Plain, and the level of air pollution is more serious than Beijing. The data of three regions are reforecasted to evaluate the stability of the hybrid models. The forecasting procedure and model parameters of three regions is the same as that in Beijing which removes the need to make some duplicate figures. The final prediction results are shown in Table 7.

As shown in Table 7, the MAPE evaluation indicators of the three regions are relatively reasonable, and the average values are close to those of Beijing. Especially, in Tianjin, the average value is 7.21%, which is the closest to Beijing's 7.45%, and the level of air pollution is similar to Beijing. This shows that the hybrid model has high stability and is suitable for air pollution concentration prediction.

5. Conclusions

Severe air pollution has a significant impact on human health, flora and fauna, and the environment. For human health, dangerous air pollution is prone to respiratory diseases and physiological dysfunction, which severely hinders human health. Therefore, scientific and accurate prediction of the concentration of atmospheric pollutants has an important practical significance, which can provide current prediction data and basis for environmental protection agencies, reduce the impact of air pollution on people's health, and guide people's work and life. In this paper, a hybrid LSTM model is established based on wavelet decomposition, information gain (IG), and long- and short-term neural network (LSTM) methods to predict the concentration of six major atmospheric pollutants in Beijing in the future. The study is summarized as follows:

(1) Using the information gain method to select the input variables is helpful to improve the accuracy of the LSTM neural network prediction results of six air pollution concentrations and use wavelet decomposition to convert the characteristic input variables into high-dimensional data. The collection effectively enhances the stability and accuracy of the predictive model.

(2) The hybrid prediction model uses wavelet decomposition of atmospheric pollutant concentration's time-series data, and the low-frequency data and high-frequency data obtained after decomposition are simultaneously used as input variables, increasing the data dimension. The information performance of the pollutant concentration's time-series data at different frequencies is better described.

(3) It can be seen from the experimental results that the hybrid prediction model has a significant improvement in the prediction accuracy of pollutant concentration and increased instability. Especially, in the prediction of burst data points, the hybrid prediction model can predict more accurately. Therefore, the high-dimensional input variable pair's data information composed of low-frequency information and high-frequency information obtained by wavelet transform can be considered as the pollutant concentration's time-series data make this expression more accurate.

(4) This study uses historical data to obtain the calculation results of the characteristic entropy of other pollutant concentrations when different pollutants are used as predictive target variables. According to

the experimental results, it can be observed that the characteristic entropy of the concentration of PM_{10} and $PM_{2.5}$ for the concentration prediction of most major atmospheric pollutants is significant, indicating that when insufficient fossil fuels lead to the increase of SO_2 and NO_2 and other pollutants, the concentration of PM_{10} and $PM_{2.5}$ in the air will be severely affected.

(5) The LSTM neural network method can be used in the model to obtain the prediction results of atmospheric pollutant concentration more accurately. This prediction model is applied to the prediction of six air pollutant concentrations in Beijing. By comparing the actual data, the average MAPE predicted is as low as 7.45%. Compared with the mechanism model with complex and high computational cost, it is more suitable for the prediction environment with robust and complex uncertainty factors. Therefore, the hybrid prediction model has strong applicability and high application value in predicting the concentration of atmospheric pollutants.

(6) By predicting the concentration of air pollutants in three different regions, it can be seen that the hybrid model is more stable and the forecast of air pollution concentration is more reliable. In the control experiment, the MAPE of the other three regions is close to the value of Beijing, indicating that the hybrid model can still obtain better prediction results under the condition of different characteristic data values, and the model has good stable prediction performance.

Data Availability

Raw data used to support the results of this study are included in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] N. Künzli, R. Kaiser, S. Medina et al., "Public-health impact of outdoor and traffic-related air pollution: a European assessment," *Lancet (North American Edition)*, vol. 356, no. 9232, pp. 795–801, 2000.

[2] D. Chen, X. Liu, J. Lang et al., "Estimating the contribution of regional transport to PM 2.5 air pollution in a rural area on the North China Plain," *Science of The Total Environment*, vol. 583, pp. 280–291, 2017.

[3] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, "Real-time air quality forecasting, part I: history, techniques, and current status," *Atmospheric Environment*, vol. 60, pp. 632–655, 2012.

[4] G. Gualtieri, F. Carotenuto, S. Finardi et al., "Forecasting PM_{10} hourly concentrations in northern Italy: insights on models performance and PM_{10} drivers through self-organizing maps," *Atmospheric Pollution Research*, vol. 9, no. 6, pp. 1309–1042, 2018.

[5] W. Sun and J. Sun, "Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm," *Journal of Environmental Management*, vol. 188, pp. 144–152, 2016.

[6] A. B. Sánchez, C. Ordóñez, F. S. Lasheras, F. J. d. C. Juez, and J. Roca-Pardiñas, "Forecasting SO_2 pollution incidents by means of Elman artificial neural networks and ARIMA models," *Abstract and Applied Analysis*, vol. 2013, Article ID 238259, 6 pages, 2013.

[7] Q. Z. Wu, W. S. Xu, A. J. Shi et al., "Air quality forecast of PM10 in Beijing with community multi-scale air quality modeling (CMAQ) system: emission and improvement," *Geoscientific Model Development*, vol. 7, no. 5, pp. 2243–2259, 2014.

[8] S. P. Das and S. Padhy, "A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting," *International Journal of Machine Learning & Cybernetics*, vol. 9, no. 1, pp. 97–111, 2018.

[9] M. A. Elangasinghe, J. A. Salmond, K. N. Dirks, and J. A. Salmond, "Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis," *Atmospheric Pollution Research*, vol. 5, no. 4, pp. 696–708, 2014.

[10] Q. Zhou, H. Jiang, J. Wang, and J. Zhou, "A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network," *Science of The Total Environment*, vol. 496, pp. 264–274, 2014.

[11] P. Wang, Y. Liu, Z. Qin, and G. Zhang, "A novel hybrid forecasting model for PM_{10} and SO_2 daily concentration," *Science of the Total Environment*, vol. 505, no. 505C, pp. 1202–1212, 2015.

[12] P. Li, R. Yan, S. Yu, S. Wang, W. Liu, and H. Bao, "Reinstate regional transport of $PM_{2.5}$ as a major cause of severe haze in Beijing," *Proceedings of the National Academy of Sciences*, vol. 112, no. 21, pp. E2739–E2740, 2015.

[13] B. Lv, W. G. Cobourn, and Y. Bai, "Development of nonlinear empirical models to forecast daily $PM_{2.5}$ and ozone levels in three large Chinese cities," *Atmospheric Environment*, vol. 147, pp. 209–223, 2016.

[14] X. Y. Ni, H. Huang, and W. P. Du, "Relevance analysis and short-term prediction of $PM_{2.5}$ concentrations in Beijing based on multi-source data," *Atmospheric Environment*, vol. 150, pp. 146–161, 2017.

[15] B.-C. Liu, A. Binaykia, P.-C. Chang, M. K. Tiwari, and C.-C. Tsao, "Urban air quality forecasting based on multi-dimensional collaborative support vector regression (SVR): a case study of Beijing-Tianjin-Shijiazhuang," *PLoS One*, vol. 12, no. 7, Article ID e0179763, 2017.

[16] D. Wang, S. Wei, H. Luo, C. Yue, and O. Grunder, "A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine," *Science of The Total Environment*, vol. 580, pp. 719–733, 2017.

[17] E. Habler and A. Shabtai, "Using LSTM encoder-decoder algorithm for detecting anomalous ADS-B messages," *Computers & Security*, vol. 78, pp. 155–173, 2018.

[18] Y. Wu, M. Yuan, S. Dong, L. Lin, and Y. Liu, "Remaining useful life estimation of engineered systems using vanilla

LSTM neural networks," *Neurocomputing*, vol. 275, no. 31, pp. 167–179, 2018.

[19] J. Zhang, Z. Yan, Z. Xiaoping, Y. Ming, and Y. Jinzhong, "Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas," *Journal of Hydrology*, vol. 561, pp. 918–929, 2018.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] H. Luo, M. Huang, and Z. Zhou, "Integration of Multi-Gaussian fitting and LSTM neural networks for health monitoring of an automotive suspension component," *Journal of Sound and Vibration*, vol. 428, pp. 87–103, 2018.

[22] H. Y. Kim and C. H. Won, "Forecasting the volatility of stock price index: a hybrid model integrating LSTM with multiple GARCH-type models," *Expert Systems with Applications*, vol. 103, no. 1, pp. 25–37, 2018.

[23] M. J. Alizadeh and M. R. Kavianpour, "Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean," *Marine Pollution Bulletin*, vol. 98, no. 1-2, pp. 171–178, 2015.

[24] Y. Chen, S. Runhe, S. Shijie, and G. Wei, "Ensemble and enhanced PM_{10} concentration forecast model based on stepwise regression and wavelet analysis," *Atmospheric Environment*, vol. 74, pp. 346–359, 2013.

[25] V. Nourani, A. H. Baghanam, J. Adamowski, and K. Ozgur, "Applications of hybrid wavelet–Artificial Intelligence models in hydrology: a review," *Journal of Hydrology*, vol. 514, pp. 358–377, 2014.

[26] X. Feng, Q. Li, Z. Yajie et al., "Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation," *Atmospheric Environment*, vol. 107, pp. 118–128, 2015.

[27] A. Vlachogianni, P. Kassomenos, A. Karppinen, S. Karakitsios, and J. Kukkonen, "Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM_{10} in Athens and Helsinki," *Science of the Total Environment*, vol. 409, no. 8, pp. 1559–1571, 2011.

[28] Y. Bai, L. Yong, W. Xiaoxue, X. Jingjing, and L. Chuan, "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.