Contents lists available at ScienceDirect

Science of the Total Environment

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

PM2.5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition

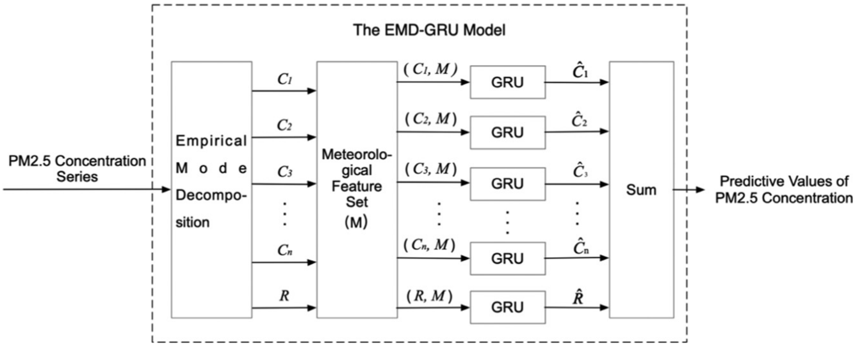
Guoyan Huang [a], Xinyi Li [a], Bing Zhang [a,b,*], Jiadong Ren [a,b]

[a] School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
[b] Key Laboratory of Software Engineering in Hebei Province, Qinhuangdao 066004, China

HIGHLIGHTS

- Propose a deep learning method based on data decomposition to make effective predictions of PM2.5 concentration.
- Check the stationarity of air pollutant concentration series by calculating the ACF and the ADF test.
- Use EMD to decompose the PM2.5 concentration series.
- Construct a multi-step prediction GRU neural network.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 17 August 2020
Received in revised form 1 December 2020
Accepted 11 December 2020
Available online 8 January 2021

Editor: Dimitra A Lambropoulou

Keywords:
Time series
Deep learning
Empirical mode decomposition
Gated recurrent unit neural network
PM2.5 concentration prediction

ABSTRACT

The main component of haze is the particulate matter (PM) 2.5. How to explore the laws of PM2.5 concentration changes is the main content of air quality prediction. Combining the characteristics of temporality and nonlinearity in PM2.5 concentration series, more and more deep learning methods are currently applied to PM2.5 predictions, but most of them ignore the non-stationarity of time series, which leads to a lower accuracy of model prediction. To address this issue, an integration method of gated recurrent unit neural network based on empirical mode decomposition (EMD-GRU) for predicting PM2.5 concentration was proposed in this paper. This method uses empirical mode decomposition (EMD) to decompose the PM2.5 concentration sequence first and then fed the multiple stationary sub-sequences obtained after the decomposition and the meteorological features into the constructed GRU neural network successively for training and predicting. Finally, the subsequences of the prediction output are added to obtain the prediction results of PM2.5 concentration. The forecast result of the case in this paper show that the EMD-GRU model reduces the RMSE by 44%, MAE by 40.82%, and SMAPE by 11.63% compared to the single GRU model.

1. Introduction

In recent years, severe global climates have occurred frequently. The serious environmental problem of air pollution has drawn worldwide attention. Fine particulate matter 2.5(PM2.5) in the atmosphere (particulate matter less than 2.5 μm in diameter in aerodynamics) is the main component that affects air quality. PM2.5 in cities mainly comes from exhaust emissions of urban traffic. Similarly, the exhaust and pollutants generated by the factory during various industrial activities are also sources of PM2.5 (Calvo et al., 2013). The increase of PM2.5 concentration level will directly lead to poor air quality and reduce visibility. Long-term exposure to high PM2.5 concentration will cause harm to human health, such as leading to respiratory disease and cardiovascular disease (Gao et al., 2015) (Pun et al., 2017) and even causing

[*] Corresponding author at: School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China.
E-mail address: bingzhang@ysu.edu.cn (B. Zhang).

https://doi.org/10.1016/j.scitotenv.2020.144516
0048-9697/© 2021 Elsevier B.V. All rights reserved.

death in severe circumstances (Burnett et al., 2018). At the same time, recent studies have shown that the harsh smog climate environment will reduce the public's subjective well-being (Zheng et al., 2019). Due to the harm caused by ultra-standard PM2.5 concentration, the problem of PM2.5 concentration prediction has received more and more attention, but the accuracy of the current prediction method is still not ideal. Therefore, timely, effective and accurate prediction of PM2.5 concentration is helpful for the formulation and implementation of early warning decision-making activities. This can help the public to reasonably arrange their time and means of travel, reducing the impact of haze climate on their actual life.

For a long time, most of current PM2.5 concentration prediction methods are based on the principle of time series analysis. For example, the ARMA model (Box and Jenkins, 2010) was commonly used to predict air pollutants. However, as a linear model, when dealing with non-linear features, the ARMA model is impossible to accurately capture the law of non-linear changes, which makes the model's prediction error high. Neural networks can fully play an important role in handling complex non-linear relationships. Foued et al. (Saâdaoui and Ben Messaoud, 2020) (Saâdaoui et al., 2020) proposed a new multiscaled Feedforward Neural Network (FNN) for nonlinear time series forecasting. Perez et al. (Pérez et al., 2000) used a multilayer neural network to predict PM2.5 concentration in downtown San Diego and achieved good results. In term of the time series prediction method based on time series decomposition. Foued and Hana (Saâdaoui and Rabbouch, 2019) used a wavelet-based hybrid neural network for short-term electricity prices forecasting. At the same time, they used the wavelet-based causal statistical model for online virtual sensors that efficient estimator the urban traffic flow (Rabbouch et al., 2018). With the development of deep learning, the application of deep learning models for air quality prediction has set off a boom. Wen et al. constructed a convolutional long short-term neural network to predict the PM2.5 concentration in Beijing (Wen et al., 2019). The accuracy of the prediction results is significantly higher than that of traditional neural networks.

The prediction of air pollutants in the time dimension can be regarded as a problem of multivariate time series prediction, but previous studies of the prediction of air pollutants have not taken the non-stationarity of air pollutants as time series, and researchers have often overlooked the impact of non-stationarity in time series prediction. This made the prediction accuracy of the model lower due to the limited prediction performance. Aiming at the non-stationarity and long-term dependence of air pollutant sequences, this paper proposed a gated unit recurrent (GRU) neural network based on the empirical mode decomposition (EMD) method for the short-term prediction of PM2.5 concentration in Beijing. Our proposed prediction model based on deep learning algorithms which have the characteristics of training and testing can be trained according to historical marked data features. As an input, the unknown characteristic data (such as tomorrow's or next hour's meteorological data) can be tested to output the PM2.5 value by the trained model effectively.

The main contributions of this paper are as follows: (1) It checked the stationarity of air pollutant concentration series by calculating the autocorrelation function (ACF) and Augmented Dickey-Fuller(ADF) test. The partial autocorrelation function (PACF) was calculated to determine the number of time step in the GRU neural network. (2) The proposed method used the concept of “divide and rule”. It first used EMD to decompose the PM2.5 concentration series, and then the multiple stationary subsequences obtained after the decomposition and the meteorological features were fed into the constructed GRU neural network for training. Finally, the subsequences of the prediction output from GRU were summed to obtain the prediction results. In this way, the defects of slow convergence and hysteresis in the neural network were further solved, for which the proposed method improved the prediction model's Goodness of Fit and robustness. (3) The air quality dataset of Beijing area from 2010 to 2014 was used to validate the effectiveness of the EMD-GRU model proposed in this paper. The prediction model effectively predicts the hourly PM2.5 concentration value of the next hour based on the meteorological data and PM2.5 concentration data of the past four hours. The experimental results show that the present model has reduced the RMSE obtained by using GRU neural networks alone by 44%. The model mainly succeeds in predicting hourly PM2.5 concentration in small area efficiently. The characteristic of our model is more fine-grained in the short-term PM2.5 forecast.

The remaining part of this article is organized as follows. Section 2 is a detailed description of the construction of the EMD-GRU model for PM2.5 concentration prediction. Section 3 introduces the experimental setup and results. Section 4 summarizes current PM2.5 prediction methods in China and abroad. Section 5 discusses the advantages and disadvantages of the method proposed in this paper. Finally, the conclusion of this article is given in Section 6.

2. Method

In this paper, a hybrid EMD-GRU model was proposed on the basis of data decomposition and GRU neural network for short-term prediction of PM2.5 concentration. The EMD-GRU model is divided into three parts. Relevant data preprocessing on the data sample is performed in the first part. The non-stationarity input sequence is decomposed by EMD method in the second part. In the third part, the GRU neural network is established to train and learn the decomposed subsequence and meteorological features, and then we make predictions on the test set and integrate forecast results from GRU output layer. The process of predicting the PM2.5 time series by using the EMD-GRU model is shown in Fig. 1. The specific operation process of the EMD-GRU model to predict PM2.5 concentration is listed as follows:

(1) In the data preprocessing stage, linear interpolation is used to fill in missing values in the data sample. Simultaneously, the non-numeric features are encoded. The linear interpolation method is used to construct linear functions like Eq. (1) from the known sample points at the left and right of the missing values. The missing value at time t is replaced by the value of $F(x_t)$ found at its corresponding point x_t.

$$F(x) = ax + b \tag{1}$$

(2) We calculate the values of Pearson correlation coefficient between each meteorological feature and PM2.5 concentration, analyze the correlation between PM2.5 concentration and meteorological features(M), and select the appropriate features as input data for prediction.

(3) We detect the stationarity of PM2.5 concentration series by calculating two different statistical functions: autocorrelation function (ACF) and augmented Dickey-Fuller(ADF). When we identify whether this series is a non-stationarity series, the EMD method is used to decompose the PM2.5 concentration series into several IMFs(C_i) and one residue(R).

(4) GRU neural network input preparation: we calculate the partial autocorrelation coefficient (PACF) of the PM2.5 concentration series to determine the time step of the series, that is, to bring the historical data into the GRU unit, and we normalize all input data to make it meet the GRU model API requirements of Keras software package.

(5) Construct the separate multi-step prediction double-layer GRU neural network, then adjust the parameters of GRU through the performance of the training set on the model, and finally, choose the best fitting model.

(6) In the end, reconstruct the predicted value output by the GRU model, that is, to use Eq. (2) to add up all these predicted sequences to obtain the PM2.5 concentration predicted sequence. Calculate prediction errors and evaluate model effects.

$$\widehat{x} = \sum_{i=1}^{n} \widehat{C}_i(t) + \widehat{R}(t) \tag{2}$$

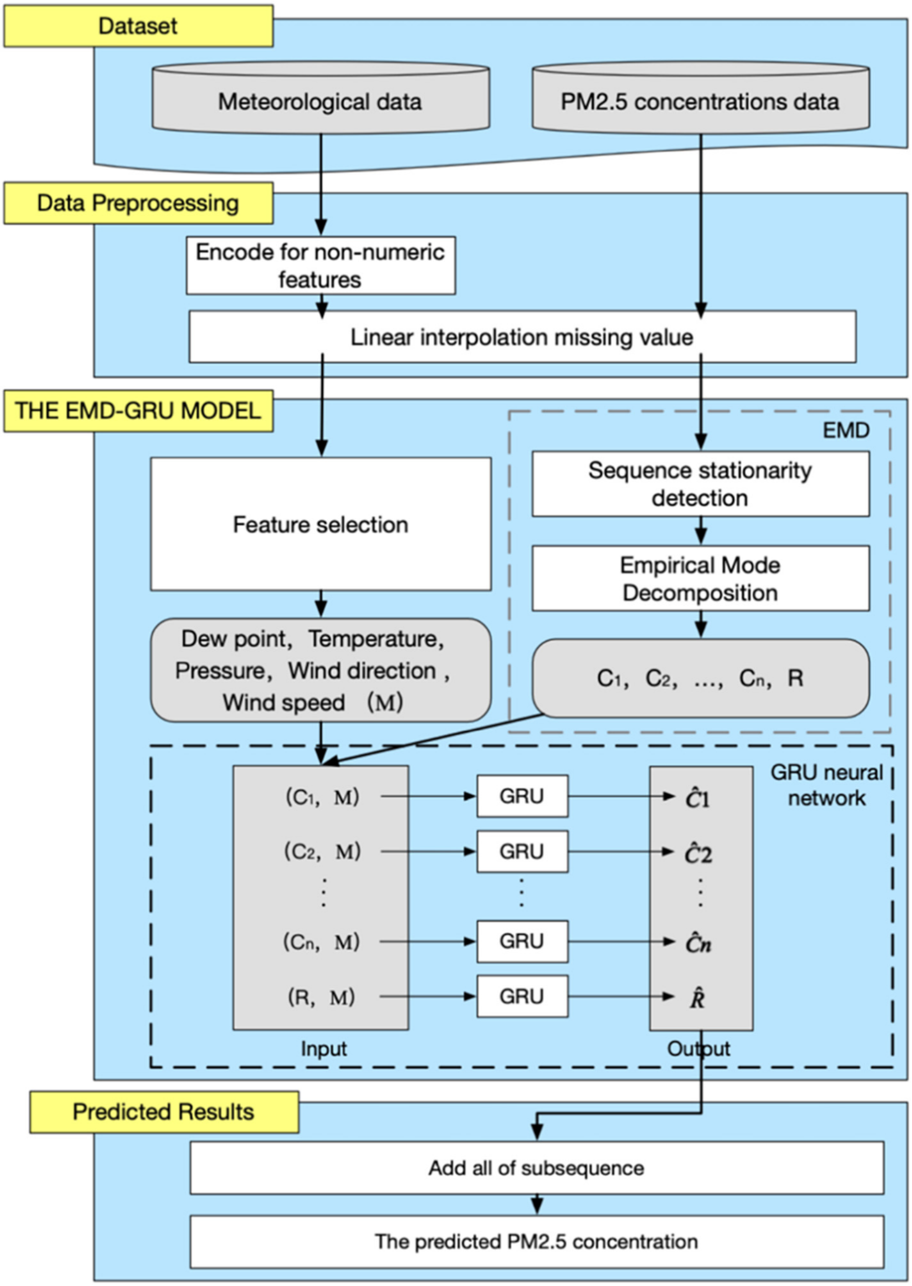

Dataset
Meteorological data
PM2.5 concentrations data
Data Preprocessing
Encode for non-numeric features
Linear interpolation missing value
THE EMD-GRU MODEL
EMD
Feature selection
Sequence stationarity detection
Empirical Mode Decomposition
Dew point，Temperature，Pressure，Wind direction，Wind speed （M）
C1，C2，…，Cn，R
GRU neural network
(C1，M)
(C2，M)
(Cn，M)
(R，M)
GRU
GRU
GRU
GRU
Ĉ1
Ĉ2
Ĉn
R̂
Input
Output
Predicted Results
Add all of subsequence
The predicted PM2.5 concentration

Fig. 1. Framework of the EMD-GRU model.

2.1. Feature selection

Meteorological conditions are important factors affecting the spread of air pollutants (He et al., 2013). To this end, the analysis of the correlation between meteorological characteristics and PM2.5 to carry out effective feature selection is crucial for the PM2.5 prediction that follows. In this paper, the Pearson correlation coefficient (Pearson, 1895) is used to express the relationship between PM2.5 concentration and meteorological features. The Pearson correlation coefficient's formula is shown as Eq. (3), where x and y denote the PM2.5 concentration series and the meteorological features; n is the number of samples in the series.

$$\rho_{x,y} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i-1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \tag{3}$$

Fig. 2 is a heat map of the correlation coefficient between PM2.5 concentration and meteorological features. As shown in Fig. 2, the absolute value of the Pearson correlation coefficient between wind direction, wind speed, humidity and PM2.5 concentration value rank in the top three among all the features. It means that the three meteorological conditions of wind direction, wind speed and humidity have the greatest influence on the PM2.5 concentration values. Besides, the temperature and air pressure, which have negative correlation with each other, also play an important role in affecting the change of PM2.5 concentration value. Under the high air pressure and low temperature weather conditions, air pollution will be more serious. Zhao et al. also verified this theory in their study by analyzing the relationship between the changes in PM2.5 pollution and meteorology in five major cities in China (Zhao et al., 2018). In the data of this paper, the value of the correlation coefficient between the accumulated snowfall time and the PM2.5 concentration value is less than 0.02. As this value is the closest to zero, the accumulated snowfall time has the weakest correlation

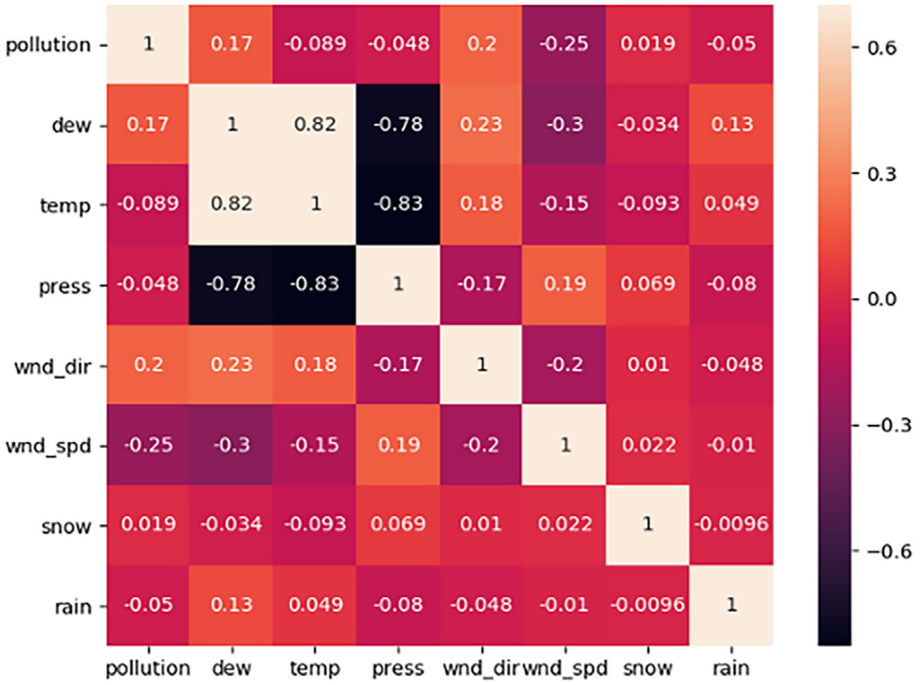


Fig. 2. Heat map of the correlation coefficient between all of features.

with PM2.5. In the Beijing PM2.5 dataset, most of the accumulated rainfall time and the accumulated snowfall time are zero. It can clearly be seen that there are fewer rainy and snowy weathers in Beijing. The two features have little reference value in the PM2.5 concentration prediction.

We selected different combinations of features to bring into our model for experiment. Table 1 shows the results of PM2.5 prediction errors under the different feature selections. As shown in Table 1, the model performs best when using the data after removing the two features of rainfall time and snowfall time as the model's input. Accordingly, dew point(f_d), historical PM2.5(f_c), temperature(f_t), air pressure (f_p), wind direction(f_w) and wind speed (f_s) are selected as the input variables of the model.

In terms of linearity test of features, this paper used Brock-Dechert-Scheinkman(BDS) statistics (Brock et al., 1992) to verify the non-linearity of PM2.5 time series $\{x_t\}$ by the software package of EViews. We set the embedding dimension 'm' of BDS test is 5. The results of the BDS statistic are represented in Table 2. From the results of BDS test, all the Z-statistics are much higher than the critical value under the 95% confidence interval and the *P*-values are all lower than 0.05. This indicates all BDS statistics reject the IID hypothesis. It is concluded that the PM2.5 concentration series is a nonlinear time series. Combining with the non-linearity of PM2.5 concentration characteristics, this paper selects the deep neural network with strong ability to deal with nonlinear features as the main body of the prediction model.

2.2. Time series stationarity detection and data decomposition

When analyzing the data as time series, it is necessary to consider whether the stochastic process reflected by the time series is steady. Stationarity is a statistical feature of time series. In other words, the mean and variance of the time series are constants independent of time. A time series with such characteristic expression is called a stationary time series. In this paper, the PM2.5 concentration series in Beijing from 2010 to 2013 in the experimental dataset is taken as a time series for the stationarity analysis. In time series stationarity detection, we determined whether the observation series was a stationary series based on the timing diagram and autocorrelation graph of the PM2.5 series in the sample. From the statistical aspect, we used ADF test on the observation sequence to further determine the stationarity of PM2.5 sequence. The temporal correlations among PM2.5 concentration time series were analyzed by the autocorrelation coefficient. The autocorrelation coefficient was statistically defined as the Pearson correlation between values at different times in a random process. For time series $\{X_t, t \in T\}$, randomly select $t, s \in T$, and the autocorrelation coefficient of time series $\{X_t\}$ is $r(s, t)$. The $r(s, t)$ calculation formula is defined as follows:

$$r(s,t) = \frac{E[(X_t - \mu_t)(X_s - \mu_s)]}{\sqrt{DX_t \cdot DX_s}} \tag{4}$$

Fig. 3 is the timing diagram of PM2.5 concentration in Beijing from 2010 to 2013. It can be observed that the waveform distribution of this series is not balanced, the fluctuation range is large, and there is no characteristic of a stationary sequence from Fig. 3. To further verify the non-stationarity of the time series, we applied the autocorrelation graph drawn by calculating the autocorrelation coefficients of the sample data to assist the identification.

The upper part of the Fig. 4 is the autocorrelation graph of the PM2.5 concentration series. No phenomenon of truncate or tailing in this time

Table 1
PM2.5 prediction error results under the each of different feature selections.

Feature selections	RMSE	MAE	R_square
Features (all)	11.52	7.42	0.9849
Features (except snow)	11.44	6.58	0.9850
Features (except rain and snow)	11.37	6.53	0.9852

Table 2
The results of the BDS test in PM2.5 concentration series.

Statistic	Z-statistic	P-value	95% CI
BDS(2)	385.861	0.000	[−1.96, 1.96]
BDS(3)	407.832	0.000	[−1.96, 1.96]
BDS(4)	433.913	0.000	[−1.96, 1.96]
BDS(5)	472.275	0.000	[−1.96, 1.96]

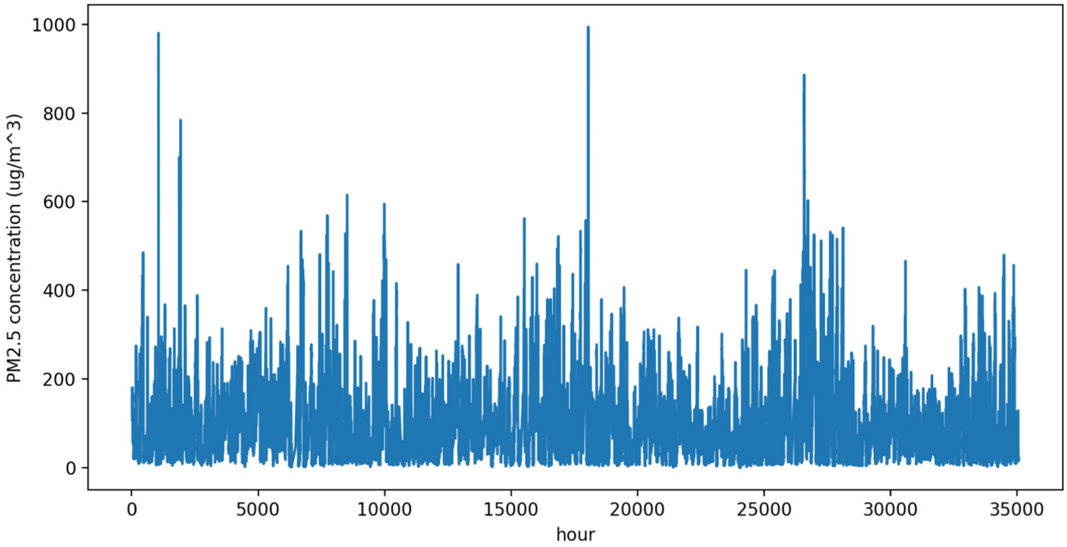


Fig. 3. Timing diagram of PM2.5 concentration in Beijing from 2010 to 2013.

series can be found in this figure. According to the stationary sequence described in literature (Wang, 2012), it usually has the properties of short-term correlation and tailing. Therefore, we can initially determine that the PM2.5 time series is a non-stationary series.

At the same time, we used the Augmented Dickey-Fuller(ADF) test method to test the original time series from an objective point of view. The ADF test is a kind of unit root test method. The principle of unit root test is when the lag operator polynomial equation of a time series has unit roots, the time series is non-stationary; conversely, when the equation does not have unit roots, the time series is stationary. The null hypothesis of ADF is that time series has a unit root. If the test statistic of ADF test is smaller than the critical value and the *P*-value is close to 0, we can reject the null hypothesis (the series is stationary). When the test statistic of ADF test is greater than the critical value, the null hypothesis cannot be rejected (this means that the series is non-stationary). The PM2.5 concentration prediction in this paper is mainly short-term forecast, so the PM2.5 concentration within a random day is taken as an example to analyze the stability of the time series. Table 3 shows the results of the ADF test. The test-statistic values calculated by the ADF test of the original time series $\{x_t\}$ are mostly greater than 10% critical values, and each *P*-value is mostly significantly greater than 0.05. So the original time series accepts the unit root null hypothesis at a significance level of 10%. It is determined that the time series $\{x_t\}$ presents a non-stationary state within four-hour interval by

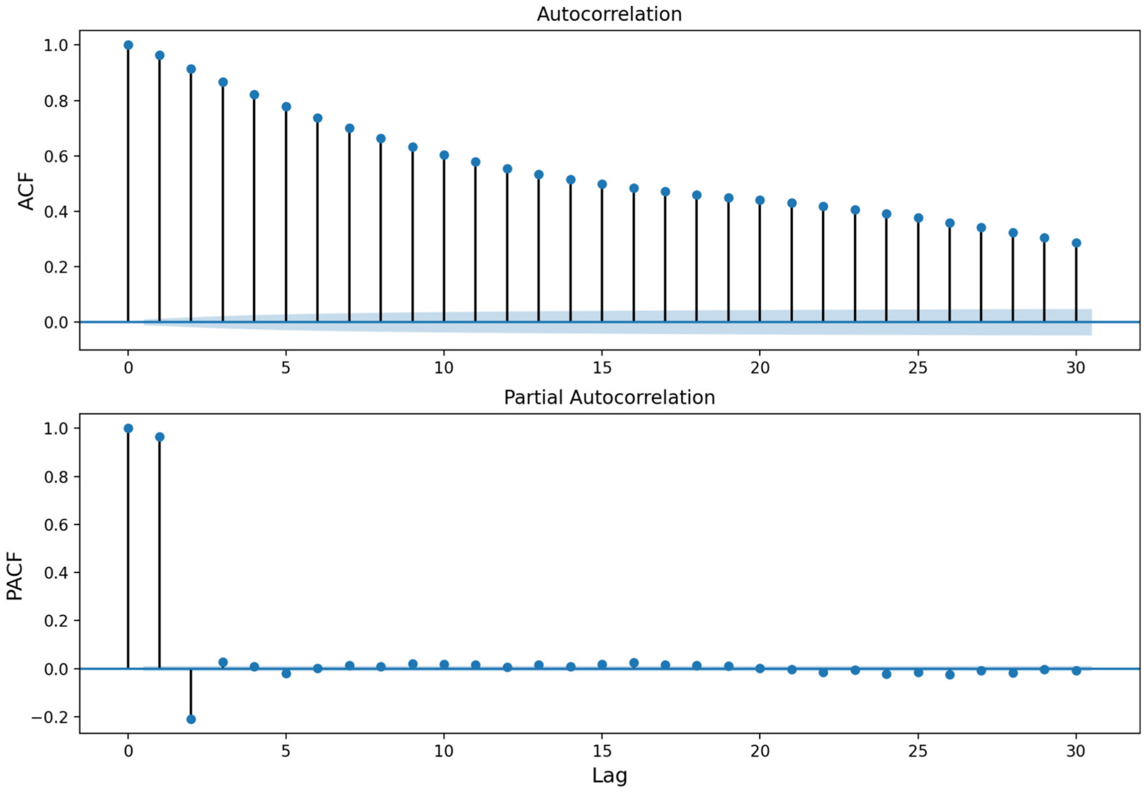


Fig. 4. Autocorrelation graph of the PM2.5 concentration series (top) and the partial autocorrelation graph (bottom).

Table 3
The results of ADF test ("*" Indicates that under the 10% significance level, "**" Indicates that under the 5% significance level. Critical Values: 1% : − 10.4172, 5% : − 5.7784, 10% : − 3.3917).

Statistic	0:00–4:00	4:00–8:00	8:00–12:00	12:00–16:00	16:00–20:00	20:00–24:00
Test-statistic	−2.6740*	−3.3067*	−12.1243	−18.5056	−3.9432**	−1.3674*
P-value	0.0786	0.0146	1.7950	2.1178	0.0017	0.5978

observed the ADF test results. Both the ACF coefficient and the results of the ADF test show that the hourly PM2.5 concentration series is a non-stationary series, so we have sufficient reasons to affirm this view.

If a time series is a non-stationary series, it means that the population from which it comes is changing. Then the analysis performed without ignoring this situation is not effective. Especially, it cannot be used to predict future events effectively. The means of decomposing a non-stationary series into multiple stationary sequences in the early stage can lay the foundation for later model training. Therefore, after identifying the PM2.5 concentration series is a non-stationary time series, we used the EMD method to decompose the PM2.5 concentration time series. The subsequences obtained by EMD decomposed are several IMFs(C_1, C_2, …, C_i) and one residue(R). All the IMFs are stationary time series of PM2.5 concentration in different frequency domains. These IMFs do not overlap with each other, and they will restore the original sequence after adding to the residual. The detailed principle of EMD's decomposition algorithm is shown in appendix A.

2.3. PM2.5 concentration prediction based on GRU neural network

In the EMD-GRU model, GRU neural network is mainly used to uniformly train the sequence obtained by EMD decomposition and selected meteorological features. After we used the trained GRU model to make predictions. The essence of PM2.5 concentration prediction is to use the series sample values observed in the past time to estimate the value of PM2.5 concentration at the next moment. It is well-known that air quality and meteorological factors belong to multi-dimensional measurement data that changes with time. As GRU neural network is particularly useful in studying time series and non-linear features, it was selected for its reliability and validity in this PM2.5 concentration prediction.

2.3.1. Time step determination

Time step refers to the difference between two adjacent time points, also known as time lag in some articles. It plays an essential role in time series prediction. The number of time step determines how many time-stamped data should be included as the input data of each unit model. Especially when constructing a multi-step recurrent neural network, the determination of the number of time step according to the characteristics of the input time series can often improve the prediction accuracy of the model. In this paper, we determine the time step of the multi-step GRU neural network by analyzing the partial autocorrelation coefficient (PACF) of the PM2.5 concentration series.

Lag k partial autocorrelation coefficient (PACF) refers to the relevant measure of the influence of X_{t-k} on X_t given the intermediate *k-1* random variables X_{t-1},X_{t-2}, …, X_{t-k+1} of time series (Wang, 2012). The calculation of the lag k partial autocorrelation coefficient (PACF) is shown in Eq. (5).

$$\rho_{X_t,X_{t-k}|X_{t-1},\dots,X_{t-k+1}} = \frac{E\left[\left(X_t - \widehat{E}X_t\right)\left(X_{t-k} - \widehat{E}X_{t-k}\right)\right]}{E\left[\left(X_{t-k} - \widehat{E}X_{t-k}\right)^2\right]} \tag{5}$$

$$\widehat{E}X_t = E[X_t|X_{t-1}, \dots, X_{t-k+1}] \tag{6}$$

$$\widehat{E}X_{t-k} = E[X_{t-k}|X_{t-1}, \dots, X_{t-k+1}] \tag{7}$$

It can be found from Fig. 4 that the autocorrelation coefficient (ACF) of PM2.5 concentration is 0.82 and the partial autocorrelation coefficient (PACF) decays to zero when the time lag $k = 4$. When k>4, the PACF always fluctuates slightly above and below zero. It is concluded that the random variables X_t, X_{t-1}, X_{t-2}, X_{t-3} and X_{t-4} have the strongest correlation in the time series.

This paper tests the RMSE and R-square of the EMD-GRU model under different number of time steps. Table 4 shows the RMSE and R-square under the different number of time step settings. With the number of time step increases, the model's prediction error appears a downward trend overall. This illustrates that inputting more information of time series into the model for training is helpful for the recurrent neural network to learn the time series better. In addition, the experimental results show that $RMSE = 11.37$ which reaches the lowest value when the time step is four hours, and then $RMSE = 11.39$ which reaches a lower value with using eight hours as time step. We find out a regular pattern from this experiment that the prediction error of the model can often be minimized when the time step is set to a multiple of four, which verifies the conclusion drawn in the partial autocorrelation graph (see Fig. 6). Taking into account the complexity of model calculation, the parameter of time step in the GRU neural network was set to four. In other words, the information of four-hour interval in the hourly PM2.5 concentration series was used as the input of the GRU unit.

2.3.2. Multi-step prediction

The process of the multi-step prediction GRU neural network deal with time series is shown in Fig. 5. At time t in the series that parameter setting of time step is four, the input data $x = \{x_{t-4}, x_{t-3}, x_{t-2}, x_{t-1}\}$ of the current moment are used as the input of the GRU neural network unit. At the same time, the hidden state value of the previous moment also was brought to the GRU network unit automatically. The output value at the current moment was obtained by unit calculating. In this way, each time step in the sequence is recursive. The parameters are defined by three matrix weights: U, V and W. They correspond to the weights of the input, output and hidden states respectively, and all of them are shared on all time steps.

In terms of neural networks, a separate multi-step prediction double-layer GRU neural network for PM2.5 prediction is built. The vector matrix $(C_i(T), M(T))$ and $(R(T), M(T))$ are respectively composed of each sequence after EMD decomposition C_i, $i \in [1.n]$, R and meteorological feature set $M(T)$. The $(C_i(T), M(T))$ and $(R(T), M(T))$ can be seen as input variables x of the GRU neural network. When the GRU neural network accomplishes the prediction, the resulting output values are $\widehat{C}_i(t)$ and $\widehat{R}(t)$. The calculation process is shown in Eqs. (9) and (10). The *GRU* function refers to the solution process of h_t in Eq. (7) of appendix B.

$$M = \left\{ f_d, f_t, f_p, f_w, f_s \right\} \tag{8}$$

Table 4
EMD-GRU model evaluation index under the setting of different time steps.

Time step	RMSE	R-square
One hour	18.28	0.9609
Two hours	13.88	0.9797
Three hours	11.82	0.9840
Four hours	11.37	0.9852
Five hours	11.51	0.9849
Six hours	11.71	0.9844
Seven hours	11.59	0.9847
Eight hours	11.39	0.9852

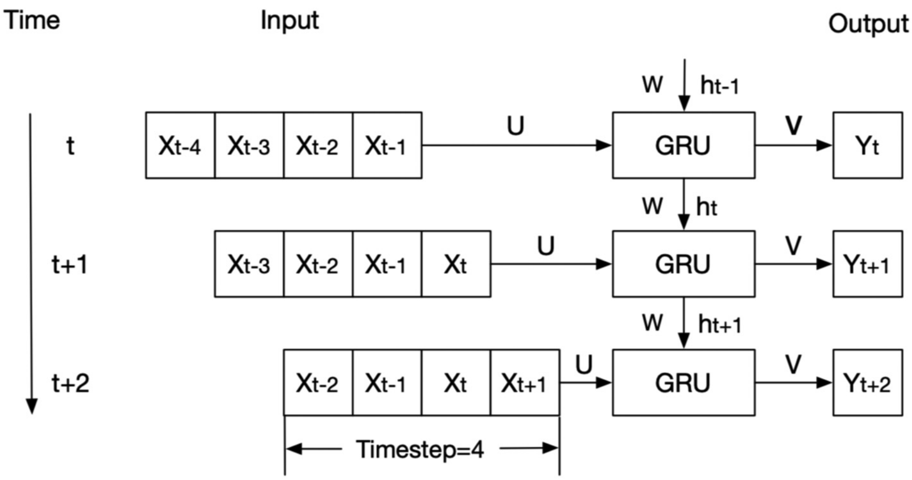


Fig. 5. Process of multi-step GRU processing time series.

$$\widehat{C}_i(t) = GRU(C_i(T), M(T)) \tag{9}$$

$$\widehat{R}(t) = GRU(R(T), M(T)) \tag{10}$$

$$T = \{t-4, t-3, t-2, t-1\} \tag{11}$$

3. Experimental results and analysis

The deep learning models involved in this article are all built using the Python programming language based on the Keras framework. The experimental environment is 64-bit Windows 7 operating system, the CPU processor is Intel Core i5-4460, and the main frequency is 3.2GHz. All experiments in this article are conducted under the same operating environment.

3.1. Data description

The dataset used in our experiment comes from the 2010/1/5–2014/12/31 PM2.5 concentration value of the US Embassy in Beijing and the meteorological data of Beijing Capital Airport provided by UCI machine learning repository (Liang et al., 2015). The data of dataset belong to time series, covering eight features including PM2.5 concentration, dew point, temperature, air pressure, wind direction, wind speed, snowfall and rainfall. The time interval for collecting all feature data is one hour, and the number of data instances is 43,800. The data types of all the features are all numerical data except that the wind direction are character data. The attribute of wind direction includes four features: NW, CV(wind speed less than or equal to 1.78 m/s), SE and NE. The experiment in this paper divides the data set into two parts: the first 80% of the data as the training data, and the remaining 20% of the data as the test data. In data preprocessing, the non-numeric feature of wind direction and filled 2067 missing values of the data set was encoded by using the linear interpolation method. Finally, the training set was transformed into the data for supervised learning, and the data of the training set and the test set were normalized.

3.2. Error measure

In terms of model evaluation, the evaluation indicators selected for all prediction models in this paper are mean square error (RMSE), mean absolute error (MAE), symmetric mean absolute percentage error (SMAPE), and R-square (R^2). RMSE is used to measure the deviation of the observed value from the true value. Compared with MSE which has the same effect in measuring the accuracy of the model, MAE is more robust for outliers. SMAPE examines the ratio between the prediction error and the true value, that is, the degree of deviation of the predicted value from the true value. The smaller the calculation results of RMSE, MAE and SMAPE are, the lower the prediction deviation is, and the better the model prediction effect is. R-square is the Goodness of Fit of the regression prediction model. The value of R-square is closer to one, indicating that the fitting effect of the forecasting model is better. The calculation formulas of evaluation indicators are shown in Eqs. (12)–(15):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |(\widehat{y}_i - y_i)| \tag{12}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\widehat{y}_i - y_i)^2} \tag{13}$$

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n} \frac{|\widehat{y}_i - y_i|}{(|\widehat{y}_i| + |y_i|)/2} \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\widehat{y}_i - y_i)^2}{\sum_{i=1}^{n} (\overline{y}_i - y_i)^2} \tag{15}$$

where y_i is the observed value of PM2.5 concentration, $\widehat{y}_i$ is the predicted value of PM2.5 concentration, $\overline{y}_i$ is the average value of PM2.5 concentration series, and n is the number of data samples.

3.3. Experimental setup

In the PM2.5 prediction of this paper, the meteorological data and the decomposed PM2.5 concentration sequence values of the past four hours are used to predict the PM2.5 concentration an hour later.

In the time series smoothing process, we applied the Python package of EMD algorithm to decompose the PM2.5 concentration series. The time series is decomposed into 18 intrinsic mode functions (IMFs) and one residual, as shown in Fig. 6.

In terms of neural network construction, a deep neural network model is constructed by stacking two layers with GRU units as hidden layers. The first GRU layer has 200 neurons, and the second one has 100 neurons. The model is continuously optimized by adjusting parameters to obtain the model with the best fitting effect on the training set. The specific parameters of the GRU neural network model are shown in Table 5.

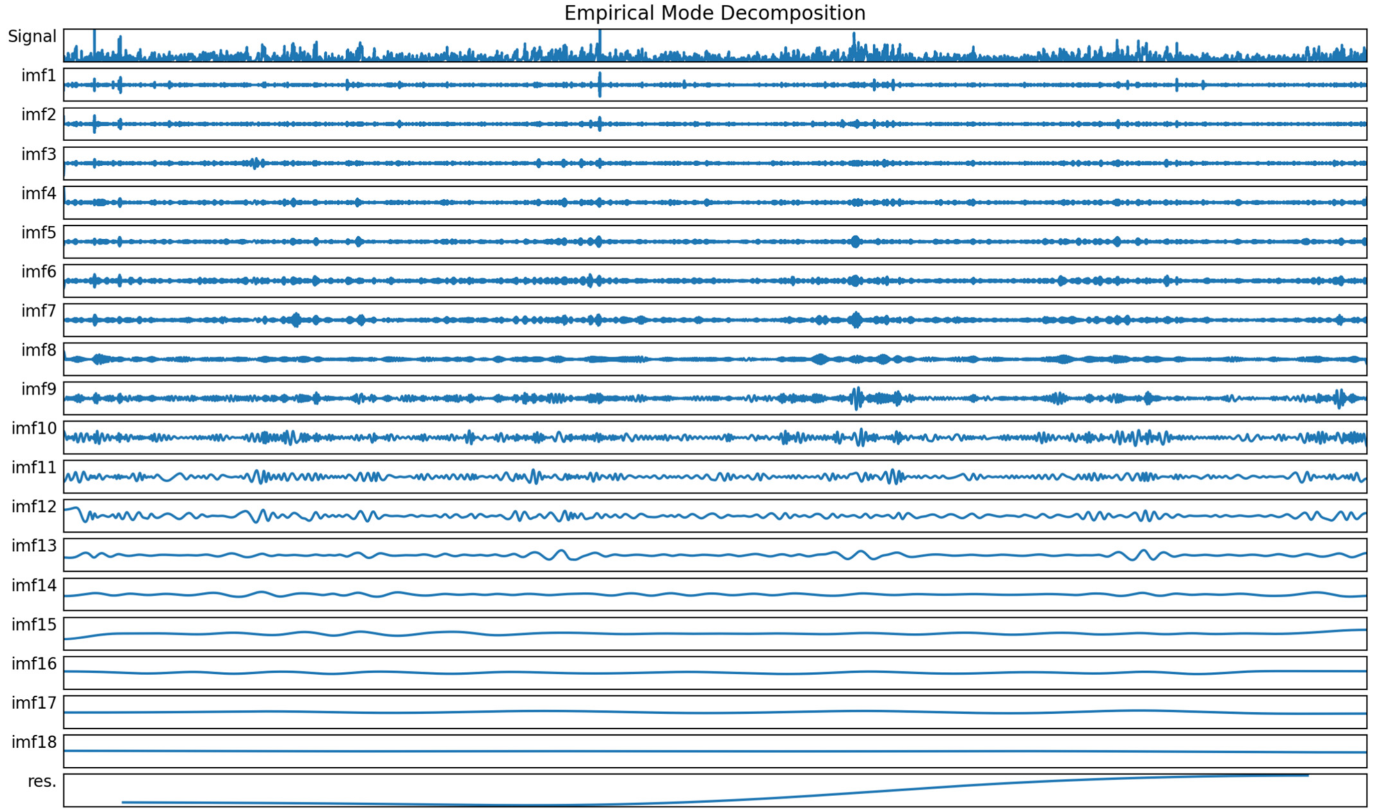


Fig. 6. IMFs and residual plot after the decomposition of 2010 to 2014 PM2.5 concentration series.

The optimizer is one of the main parameters needed to build a neural network in the Keras framework. The optimizer selected for GRU training in this paper is Adam optimization function. The learning rate of GRU is set as 0.001, and the learning rate decay over each update is 0. Loss function is defined by mean absolute error (MAE). Since the ReLU function has no gradient disappearance problem when the input data are positive numbers, the activation function in the GRU unit is set as the ReLU function in the experiment. In order to achieve the best training effect of recurrent neural network and reduce the training set error to the most stable value, a total of 30 epochs were performed in the experiment. Since the data of training set are too much, the batch size is set to 128 in each epoch. In other words, the model will update parameters after processing 128 sets of data. This operation makes the direction of gradient descent in neural network learning more accurate, and the resulting training fluctuations reach a relatively stable state. To prevent the model from overfitting, dropout technology is used in the hidden layers. The parameter of dropout is 0.2, which makes 20% of random neuron nodes in each hidden layer become invalid. This technology is used to weaken the strong dependence of some nodes and distribute the backpropagation correction value to each parameter in a balanced manner. The learning rate is reduced by using the callback function ReduceLROnPlateau to further improve the performance of the neural network when the evaluation indicators stop changing in the network. The specific operation of the callback method is that if the model performance is not seen in three batches, the learning rate is reduced at a rate of 0.7 times. The lower limit of the learning rate is set to 0.00001. After the model training, each sample point of the test set is predicted, and all the sequences obtained by the prediction are added as the final prediction values. Finally, the RMSE, MAE, SMAPE and R-square are calculated to evaluate our model.

Table 5
PM2.5 prediction error results under the each of different feature selections.

Parameter name	Value
Training set	35,040
Test set	8756
Number of GRU units	200&100
Batch size	128
Loss	MAE
Optimizer	Adam
Epochs	30
Sample weight mode	1D
Dropout	0.2
Learning rate	0.001

3.4. Results and analysis

We selected four kinds of machine learning models and three kinds of deep learning models as comparative models to compare the predictive performance of models. The four regression prediction models of machine learning are the Support Vector Machine(SVM), the Decision Tree Regressor(DTR), the Gradient Boosted Decision Trees(GBDT) and the Random Forest(RF). All of these machine learning models are built under the scikit-learn framework. Models based on deep learning algorithms include RNN, LSTM and GRU. The neural network architectures of these models are consistent with the EMD-GRU model. To ensure the validity of the experiment, all experiments are conducted under the same experimental setup, with the same training set and test set applied in the data. By the way, in order to reduce the influence of the randomness of some models on the prediction results, GBDT, RF, RNN, LSTM, GRU and the proposed EMD-GRU model in this paper have all carried out ten random repeated experiments. Therefore, the numerical results of the experiments shown in Table 6 are the average values of the evaluation index obtained from ten experiments plus its upper and lower limits.

Table 6 lists the quantitative results by RMSE, MAE, SMAPE and R-square, with comparative analyses of SVM, DTR, GBDT, RF, RNN, LSTM,

Table 6
The EMD-GRU model evaluation index under the setting of different time steps.

	Method	Parameter setting	RMSE	MAE	SMAPE(%)	R-square
ML	SVM	Kernel = RBF C = 14, gamma = 0.01	30.627±0.000	23.346±0.000	39.790±0.000	0.8934±0.0000
	DTR	Criterion = 'mse' Max depth = 3, Max leaf nodes = 8	26.299±0.000	16.578±0.000	26.553±0.000	0.9117±0.0000
	GBDT	Loss = 'huber' Min samples split = 2, Learning rate = 0.1	20.872±0.028	11.447±0.012	17.188±0.001	0.9478±0.0001
	RF	n estimators = 50 Max depth = 6 Min samples split = 3	24.841±0.175	15.188±0.162	24.773±0.175	0.9210±0.0011
DL	RNN	See Table 5	21.225±0.044	11.358±0.098	17.357±0.044	0.9488±0.0002
	LSTM	See Table 5	20.872±0.038	11.184±0.023	16.759±0.116	0.9506±0.0006
	GRU	See Table 5	20.309±0.053	11.039±0.049	16.758±0.227	0.9531±0.0002
Proposed	EMD-GRU	See Table 5	11.372±0.145	6.532±0.073	14.809±0.646	0.9852±0.0004

GRU and the proposed EMD-GRU model. The hyperparameter setting of RNN, LSTM and GRU are basically the same as that in the EMD-GRU model. From the experimental results statistically shown in Table 6, the RMSE, MAE and SMAPE of the EMD-GRU model are lower than the other seven models, and R-square is closer to 1. This proves that the EMD-GRU model has a better model fitting effect and higher prediction accuracy than traditional machine learning and deep learning models. In the experimental comparative analysis of the three models of deep learning, the average RMSE of the GRU model is 20.309, which is 1.916 lower than the average RMSE of the RNN model. It is verified that the GRU model can solve the problem that RNN cannot achieve long-term dependence in sequence prediction. Fig. 7 shows a line graph of the predicted PM2.5 concentration of the EMD-GRU model and the predicted PM2.5 concentration of three traditional deep learning models and observed PM2.5 concentration. The figure can visually show that the predicted value sequence of the EMD-GRU model represented by the red broken line is more consistent with the observed value sequence represented by the purple broken line, that is, the predicted value of the EMD-GRU model is closer to the observed value. At the same time, from the experimental data results shown in Table 6, compared with the GRU model, the prediction error evaluation index of the EMD-GRU model decreased the RMSE by 44.5%, MAE by 40.82%, and SMAPE by 11.63%. It can be explained from many aspects that the EMD-GRU model effectively solves the time lag caused by the GRU model when dealing with non-stationary time series, and further improves the accuracy of model prediction.

3.5. Comparative experiment

In order to verify the prediction accuracy of the model proposed in this article, we compared the prediction results and time cost of the model proposed in this article with other three different hybrid models, among them, the prediction results of the CNN-BGRU model are provided from the research of Tao et al. (Tao et al., 2019), and the other two are comparison models constructed by us. The data sets used in all models are the PM2.5 concentration values of the US Embassy in Beijing from January 1, 2010 to December 31, 2014, and the meteorological data of Beijing Capital Airport. In this comparative experiment, the features we selected remain the same as those used by Tao's model, and the data set is also divided into a training set, a verification set, and a test set according to the division ratio in Tao's experiment. Then the meteorological data and PM2.5 data of the past eight hours is used to predict the PM2.5 concentration after two hours by using the proposed model. Finally, two error evaluation indexes RMSE and MAE were selected to evaluate the model performance. The time cost of the model is also used as an evaluation indicator of computational cost.

The main difference between these four PM2.5 prediction models is that different feature extraction methods are used in the prediction

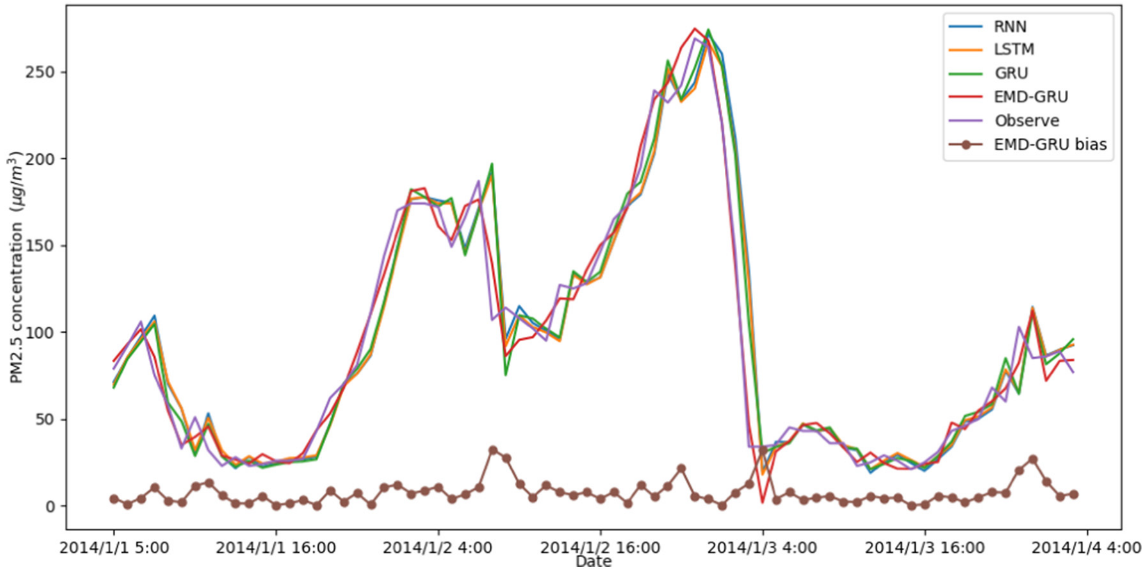


Fig. 7. Line graph of January 1, 2014 to January 4, 2014 hourly PM2.5 concentration prediction results.

Table 7
The comparison analysis of prediction errors and time cost between the CNN-BGRU, the Wavelet-GRU, the VMD-GRU and the EMD-GRU.

Method	RMSE	MAE	Time cost(min)
CNN-BGRU (Tao et al., 2019)	14.5319	10.7498	–
Wavelet-GRU	14.0386	8.8671	18.36
VMD-GRU	14.7882	9.497	12.80
EMD-GRU	11.8334	7.7031	43.00

process. The Wavelet-GRU, the VMD-GRU and the EMD-GRU model are all based on the GRU neural network and time series decomposition to estimate the PM2.5 concentration value. We respectively combined wavelet analysis and variational mode decomposition(VMD) instead of EMD method with the GRU neural network constructed in this paper to generate the Wavelet-GRU and the VMD-GRU models. Then conducted the same number of experiments on the same data set to verify the prediction effect of the model. In the Wavelet-GRU model, we applied 5-level stationary wavelet transform (5-level SWT) to decompose the original time series of PM2.5 concentration, then obtained five sequences at low frequency and one sequence at high frequency. They were obtained by applying Db1 wavelets implemented in the wavelet library Pywavelets of Python. Db1 was chosen as the wavelet function as it provided the smallest variability of time series at the particular levels. The parameters of the GRU neural network in all hybrid methods are set according to the parameter setting table in Table 5.

Table 7 is the comparison analysis of prediction errors and time cost between the CNN-BGRU, the Wavelet-GRU, the VMD-GRU and the EMD-GRU. We analyze compare experimental results from two aspects. In term of model's time cost, Since Tao et al. had did not show the time consumption of the CNN-BGRU model in their article, we did not use it as a comparison object. The VMD-GRU model has the least number of sub-sequences and the least number of iterations, so the running time cost is the lowest among the three models. In terms of error in forecast results, the EMD-GRU model proposed in this paper is the smallest in both RMSE and MAE indicators. Comprehensive comparison although the EMD-GRU model proposed in this paper has the largest time cost, the model prediction error is the smallest. But the computation course of the EMD-GRU is offline, the time consumption can be acceptable.

4. Related work

Faced with the crises caused by PM2.5 to various aspects of the ecological environment, human health, and social activities, how to use the monitored PM2.5 concentration of the city to accurately predict the future PM2.5 change trend has become a hot issue in atmospheric pollution research. The existing research methods for air prediction are mainly divided into two methods: deterministic method represented by three-dimensional air quality model and statistical methods for data-driven models. The deterministic method is mainly used by the simulation model which based on the principle of simulating atmospheric flow in three-dimensional space to predict pollutants. One of the most classic models is the CMAQ model developed by the US Environmental Protection Agency (EPA) in the late 1990s. The CMAQ (Byun and Schere, 2006) broke through the previous simulation of only a single species, realized the air quality prediction and assessment of multiple angles and multiple pollution sources (Binkowski and Roselle, 2003). The advantage of the simulation model is that it can be carried out in the absence of historical data. However, the complexity of its establishment is relatively high, and the use of default parameters that lack actual observations limits the performance of the model. Until now, the air quality predictions in the actual production process are still mainly realized by the atmospheric simulation model CMAQ or WRF-Chem at regional scale. The models by statistical approach are suitable for PM2.5 analysis of health research. Furthermore, the statistical approach now is also combined with WRF-Chem or other chemistry transport model to forecast PM 2.5. Statistical methods are supported by strong mathematical theories, so statistical models are easier to interpret than other air quality models, and are more applicable when data reserves are large. The non-linear change of air quality in urban space depends on various environmental factors, such as temperature, humidity, wind direction and other meteorological conditions. In order to better deal with these non-linear features that affect air quality, more and more machine learning methods that deal with non-linear changes are now applied to the prediction of air quality. Hou et al. (Hou et al., 2014) proposed to predict PM2.5 and PM10 concentrations in Beijing based on support vector regression (SVR) method used daily average aerosol optical depth (AOD) and meteorological parameters as reference data. Similarly, Jun Wang and Sundar A. Christopher (Wang and Christopher, 2003) confirmed that there is a linear relationship between aerosol optical thickness (AOT) and PM2.5 concentration. After that, they used the atmospheric aerosol thickness (AOT) measured by satellite-mounted MODIS to estimate the PM2.5 concentration in the Alabama area, and the accuracy of the prediction under cloudless conditions was as high as 90%.Â It can be seen that Satellite-derived AOT is a useful tool for air quality studies over large spatial domains to track and monitor aerosols. Besides, Stafoggia et al. (Stafoggia et al., 2019) used a phased random forest algorithm to estimate the average daily concentrations of PM10, PM2.5 and PM2.5–10 in a grid of one kilometer in Italy from 2013 to 2015. This method can capture the variability of most particulate matter, and raise the experimental prediction error index to 0.86. Nevertheless, the model did not achieve a good fit in the prediction of PM2.5–10 concentration levels. Zhang et al. (Zhang et al., 2020) used Kalman filter (KF) to perform deviation correction on the output of GEOS-Chem, WRF-Chem and WRF-CMAQ chemical transmission models to adjust the model's predicted output. Then two different integration methods arithmetic mean ensemble(AME) and optimized ensemble(OPE) were used to integrate the predicted output of the three models to obtain the predicted values of the daily PM2.5 concentration. In their research, KF-OPE model which their paper proposed showed the best results with the RMSE decreasing from 5.61 to 3.52 μgm^{-3} (37%). The biggest contribution of this method is that multiple model outputs and multiple satellite data products are used as ensemble members for Kalman filter processing, and the emphasis is on the synergy between surface observation networks and satellite observations for improving air quality forecasts in rural areas. In terms of neural networks, the artificial neural network (ANN) model was used to predict the PM10 concentration in Seoul Metro Station and achieved good results (Park et al., 2018). At the same time, the hybrid model based on artificial neural network further improves the accuracy of model prediction. The superiority of ANN in dealing with the functional relationship between non-linear influence factors is often impossible to achieve by other machine learning methods. However, in the current prediction of air quality, when the concentration of pollutants changes rapidly within a short period of time, the prediction results of pollutant concentration have a large deviation. Existing artificial neural networks often fail to capture the abrupt changes in particle concentration and the dependence of particle concentration as a time series. In solving this problem, this paper chose to apply deep GRU neural network in the construction of neural network. The reason is that GRU can continuously retain the valuable historical information of the input sample series during training and learning and continue these to a new round of learning. Application of GRU improves the accuracy of our prediction models.

In recent years, neural network technology has once again been pushed to a new climax with the rise of deep learning. The proposed recurrent neural network makes up for the shortcomings of artificial neural networks that cannot achieve long-term dependence. Deep learning algorithms have shown outstanding performance in applied research in multiple fields such as natural language processing, image recognition, medical diagnosis (Young et al., 2018) (Norouzzadeh et al., 2018) (Zhu et al., 2019). The air quality prediction methods based on RNN, LSTM and GRU models have also achieved good results in the past two years. For example, Bun et al. proposed a deep recurrent neural network using a self-coding pre-training method based on time series prediction to train the data collected at Japanese environmental monitoring stations. This method effectively predicted the PM2.5 concentration, and its prediction error is lower than that of the RNN model used alone (Ong et al., 2016). Due to the problem of gradient disappearance in the RNN model, S. Hochreiter and J. Schmidhuber proposed the LSTM model in 1997 to solve this problem (Hochreiter and Schmidhuber, 1997). Ma et al. (Ma et al., 2019) used bidirectional long short-term memory (BLSTM) network and inverse distance weighting (IDW) technology to predict spatially and temporally air pollutants at different time intervals. Zhao et al. (Zhao et al., 2019) implemented a long short-term memory fully connected (LSTM-FC) neural network for predicting PM2.5 pollutions at a specific air quality monitoring station within 48 h. Deep learning is stronger than machine learning in dealing with problems of highly non-linear decision functions. In the article of Karimian et al. (Karimian et al., 2019), three models based on long short-term memory network (LSTM), multiple regression trees (MRT) and deep feedforward neural network (DFNN) were applied to the regression of PM2.5 concentration values respectively. The experimental results show that the LSTM obtained the best results. The working principle of the Gated Recursive Unit (GRU) proposed in 2014 is basically the same as that of LSTM, but the GRU unit structure design is more streamlined. In the training of certain data sets, the GRU with fewer parameters can exceed the LSTM unit, no matter whether it is convergence in CPU time, or parameter update and generalization (Chung et al., 2014). Tao et al. (Tao et al., 2019) applied the hybrid model of convolutional neural network and two-layers GRU to the prediction of PM2.5 concentration levels in Beijing. These comparison experiments show that the standard error RMSE of the hybrid model is lower than that predicted by RNN and LSTM models. After studying and analyzing PM2.5 prediction methods in China and abroad, we have proposed a new hybrid model of EMD-GRU model for PM2.5 prediction. This model effectively solves the time lag generated by the recurrent neural network in predicting non-stationary series like PM2.5 concentration series and improves the performance of the cyclic neural network when processing non-stationary series. Compared with other deep learning models, the EMD-GRU model's prediction error of PM2.5 concentration is lower.

5. Discussion

This paper proposed the EMD-GRU hybrid model combined with meteorological data to predict PM2.5 concentration in Beijing. Firstly, through observation of Table 6, it can be easily found that the deep learning model is significantly better than the machine learning model in predicting PM2.5 concentration series. This proves that deep learning has a strong advantage in the prediction of time series and can effectively capture the time dependence of time series data. Secondly, when the three basic deep learning models RNN, LSTM and GRU are used to predict the PM2.5 concentration, the GRU model has the highest prediction accuracy. This indicates that GRU is more suitable for PM2.5 concentration prediction in Beijing than other deep learning models such as RNN and LSTM. The EMD-GRU model proposed in this paper is a combination of data decomposition and neural network, which further solves the performance defects of the GRU model in predicting the non-stationarity time series. Thirdly, from the results of the comparative experiment, although the EMD-GRU model has the most iterative calculations and consumes longer time, the prediction error is the smallest. Among all the hybrid models based on the signal decomposition method and GRU neural network which were proposed in our comparative experiment, the EMD-GRU model performs most prominently in the PM2.5 concentration sequence prediction in Beijing.

Due to the proposed model considers much more on fine-grained and short-term PM2.5 forecast at hour level, in the case analysis here, we chose a separate site to verify the accuracy of model prediction. But this does not mean that our method has geographic limitations in PM2.5 prediction, our model also can be applied to many different stations of different cities. In future research, air quality data and weather data from multiple stations will be used in our model to predict hourly PM2.5 concentration. Because the data set which we used lacks the relevant characteristics of PM2.5 in the spatial dimension, the EMD-GRU model proposed in this paper only analyzes the PM2.5 concentration from the perspective of time series, and the spatial prediction is not involved.

6. Conclusion

This paper proposed the EMD-GRU model based on data decomposition and recurrent neural network for predicting PM2.5 concentration in Beijing. It fully considers the relationship between meteorological features and PM2.5 concentration and the impact of non-stationarity of time series on prediction. In the early experiments, by comparing the predicted values and the observed values obtained by using the GRU model alone, it was found that most of the predicted values often lags behind the true values. To solve this problem of the GRU model, an improved GRU network based on EMD(EMD-GRU) was proved. In this method, in the form of data decomposition and neural network, the time series data was put after steadying into the GRU network to training and testing. The results of this study indicate that the proposed method greatly reduces the prediction error of the model and improves the fit of the model. By the way, the proposed method initially solved the time lag caused by the original GRU model. At the same time, the prediction accuracy of the proposed EMD-GRU model method is higher than that of traditional machine learning and a single deep learning method. All these findings suggest a role for data decomposition in promoting deep learning method. However, the present study is limited by the lack of information on spatial features related to PM2.5, and the generalization of the model in different regions cannot be explained. To this end, a lot of real and effective air quality related data will be collected later, and the model proposed in this paper will be applied to multiple different regions of data sets for verification.

CRediT authorship contribution statement

Guoyan Huang: Project administration, Funding acquisition, Investigation, Data curation. **Xinyi Li:** Conceptualization, Methodology, Software, Writing – original draft. **Bing Zhang:** Writing – review & editing, Validation, Formal analysis. **Jiadong Ren:** Supervision, Resources.

Declaration of competing interest

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

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61772451 and 61802332. The authors are grateful to the valuable comments and suggestions of the reviewers.

Appendix A. Empirical mode decomposition (EMD)

The Empirical mode decomposition (EMD) is an adaptive signal decomposition method for non-linear and non-stationary signal processing proposed by Huang (Huang et al., 1998). This method can decompose the original signal into many finite oscillation time scale components called intrinsic mode functions (IMFs) and a residual component in a self-adaptive way (Huang et al., 1999). The EMD method has a high Signal-Noise Ratio (Yeh et al., 2010). Compared with the wavelet transform (WT) method for signal decomposition, the EMD does not have the problem of preselecting wavelet basis functions like wavelet transform. On the other hand, the EMD does not require pre-defined number of IMFs like the VMD. It avoid the impact on results of time series decomposition due to unreasonable parameter settings. At the same time, because the EMD is based on the local characteristics of the time scale of the signal sequence, it has good time-frequency resolution and adaptability. Giving original time series *X(t)* ($t = 1,2, \ldots,$n) the procedure of EMD can be described as follows:

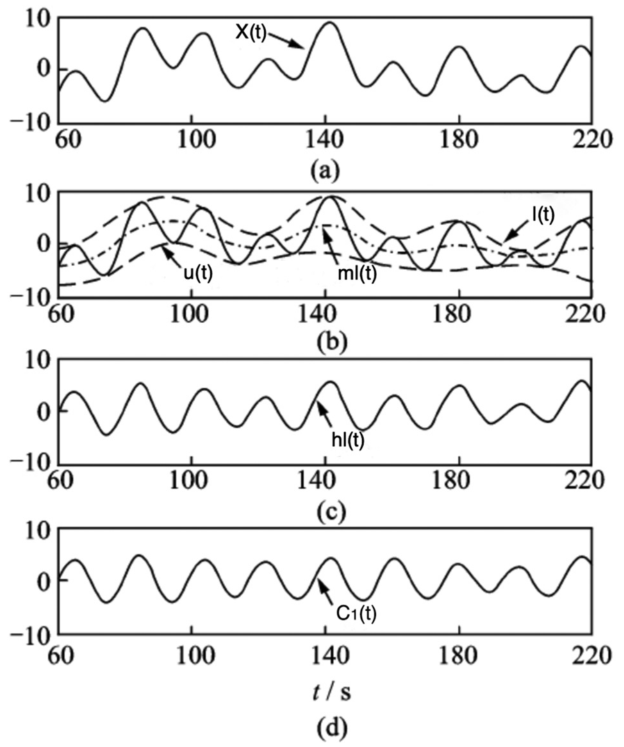


Fig. 1. Series decomposition in the EMD method.

(1) Identify all the local maxima of the original signal data series *X (t)*, and then use the three-spline interpolation function to create the upper envelope *u(t)* of the original data series.

(2) Identify all the local minima of the original signal data series *X(t)*, and then use the three-spline interpolation function to create the lower envelope *l(t)* of the original data series.

(3) Calculate average envelope *ml(t)* of the upper and lower envelopes. The mean value *ml(t)* can be computed using the following formula:

$$ml(t) = \frac{l(t) + u(t)}{2} \tag{1}$$

(4) Subtract the average envelope ml(t) from the original data series *X(t)*. The result is a new data series *hl(t)*. The *hl(t)* can be computed using the following formula:

$$hl(t) = X(t) - ml(t) \tag{2}$$

(5) Check *hl(t)*: If *hl(t)* does not exist negative local maxima and positive local minima, then *hl(t)* is defined as an IMF. The *X(t)* is replaced by the residue item $r(t) = X(t) - hl(t)$. Here, the IMF is represented as; If *hl(t)* exists negative local maxima and positive local minima, it means that *hl(t)* is not an IMF, so the *X(t)* is replaced by *hl(t)*.

(6) Repeat steps (1) to (5), until the residue item *r(t)* becomes a monotone function or the number of extrema is less than one or equal to one, so that no more IMFs can be extracted. *R(t)* indicates the tendency of the original signal data series.

Finally, the original signal data series can be reconstructed through all the decomposition IMFs and a residue. It can be expressed as the following formula:

$$X(t) = \sum_{i=1}^{n} C_i(t) + R(t) \tag{3}$$

Appendix B. GRU neural network

The Gated recurrent unit neural network (GRU) is a kind of recurrent neural network (Chung et al., 2014). Essentially, it effectively integrates and screens information in a chronological order, of which some is retained, and the other is discarded. GRU can be seen as an improved model of LSTM. Its special gating structure can effectively improve the problem of gradient disappearance caused by long time series during back propagation. The problem of gradient disappearance means that the gradient will continue to decline when the gradient propagates back with time. When the gradient value becomes very small (infinitely close to zero), there will be no further learning, and long-term dependence cannot be achieved. The GRU model effectively avoids the shortcomings of the recurrent neural network when processing long sequences.

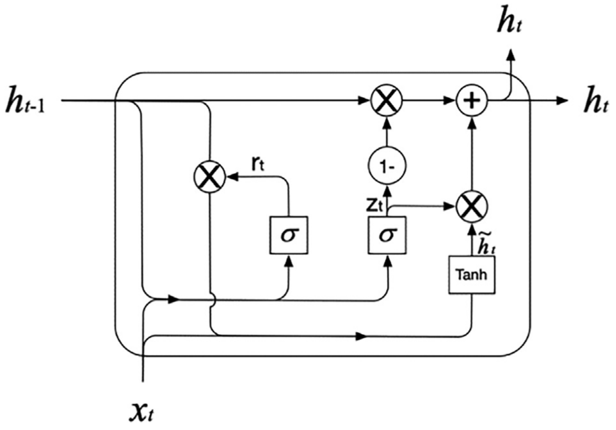


Fig. 2. Network structure of the GRU.

GRU as a variant of LSTM, uses the same design concept to solve the gradient disappearance problem. It is worth mentioning that its internal structure is simpler than LSTM. GRU's unit structure contains only two "gates": the update gate(r_t) and the reset gate(z_t). In details, the purpose of the update gate is to decide which information to forget and which information need to be retained. The reset gate is used to decide the degree of forgetting of previous information. Fig. 2 presents the network structure of GRU. Each gate of GRU and output of the hidden layer are calculated as follows:

$$r_t = \sigma(x_t W_{xr} + h_{t-1} W_{hr} + b_r) \tag{4}$$

$$z_t = \sigma(x_t W_{xz} + h_{t-1} W_{hz} + b_z) \tag{5}$$

$$\widetilde{h}_t = \ tanh\,(x_t W_{xh} + r_t \otimes h_{t-1} W_{hh} + b_h) \tag{6}$$

$$h_t = z_t + (1 - z_t) \otimes \widetilde{h}_t \tag{7}$$

where σ is the activation function; x_t is the present input; h_{t-1} is the previous output; W_{xr} and W_{hr} are the weights of the update gate; W_{xz} and W_{hz} are the weights of the reset gate; W_{xh} and W_{hh} are the weights of the output candidate value; b_r, b_z and b_h are bias vectors of the update gate, the reset gate and the output candidate value $\widetilde{h}_t$ respectively. The operator "⊗" means to sequentially multiply the elements of the array.

References

Binkowski, F.S., Roselle, S.J., 2003. Models-3 Community Multiscale Air Quality(CMAQ) model aerosol component 1. Model description. J. Geophys. Res. - Atmos. 108, 463. https://doi.org/10.1029/2001JD001409.

Box, G.E.P., Jenkins, G.M., 2010. Time series analysis, forecasting and control. Journal of Time 31 (3). https://doi.org/10.1111/j.1467-9892.2009.00643.x.

Brock, William A., Hsieh, D.A., Lebaron, B., 1992. Nonlinear dynamics, chaos, and instability. Journal of Finance 48 (1), 202–203. https://doi.org/10.2307/2328898.

Burnett, R., et al., 2018. Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proc. Natl. Acad. Sci. 115, 9592–9597. https://doi.org/10.1073/pnas.1803222115.

Byun, D., Schere, K.L., 2006. Review of the governing equations, computational algorithms, and other components of the models-3 Community Multiscale Air Quality (CMAQ). Appl. Mech. Rev. 59, 51–77. https://doi.org/10.1115/1.2128636.

Calvo, A.I., Alves, C., Castro, A., Pont, V., Vicente, A.M., Fraile, R., 2013. Research on aerosol sources and chemical composition: past, current and emerging issues. Atmos. Res. 120–121, 1–28. https://doi.org/10.1016/j.atmosres.2012.09.021.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., December 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS 2014 Deep Learning and Representation Learning Workshop http://arxiv.org/licenses/nonexclusive-distrib/1.0/.

Gao, M., Guttikunda, S.K., Carmichael, G.R., Wang, Y., Liu, Z., Stanier, C.O., Saide, P.E., Yu, M., 2015. Health impacts and economic losses assessment of the 2013 severehaze event in Beijing area. Sci. Total Environ. 511, 553–561. https://doi.org/10.1016/j.scitotenv.2015.01.005.

He, J., Yu, Y., Liu, N., Zhao, S., 2013. Numerical model-based relationship between meteorological conditions and air quality and its implication for urban air quality management. Int. J. Environ. Pollut. 53 (3/4), 265–286. https://doi.org/10.1504/IJEP.2013.059921.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hou, W., Li, Z., Zhang, Y., Xu, H., Zhang, Y., Li, K., Li, D., Wei, P., Ma, Y., 2014. Using support vector regression to predict PM10 and PM2.5. 35th International Symposium on Remote Sensing of Environment (ISRSE-35), IOP Conference Series-Earth and Environmental Science. 17, p. 012268. https://doi.org/10.1088/1755-1315/17/1/012268.

Huang, H., et al., 1998. The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis. Procedings of the Royal Society. 454, 903–995. https://doi.org/10.1146/annurev.fluid.31.1.417.

Huang, N.E., Shen, Z., Long, S.R., 1999. A new view of nonlinear water waves: the hilbert spectrum. Annu. Rev. Fluid Mech. 31, 417–457. https://doi.org/10.1146/annurev.fluid.31.1.417.

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X., Sachdeva, S., 2019. An improved method for monitoring fine particulate matter mass concentrations via satellite remote sensing. Aerosol Air Qual. Res. 19, 1400–1410. https://doi.org/10.4209/aaqr.2015.06.0424.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., Chen, S., 2015. Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. P. Roy. Soc. A-Math. Phy. 471 (2182). https://doi.org/10.1098/rspa.2015.0257.

Ma, J., Ding, Y., Gan, V., Lin, C., Wan, Z., 2019. Spatiotemporal prediction of PM2.5 concentrations at different time granularities using IDW-BLSTM. IEEE Access 7, 107897–107907. https://doi.org/10.1109/ACCESS.2019.2932445.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J., 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. 115 (25), E5716–E5725. https://doi.org/10.1073/pnas.1719367115.

Ong, B.T., S., K., Zettsu, K., 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. Neural Comput. & Applic. 27, 1553–1566. https://doi.org/10.1007/s00521-015-1955-3.

Park, S., Kim, M., Kim, M., Namgung, H., Kim, K., Cho, K.H., Kwon, S., 2018. Predicting PK10 concentration in Seoul metropolitan subway stations using artificial neural network (ANN). J. Hazard. Mater. 341, 75–82. https://doi.org/10.1016/j.jhazmat.2017.07.050.

Pearson, K., 1895. Note on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. 58, 240–242. https://doi.org/10.1098/rspl.1895.0041.

Pérez, P., Trier, A., Reyes, J., 2000. Prediction of PM 2.5, concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. 34 (8), 1189–1196. https://doi.org/10.1016/S1352-2310(99)00316-7.

Pun, V.C., Kazemiparkouhi, F., Manjourides, J., Suh, H.H., 2017. Long-term PM2.5 exposure and respiratory, cancer, and cardiovascular mortality in older US adults. Am. J. Epidemiol. 186 (8), 961–969. https://doi.org/10.1093/aje/kwx166.

Rabbouch, H., Saâdaoui, F., Mraihi, R., 2018. A vision-based statistical methodology for automatically modeling continuous urban traffic flows. Adv. Eng. Inform. 38, 392–403. https://doi.org/10.1016/j.aei.2018.08.006.

Saâdaoui, F., Ben Messaoud, O., 2020. Multiscaled neural autoregressive distributed lag: a new empirical mode decomposition model for nonlinear time series forecasting. International Journal of Neural Systems 30 (8). https://doi.org/10.1142/S0129065720500392 2050039.

Saâdaoui, F., Rabbouch, H., 2019. A wavelet-based hybrid neural network for short-term electricity prices forecasting. Artif Intell Rev. 52, 649–669. https://doi.org/10.1007/s10462-019-09702-x.

Saâdaoui, F., Saadaoui, H., Rabbouch, H., 2020. Hybrid feedforward ANN with NLS-based regression curve fitting for US air traffic forecasting. Neural Comput. & Applic. 32, 10073–10085. https://doi.org/10.1007/s00521-019-04539-5.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., Hoogh, K., de'Donato, F., et al., 2019. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013-2015, using a spatio-temporal land-use random-forest model. Environ. Int. 124, 170–179. https://doi.org/10.1016/j.envint.2019.01.016.

Tao, Q., Liu, F., Li, Y., Sidoro, D., 2019. Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. IEEE Access. 7, 76690–76698. https://doi.org/10.1109/ACCESS.2019.2921578.

Wang, Y., 2012. Applied Time Series Analysis(3rd Ed). Renmin University of China Press, China.

Wang, J., Christopher, S.A., 2003. Intercomparison between satellite-derived aerosol optical thickness and PM2.5 mass: Implications for air quality studies. Geophys. Res. Lett. 30, 2095. https://doi.org/10.1029/2003GL018174.

Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., Chi, T., 2019. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Sci. Total Environ. 654, 1091–1099. https://doi.org/10.1016/j.scitotenv.2018.11.086.

Yeh, J.R., Shieh, J.S., Huang, N.E., 2010. Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method. Adv. Adapt. Data Anal. 2, 135–156. https://doi.org/10.1142/S1793536910000422.

Young, T., Hazarika, D., Poria, S., Cambria, E., 2018. Recent trends in deep learning based natural languageprocessing. IEEE Comput. Intell. M. 13, 55–75. https://doi.org/10.1109/MCI.2018.2840738.

Zhang, H., Wang, J., García, L.C., Ge, C., Plessel, T., Szykman, J., et al., 2020. Improving surface PM2.5 forecasts in the united states using an ensemble of chemical transport model outputs: 1. bias correction with surface observations in nonrural areas. Journal of Geophysical Research: Atmospheres 125 (14). https://doi.org/10.1029/2019JD032293.

Zhao, D., Chen, H., Sun, X., Shi, Z., 2018. Spatio-temporal variation of PM2.5 pollution and its relationship with meteorology among five megacities in China, aerosol and air quality. Research. 18, 2318–2331. https://doi.org/10.4209/aaqr.2017.09.0351.

Zhao, J., Deng, F., Cai, Y., Chen, J., 2019. Long short-term memory—fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. Chemosphere. 220, 486–492. https://doi.org/10.1016/j.chemosphere.2018.12.128.

Zheng, S., Wang, J., Sun, C., Zhang, X., Kahn, M.E., 2019. Air pollution lowers Chinese urbanites' expressed happiness on social media. Nat. Hum. Behav. 3, 237–243. https://doi.org/10.1038/s41562-018-0521-2.

Zhu, Y., Wang, Q., Xu, M., Zhang, Z., Cheng, J., Zhong, Y., et al., 2019. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointest. Endosc. 89, 806–815. https://doi.org/10.1016/j.gie.2018.11.011.