Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-021-12657-8

RESEARCH ARTICLE

A hybrid deep learning technology for $PM_{2.5}$ air quality forecasting

Zhendong Zhang [1] · Yongkang Zeng [1] · Ke Yan [1,2]

Received: 23 October 2020 / Accepted: 20 January 2021

Abstract

The concentration of $PM_{2.5}$ is one of the main factors in evaluating the air quality in environmental science. The severe level of $PM_{2.5}$ directly affects the public health, economics and social development. Due to the strong nonlinearity and instability of the air quality, it is difficult to predict the volatile changes of $PM_{2.5}$ over time. In this paper, a hybrid deep learning model VMD-BiLSTM is constructed, which combines variational mode decomposition (VMD) and bidirectional long short-term memory network (BiLSTM), to predict $PM_{2.5}$ changes in cities in China. VMD decomposes the original $PM_{2.5}$ complex time series data into multiple sub-signal components according to the frequency domain. Then, BiLSTM is employed to predict each sub-signal component separately, which significantly improved forecasting accuracy. Through a comprehensive study with existing models, such as the EMD-based models and other VMD-based models, we justify the outperformance of the proposed VMD-BiLSTM model over all compared models. The results show that the prediction results are significantly improved with the proposed forecasting framework. And the prediction models integrating VMD are better than those integrating EMD. Among all the models integrating VMD, the proposed VMD-BiLSTM model is the most stable forecasting method.

Keywords Air quality prediction · Deep learning · Variational mode decomposition · Bidirectional long · Short-term memory network

Introduction

Due to the acceleration of global urbanization, the air quality decreases significantly around the world in general. Air pollution incidents or abnormal weather have significantly increased in many countries, and such pollution severely threatens local lives and social development (Chan and Yao 2008; Zhao et al. 2019; Li et al. 2019a). Therefore, in recent years, strengthening the monitoring of air quality and the forecasting of air pollution have become an increasingly popular topic in the rapidly developing artificial intelligence (AI) technology. $PM_{2.5}$ refers to fine particles in the air with an aerodynamic equivalent diameter of 2.5 microns or less. Compared with other coarser atmospheric particles, $PM_{2.5}$ particles are smaller and have a longer transmission distance (Morillas et al. 2019; Wu et al. 2019). The $PM_{2.5}$ level therefore is important for human health (O'Donnell et al. 2011; Morillas et al. 2019).

In recent years, the forecasting of $PM_{2.5}$ has aroused the attention of researchers. There are mainly two types of forecasting methods of $PM_{2.5}$ in the literature, including numerical modeling and statistical modeling. The numerical modeling method builds models based on physical and chemical properties through numerical calculations. The prediction is conducted by analysing distributions of $PM_{2.5}$ in the atmosphere and solving the conservation equations. The numerical modeling method is usually more practical, providing systematic prediction results. However, it requires detailed and precise information on regional climate, mountain landforms and the distribution of pollution sources (Li et al. 2019a, b; Hao et al. 2020).

With the continuous development of AI technology, statistical modeling, especially models that integrate machine learning (ML) and deep learning (DL) techniques, attracts attentions from a wide range of scientists belonging to various fields (Ding et al. 2019; Wang and Zhang 2020). Most of

Responsible Editor: Marcus Schulz

✉ Ke Yan
keddiyan@gmail.com

[1] Key Laboratory of Electromagnetic Wave Information Technology and Metrology of Zhejiang Province, College of Information Engineering, China Jiliang University, Hangzhou 310018, China

[2] National University of Singapore, Singapore 117566, Singapore

Published online: 23 March 2021

the statistical models are also called data-driven methods. These data-driven methods fit the target data by utilizing many samples to refine and approximate the real model continuously. The state-of-art data-driven method is called long short-term memory neural network (LSTM) (Hochreiter and Schmidhuber 1997). In recent years, the LSTM neural network is widely applied in prediction problems (Wu and Lin 2019a; Chang et al. 2020). LSTM is a recurrent neural network (RNN) with a specialized internal structure. With its unique gated structure, LSTM prevents the gradient disappearance and gradient explosion by controlling the selective processing of information, thus avoiding the problem of long-term dependence (Gers et al. 2000).

In this study, the cutting-edge bidirectional long short-term memory neural network (BiLSTM) is proposed to predict time series data of $PM_{2.5}$. The principle of BiLSTM is that each training sequence comprises two LSTM neuron sequences (forward and backward), both connecting to the same output layer. BiLSTM effectively learns the input data from two temporal directions (Wu et al. 2020). In addition, the variational mode decomposition (VMD) is employed to decompose the $PM_{2.5}$ non-periodic signal in the frequency domain. The input signal frequency domain complexity is reduced by decomposing the complex signal into multiple harmonic subsequences (Han et al. 2019; Gendeel et al. 2018). The real-world air quality data collected in Beijing, China, during 2013–2017 is utilized for the comparative study conducted in Section IV (experimental process and results). And five evaluation metrics are employed to evaluate the forecasting performance. Experimental results show that the proposed method is superior to existing data-driven $PM_{2.5}$ forecasting methods.

Related works

Time series forecasting is a popular topic in AI, especially when the time series is dynamic and non-linear. For non-linear time series data forecasting problems, AI-based methods usually achieve better performance compared with the traditional approaches.

Based on a specialized internal structure of neurons, LSTM is widely implemented in a variety of time series forecasting fields. In terms of household electricity forecasting and photovoltaic power generation forecasting, LSTM and its extensions show outstanding performance. Yan et al. (2018) employed CNN to accurately extract one-dimensional household electricity data features and apply LSTM to predict household energy consumption. The wavelet decomposition was introduced to decompose the household electric energy signal into signals of different frequencies for processing (Yan et al. 2019). The process of wavelet decomposition reduced the complexity of the frequency domain. The decomposed data then input it into the LSTM unit for prediction. Compared with the model without wavelet decomposition, the forecasting performance is effectively promoted. Given the periodicity and volatility, Zhou et al. (2019) added the attention mechanism to the LSTM unit to increase the focus upon CNN feature extraction. In addition, LSTM is also applied to other fields. Ding and Qin (2019) proposed a LSTM-based multiple-input multiple-output recurrent neural network model to predict stock price fluctuation trends and the highest and lowest opening prices. Jin et al. (2020) employed LSTM to recognize text data. Xie et al. (2019) applied the LSTM neural network to classify language emotions (SEC).

LSTM and its extensions are widely adopted for air quality forecasting. Li et al. (2020) proposed a hybrid deep learning model combining convolutional neural network (CNN) and LSTM to predict the concentration of air pollutants $PM_{2.5}$ in the next 24 h. Wang et al. (2020) combined chi-square test (CT) with LSTM for air quality forecasting. The prediction accuracy is higher than traditional machine learning methods, including support vector machine regression (SVR), multi-layer perceptron (MLP), back-propagation (BP) neural network and RNN. Wu and Lin (2019a) used LSTM to predict the high-frequency sub-sequence WD(D) after wavelet decomposition in AQI and used the least square support vector machine (LSSVM) to predict the low-frequency sub-sequence WD(A), by combining the two methods to achieve efficient prediction accuracy. Xu and Yoneda (2019) proposed a multi-tasking LSTM auto-encoder model for air quality prediction. The PM2.5 index in Beijing, China, was investigated. The LSTM auto-encode model was capable of encoding key evolution pattern of urban meteorological systems and providing superior performance compared with traditional prediction models.

From many of the above works, for unstable and non-linear time series data forecasting, a good prediction model includes both signal processing and deep learning models. Useful signal processing functions can effectively improve the accuracy of prediction. Current effective signal processing methods include variational mode decomposition (VMD), empirical mode decomposition (EMD), Kalman filter (KF), singular value decomposition (SVD) and wavelet transform (WT). Zheng et al. (2017) proposed an EMD-LSTM hybrid model combining empirical mode decomposition (EMD) and LSTM, which decomposes the power load data signal into multiple intrinsic mode functions (IMF) for prediction and evaluation. Wu et al. (2020) employed singular value decomposition to reconstruct the original cutting force signal of the tool and then used BiLSTM to predict the characteristic subsignal, which effectively improves the prediction accuracy. Chang et al. (2019) applied wavelet transform (WT) to decompose the electricity price data, and the processed data will have a more stable variance. Combined with the Adam optimizer, a hybrid model WT-Adam-LSTM is proposed. Song

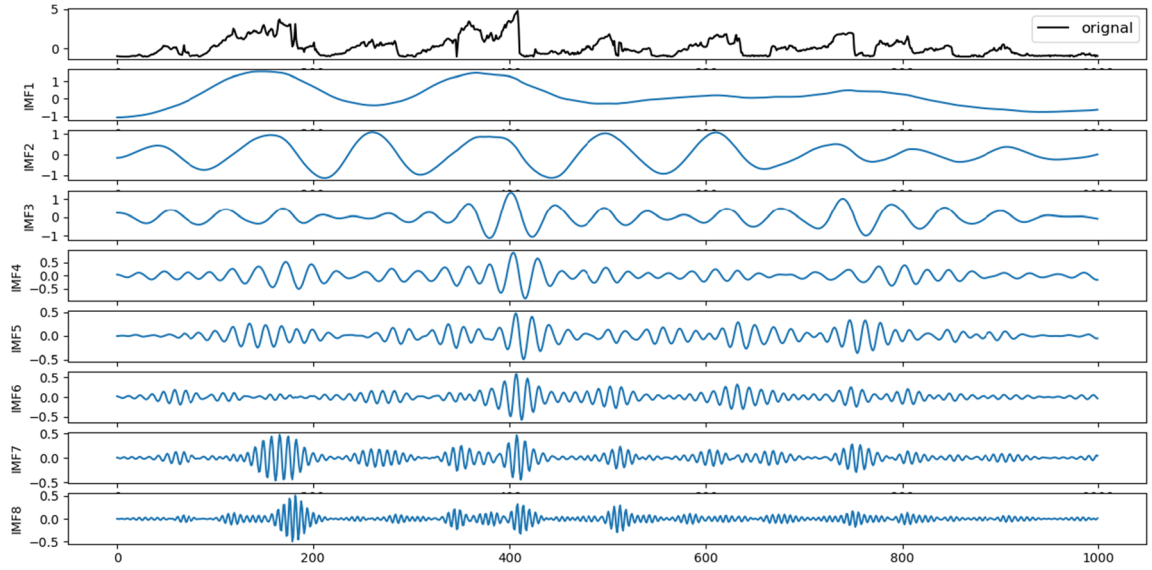

Fig. 1 The result of variational mode decomposition

et al. (2019) employed Kalman filter (KF) to process the original data then input it into LSTM units, forming an LSTM-Kalman hybrid model to predict the concentration of air pollutants. Wu and Lin (2019b) used sample entropy (SE) to reorganize the sub-sequences decomposed by VMD, and then used LSTM to predict the new sequence. In such a way, the computational complexity was reduced. And the signal over-decomposition problem was prevented.

Methodology

Data and standardization

The dataset used in this study is from the UCI machine learning knowledge base, which is the air quality data of 12 observatories around Beijing recorded by the US Embassy from 2013 to 2017 (Zhang et al. 2017). The original data set was sampled and recorded at hourly intervals, with 35,064 samples (1461 days).

In the test experiment, the first 24,000 data samples (1000 days) of the first 26,400 data from Changping Observatory are selected as the training set, and the last 2400 data (100 days) as the test set for the experiment.

Data standardization is one of the critical steps of time series forecasting. The zero-mean standardization method is applied to process the data. Data standardization with a mean value of 0 and a standard deviation of 1 that obey the standard normal distribution is obtained. The standardization and de-standardization formulas are:

$$x^{*} = \frac{x - \bar{x}}{\sigma} \tag{1}$$

$$x = \sigma x^{*} + \bar{x} \tag{2}$$

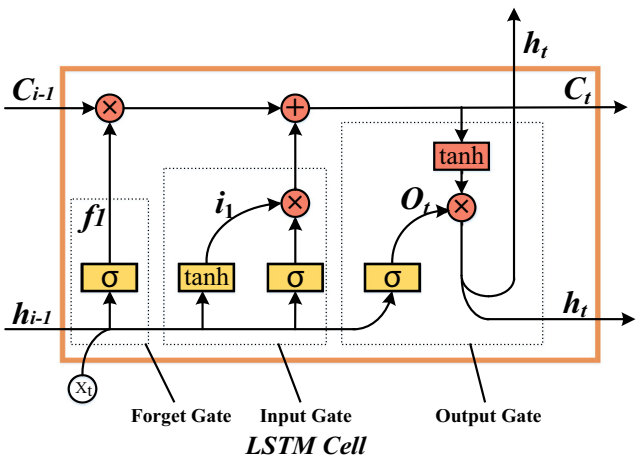

Fig. 2 Internal structure of a traditional LSTM neuron

Variational mode decomposition (VMD)

Due to the violent fluctuation and the complex distribution in the original $PM_{2.5}$ data frequency domain, it is challenging to achieve high-precision prediction. Therefore, the method of variational mode decomposition (VMD) is employed to stationarize the input data. The original $PM_{2.5}$ data signal is adaptively decomposed into several intrinsic mode functions (IMFs) by VMD. This decomposition effectively reduces the nonlinearity and volatility, thereby achieving signal stability (Liang et al. 2020). When ensuring that the decomposition sequence is IMFs with a limited bandwidth with a center frequency, VMD must finally minimize the sum of the estimated bandwidth of each mode.

The specific resolution process is as follows.

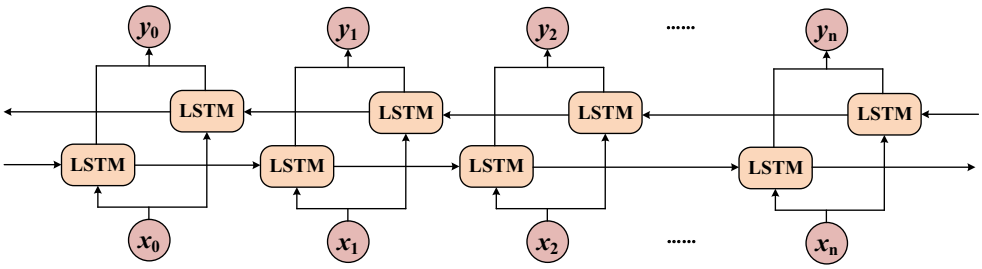

Fig. 3 Internal structure of BiLSTM unit

Construct a variational problem, decompose $PM_{2.5}$ data into K mode variables, each mode $u_k(t)$ is a finite bandwidth with a center frequency ω_k. The variational problem's objective function is the minimum sum of the estimated bandwidth of each mode, and the constraint condition is that the sum of the mode components is equal to the original signal. Then, the corresponding expressions are

$$\min_{u_k,\omega_k} \sum_k \left|\left|\partial_t[(\delta(t)+j/\pi t)*u_k(t)]e^{-j\omega_k t}\right|\right|_2^2 \tag{3}$$

$$s.t. \sum_{k=1}^{K} u_k = f \tag{4}$$

where $\delta(t)$ is the Dirac function, f is the original signal and * is the convolution operator.

To obtain the constrained variable problem's optimal solution, the penalty factor α and the Lagrangian multiplication operator $\lambda(t)$ are introduced to transform the constrained variational problem into an unconstrained variational problem. The definitions are

$$\begin{aligned} L(u_k,\omega_k,\lambda) = \alpha \sum_k \left|\left|\partial_t[(\delta(t)+j/\pi t)*u_k(t)]e^{-j\omega_k t}\right|\right|_2^2 \\ + \left\|f(t)-\sum_k u_k(t)\right\|_2^2 \\ + \left[\lambda(t), f(t)-\sum_k u_k(t)\right] \end{aligned} \tag{5}$$

Finally, the alternating direction multiplier (ADMM) iterative algorithm is applied to optimize the obtained mode components and center frequencies to solve the optimal solution of the augmented Lagrangian equation iterator of the equation. The iteration process is as follows:

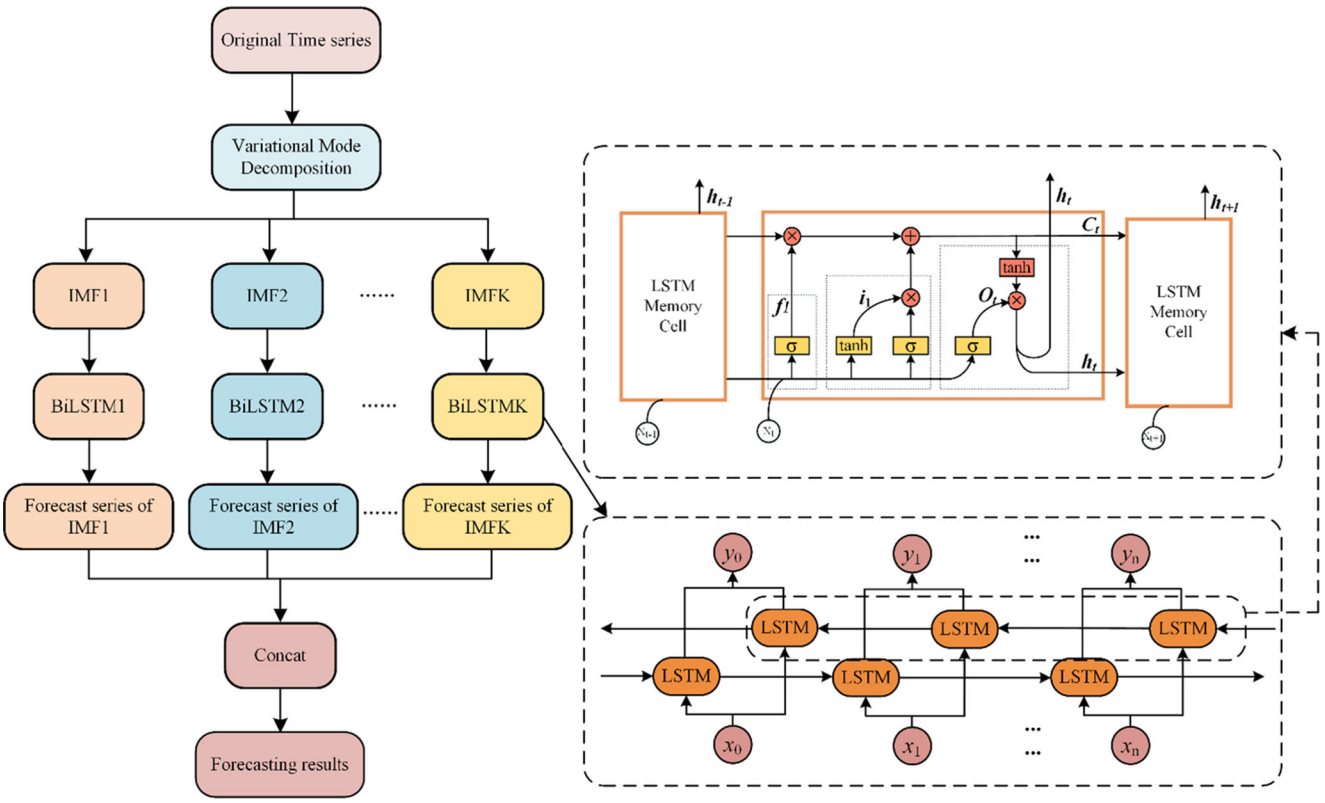

Fig. 4 Experiment flow chart of VMD-BiLSTM

Table 1 Test results of different K value

Value of K	MAE	RMSE%	MAPE	R^2	ACC	Time(s)
3	9.61	19.03	26.22	0.9772	61.358	76.972
4	8.23	15.59	24.49	0.9847	66.24	105.06
5	7.339	13.732	22.233	0.9881	68.112	127.416
6	6.743	12.296	19.922	0.9904	70.071	157.598
7	6.232	11.818	18.024	0.991	73.155	180.792
8	5.482	10.46	16.889	0.993	75.364	210.507
9	5.73	10.157	16.516	0.9935	76.2405	242.55

$$\widehat{u}_{\mathrm{k}}^{\mathrm{n+1}}(\omega)=\frac{\widehat{f}(\omega)-\sum_{\mathrm{i}=\mathrm{k}}\widehat{u}_{\mathrm{i}}(\omega)+\frac{\widehat{\lambda}(\omega)}{2}}{1+2\alpha(\omega-\omega_{\mathrm{k}})^{2}} \tag{6}$$

$$\omega_{k}^{n+1}=\frac{\int_{0}^{\infty}\omega\left|\widehat{u}_{k}(\omega)\right|^{2}d\omega}{\int_{0}^{\infty}\left|\widehat{u}_{\mathrm{k}}(\omega)\right|^{2}d\omega} \tag{7}$$

$$\widehat{\lambda}^{n+1}(\omega)=\widehat{\lambda}^{n}(\omega)+\gamma\left(\widehat{f}(\omega)-\sum_{k}\widehat{u}_{\mathrm{k}}^{\mathrm{n+1}}(\omega)\right) \tag{8}$$

where γ is the noise tolerance, $\widehat{u}_{\mathrm{k}}^{\mathrm{n+1}}$, $\omega_{\mathrm{k}}^{\mathrm{n+1}}$ and $\widehat{\lambda}^{n+1}$ are the Fourier transform values newly generated in the iteration.

Figure 1 shows the original PM2.5 time series data waveform after VMD decomposition, and the waveform is distributed from low frequency to high frequency.

Bidirectional LSTM

In the traditional recurrent neural network (RNN) model and the long- and short-term memory recurrent neural network (LSTM) model, information can only propagate forward. This makes the current state information output by the model solely depend on the lead at present. BiLSTM can obtain information about the distribution of periods before and after combining the bidirectional (BiRNN) model and the LSTM unit (Fig. 2).

This paper chooses the BiLSTM model to predict PM2.5 data and obtain the forward and backward characteristics of PM2.5, respectively. This mechanism enables BiLSTM to receive more comprehensive feature information and improve the prediction accuracy of experimental results.

In the LSTM unit, the decisive role is the forget gate, which ignores the historical information selectively. The second is the input gate and output gate, which are used to input and output the current unit (Moniz and Krueger 2018). The expression of each equation is

$$i_{\mathrm{t}}=\sigma_{\mathrm{i}}(x_{\mathrm{t}}W_{\mathrm{xi}}+h_{\mathrm{t-1}}W_{\mathrm{hi}}+b_{\mathrm{i}}) \tag{9}$$

$$f_{\mathrm{t}}=\sigma_{\mathrm{f}}(x_{\mathrm{t}}W_{\mathrm{xf}}+h_{\mathrm{t-1}}W_{\mathrm{hf}}+b_{\mathrm{f}}) \tag{10}$$

$$c_{\mathrm{t}}=f_{\mathrm{t}}\times c_{\mathrm{t-1}}+i_{\mathrm{t}}\times\sigma_{\mathrm{c}}(x_{\mathrm{t}}W_{\mathrm{xc}}+h_{\mathrm{t-1}}W_{\mathrm{hc}}+b_{\mathrm{c}}) \tag{11}$$

$$o_{\mathrm{t}}=\sigma_{\mathrm{o}}(x_{\mathrm{t}}W_{\mathrm{xo}}+h_{\mathrm{t-1}}W_{\mathrm{ho}}+b_{\mathrm{o}}) \tag{12}$$

$$h_{\mathrm{t}}=o_{\mathrm{t}}\times\sigma_{\mathrm{h}}(c_{\mathrm{t}}) \tag{13}$$

Figure 3 shows the internal unit structure of BiLSTM. It is seen from the figure that the BiLSTM model consists of a forward LSTM and a backward LSTM. This makes it obtain the forward and backward characteristics of the current time, respectively, and compared with LSTM, the output of BiLSTM is more affected by the data before and after. Because the air quality data fluctuates significantly over time and strongly correlates with the state before and after, this paper intends to use BiLSTM to predict PM2.5 data.

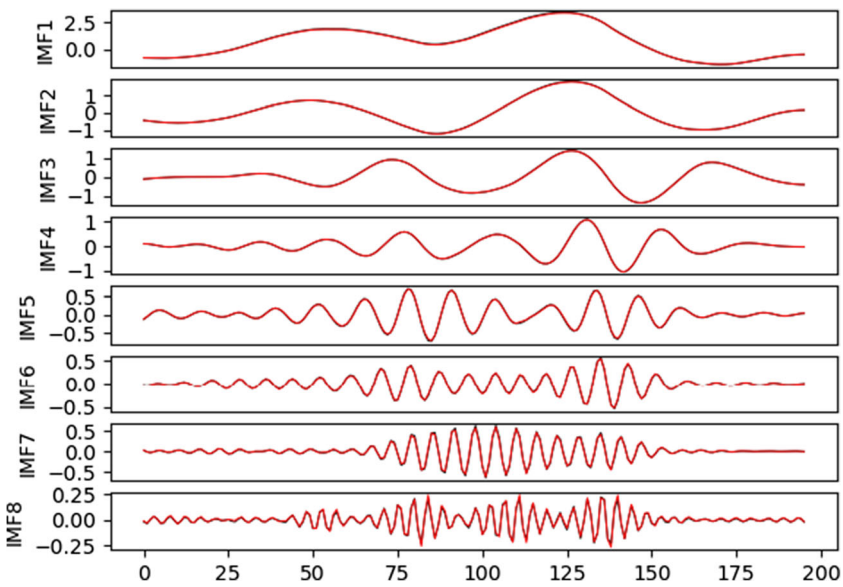

Fig. 5 Prediction curve of each IMF

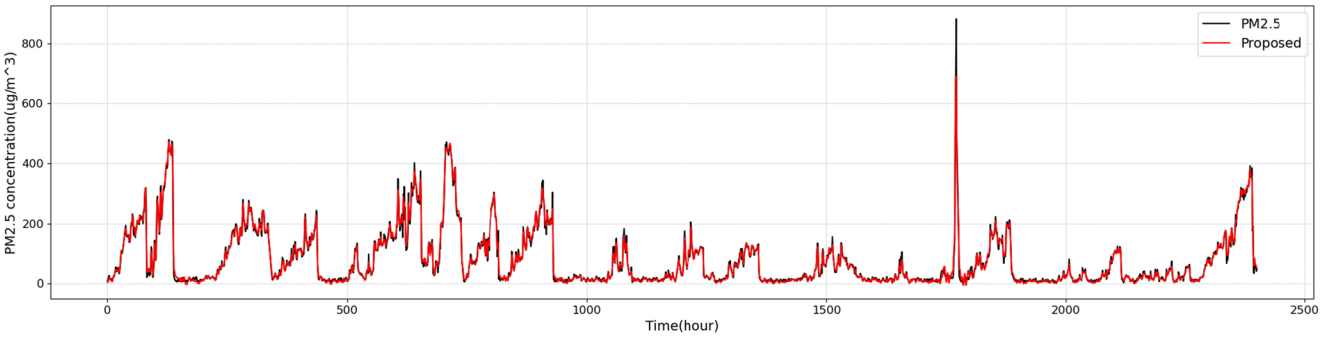

Fig. 6 The overall prediction performance of the VMD-BiLSTM model in Changping station

Evaluation metric

To evaluate the models' forecasting performance and justify the effectiveness of the proposed method, five evaluation metrics are employed in the experiment. The evaluation metrics include mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), coefficient of determination (R^2) and forecasting trend accuracy (ACC). The calculation formula for each evaluation metrics are

$$\mathrm{MAE}\left(y,\widehat{y}\right)=\frac{1}{n}\sum_{i=1}^{n}\left|y_{\mathrm{i}}-\widehat{y_{\mathrm{i}}}\right| \tag{14}$$

$$\mathrm{RMSE}\left(y,\widehat{y}\right)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|y_{\mathrm{i}}-\widehat{y_{\mathrm{i}}}\right|\right)^2} \tag{15}$$

$$\mathrm{MAPE}\left(y,\widehat{y}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_{\mathrm{i}}-\widehat{y_{\mathrm{i}}}\right|}{y_{\mathrm{i}}} \tag{16}$$

$$R^2\left(y,\widehat{y}\right)=1-\frac{\sum_{i=1}^{n}\left(y_{\mathrm{i}}-\widehat{y_{\mathrm{i}}}\right)^2}{\sum_{i=1}^{n}\left(y_{\mathrm{i}}-\overline{y}\right)^2} \tag{17}$$

$$\mathrm{ACC}=\frac{i_{\mathrm{correct}}}{i_{\mathrm{all}}}\times 100 \tag{18}$$

where y is the true value, $\widehat{y}$ is the predicted value, $\overline{y}$ is the average of the true values, n is the total number of samples. In (22), i_{correct} is the number of times samples that the PM2.5 increase/decay trend is correct, and i_{all} is the total number of times samples, which is the length of the test set.

MAE, RMSE and MAPE are employed to evaluate the error level of the prediction results. The smaller the value, the more accurate the prediction effect. R^2 is used to assess the degree of fit of the prediction result to the overall original data. The larger the value, the higher the degree of data fit and the better prediction effect. ACC is employed to evaluate the model's forecasting accuracy of trends in the short term.

Table 2 The evaluation of forecasting performance at Changping Station

Algorithms	MAE	RMSE%	MAPE	R^2	ACC
Decision tree	18.231	38.777	41.070	0.828	47.353
Random forest	13.642	28.678	31.331	0.906	48.770
SVR	12.403	29.858	29.179	0.898	48.479
MLP	12.026	26.251	27.702	0.921	48.187
LSTM	13.613	27.701	31.750	0.912	48.020
NLSTM	12.774	26.959	37.866	0.917	47.812
BiLSTM	12.058	25.469	26.932	0.926	48.312
EMD+LSTM	9.676	18.794	25.756	0.960	68.862
EMD+NLSTM	9.045	18.632	23.869	0.960	69.737
EMD+ BiLSTM	9.383	17.751	26.312	0.964	69.737
VMD+LSTM	5.792	11.089	18.185	0.986	74.364
VMD+NLSTM	5.212	9.436	16.845	0.990	75.490
Proposed	5.359	9.398	16.408	0.992	76.365

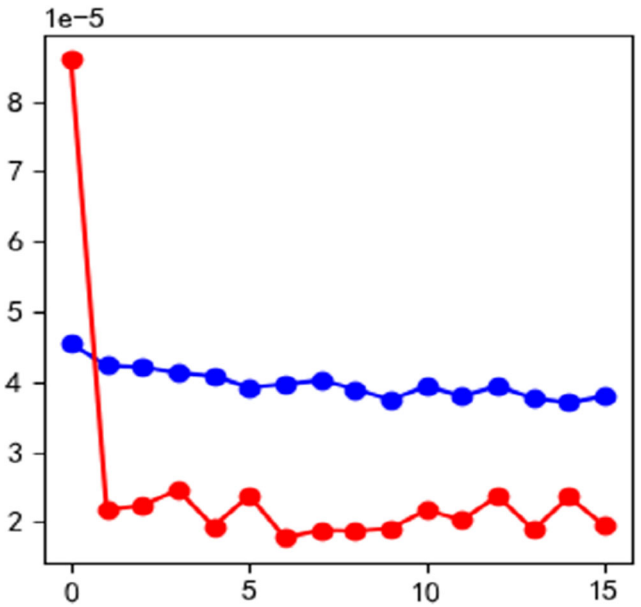

Fig. 7 The training (red) and validation (blue) loss curves of the VMD-BiLSTM model

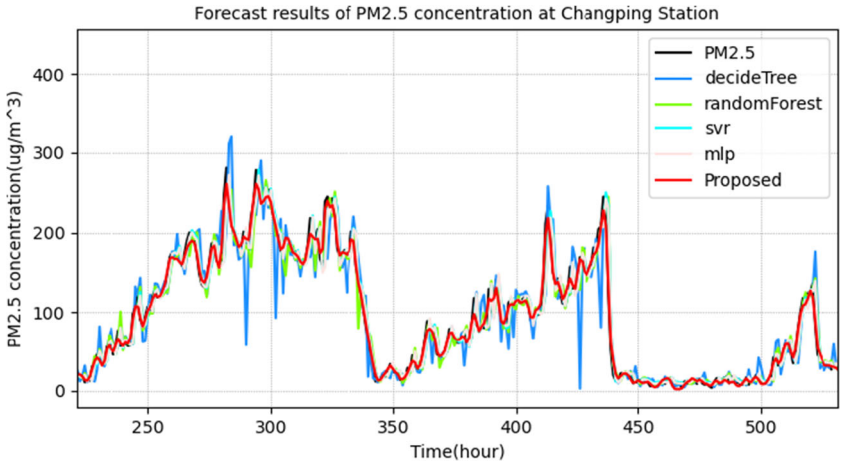

Fig. 8 The forecasting performance of VMD-BiLSTM compared with traditional machine learning algorithms

Experimental process and results

The overall flowchart of the proposed is shown in Fig. 4. The experiment process is divided into three stages: data preprocessing, decomposition prediction, and result analysis. In the first stage, the original data is standardized and divided. In the second stage, the univariate PM2.5 data is decomposed into K IMFs, then apply the BiLSTM model to learn and predict the decomposed data components. In the last stage, all outputs from BiLSTMs are concatenated and compared with the actual $PM_{2.5}$ data. All K BiLSTMs have the same internal structure. The hyper-parameters in the actual implementation of each BiLSTM include epoch=16, dropout=0.1, output_dim=64, activation='linear', validation_split=0.05, learning rate=0.001 (static), decay=0. All hyper-parameters are tuned for optimized performance during the experimental process. The mean square error (MSE) is used as the loss function. The root mean square propagation (RMSProp) optimizer is chosen to train the model. The RMSProp optimizer is selected maximally avoiding the swing effect during the gradient descent process and consequently accelerating the convergence process.

The value of K is first determined, which decides the number of IMF decomposition for the original time series data. The five evaluation metrics and the time of training are applied to evaluate the forecasting performance of VMD-BiLSTM with different values of K. the results are shown in Table 1.

According to Table 1, with the increment of K's value, the prediction performance is continuously improved with the trend been saturated. Additionally, with the increment of K, the consumption of calculation and time again rises. To ensure that the ideal prediction effect and reduce the calculation cost, the decomposition number K is finally chosen to be $K = 8$ in the experiment.

The prediction performance and fitting effect of each IMF are shown in Fig. 5. According to Fig. 5, the IMF8 curve fluctuates sharply, and the amplitude is relatively large. Such fluctuation leads to unsatisfactory forecasting results, which is also one of the main factors affecting forecasting performance. After the sample is decomposed, the range of changes is reduced, conducive to improving the prediction accuracy. However, due to the superposition and accumulation of the errors of eight independent forecasting processes, large errors

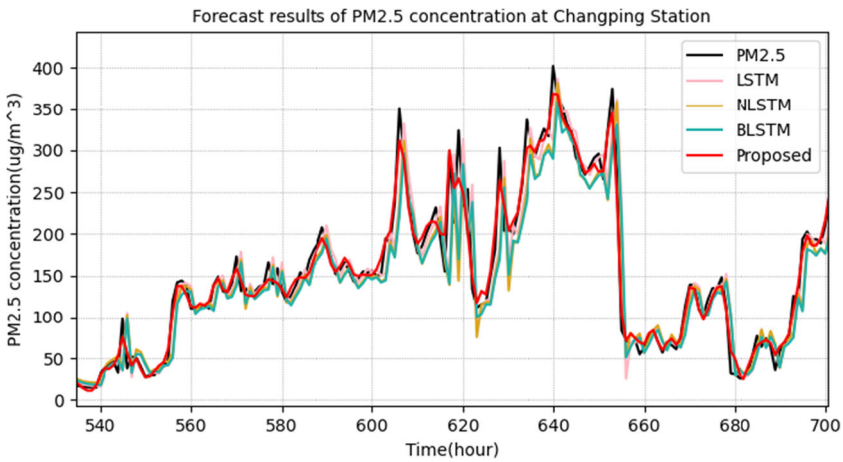

Fig. 9 The forecasting performance of VMD-BiLSTM compared with LSTM and its extended algorithms

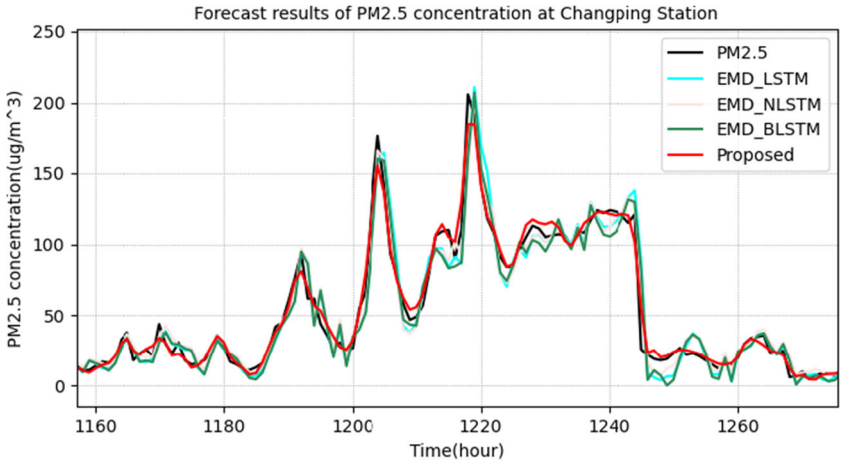

Fig. 10 The forecasting performance of VMD-BiLSTM compared with hybrid models combining LSTM and EMD

are still observed in a specific interval. A comparison of the final prediction result with the actual PM2.5 data is demonstrated in Fig. 6. According to it, at the peak of the 1765–1780 interval, the model achieved an unsatisfactory prediction effect.

To further justify the proposed model's effectiveness, several existing models for time series forecasting are employed to the comparison. The compared models include traditional machine learning models, such as the decision tree, random forest and support vector regression (SVR), LSTM models, including the conventional LSTM, nested LSTM (NLSTM) and BiLSTM, hybrid models combining LSTM with empirical mode decomposition (EMD), or VMD. The evaluation results of the models' forecasting performance are recorded in Table 2. The training and validation loss curves of the proposed VMD-BiLSTM model are shown in Fig. 7, which show no symptoms of over-fitting or under-fitting.

The forecasting performance of VMD-BiLSTM compared with traditional machine learning algorithms are shown in Fig. 8. The compared standard machine learning algorithms include decision tree, random forest, support vector machine regression (SVR), multi-layer perceptron (MLP). Experiment results show that the machine learning method cannot predict the actual value of the sample well. The machine learning algorithm's prediction in the interval where the data sample fluctuates frequently has a large deviation from the actual value. The VMD-BiLSTM prediction curve is more in line with the real value, and the prediction results are better than traditional machine learning models.

The forecasting performance of VMD-BiLSTM compared with LSTM and its extended algorithms are shown in Fig. 9. The extensions of LSTM model include nested LSTM (NLSTM) and bidirectional LSTM (BiLSTM). According to the experiment result, there is a strong lag on the prediction curve of LSTM models without signal decomposition. Therefore, the prediction results are not ideal, and the MAPE values are all around 30, while the trend prediction accuracy is less than 50%.

The forecasting performance of VMD-BiLSTM compared with hybrid models combining LSTM and empirical mode decomposition (EMD) are shown in Fig. 10. The hybrid models include EMD-LSTM, EMD-NLSTM and EMD-BiLSTM. According to Fig. 9, the LSTM expansion

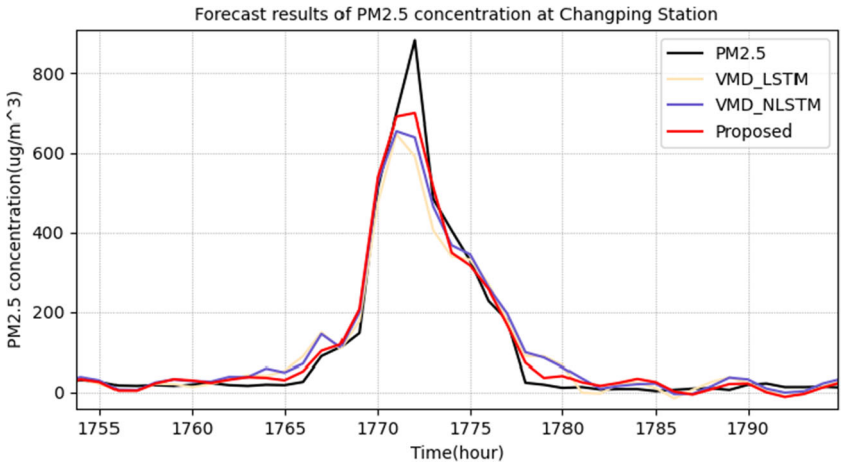

Fig. 11 The forecasting performance of VMD-BiLSTM compared with hybrid models combining LSTM and VMD

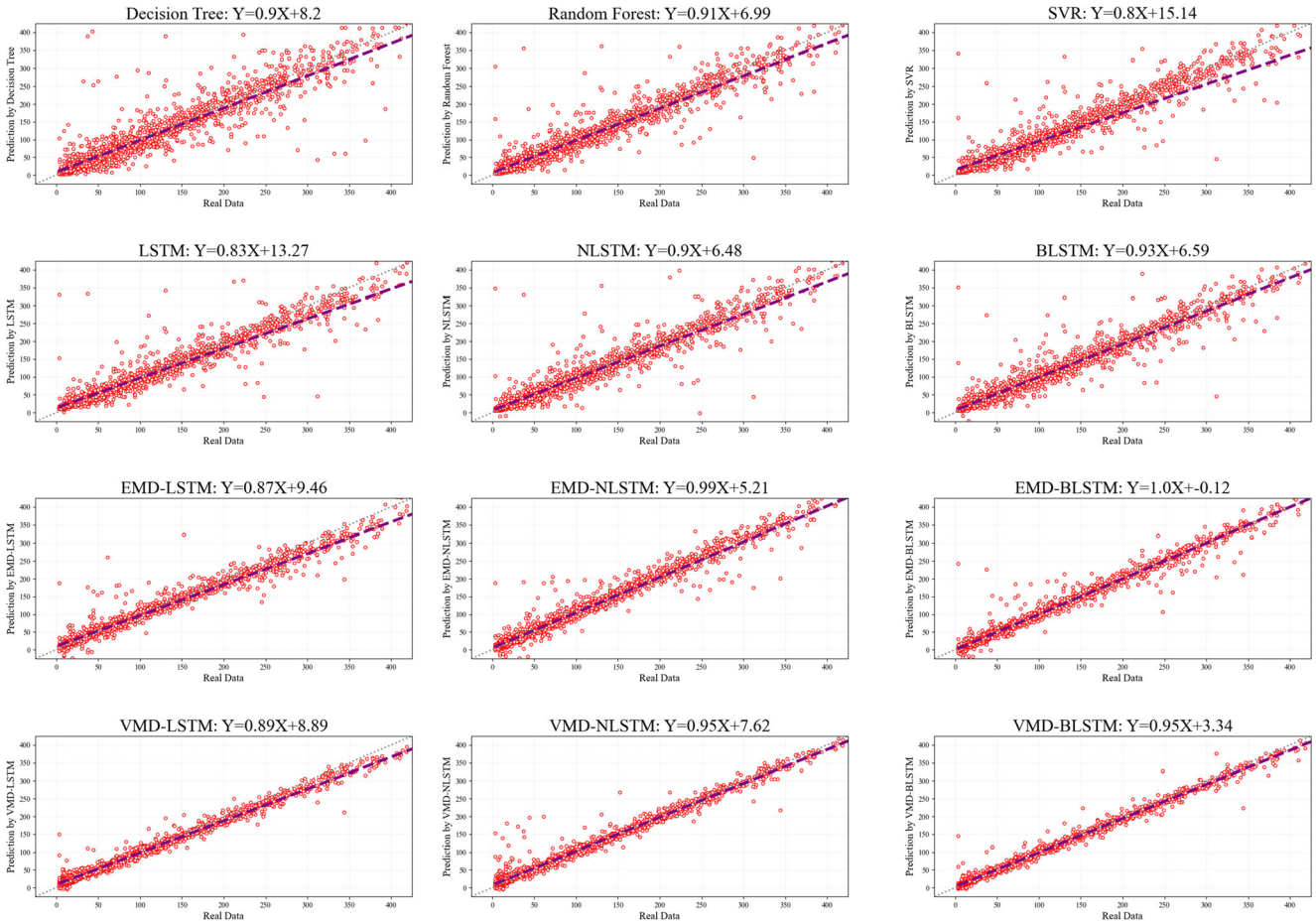

Fig. 12 Scatter plot of predicted $PM_{2.5}$ and true value in the model training phase

algorithm combined with EMD decomposition has an ideal result in predicting the peak. However, at the beginning and end of the peak, EMD results in severe delays and fluctuations. This drawback leads to the fact that the LSTM extension algorithm combined with EMD cannot have ideal prediction results. However, the VMD-BiLSTM model generates no signal delay phenomenon, and it is better than the models combining with EMD in each evaluation result.

The forecasting performance of VMD-BiLSTM compared with hybrid models combining VMD and other LSTM models are shown in Fig. 11; the compared models include VMD-LSTM and VMD-NLSTM. According to the result, the three algorithms show little difference in the smoothing interval. When there is a wave crest, VMD-BiLSTM has higher prediction accuracy in the beginning and end periods of the range, and the prediction results are closer to the actual value of the range.

In order to further study the VMD-BiLSTM model, we use scatter plots to visualize the prediction results and combine the absolute prediction error box plot and the relative prediction

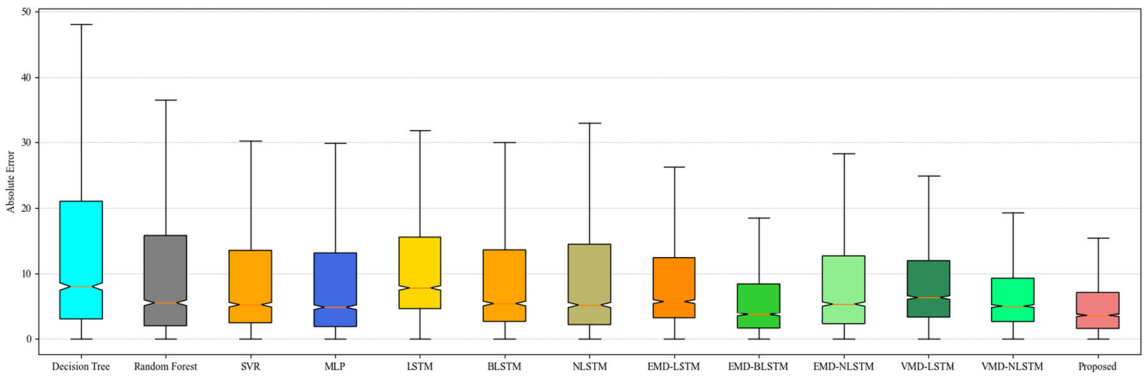

Fig. 13 Absolute prediction error (%) of predicted PM2.5 and true value in the testing phase

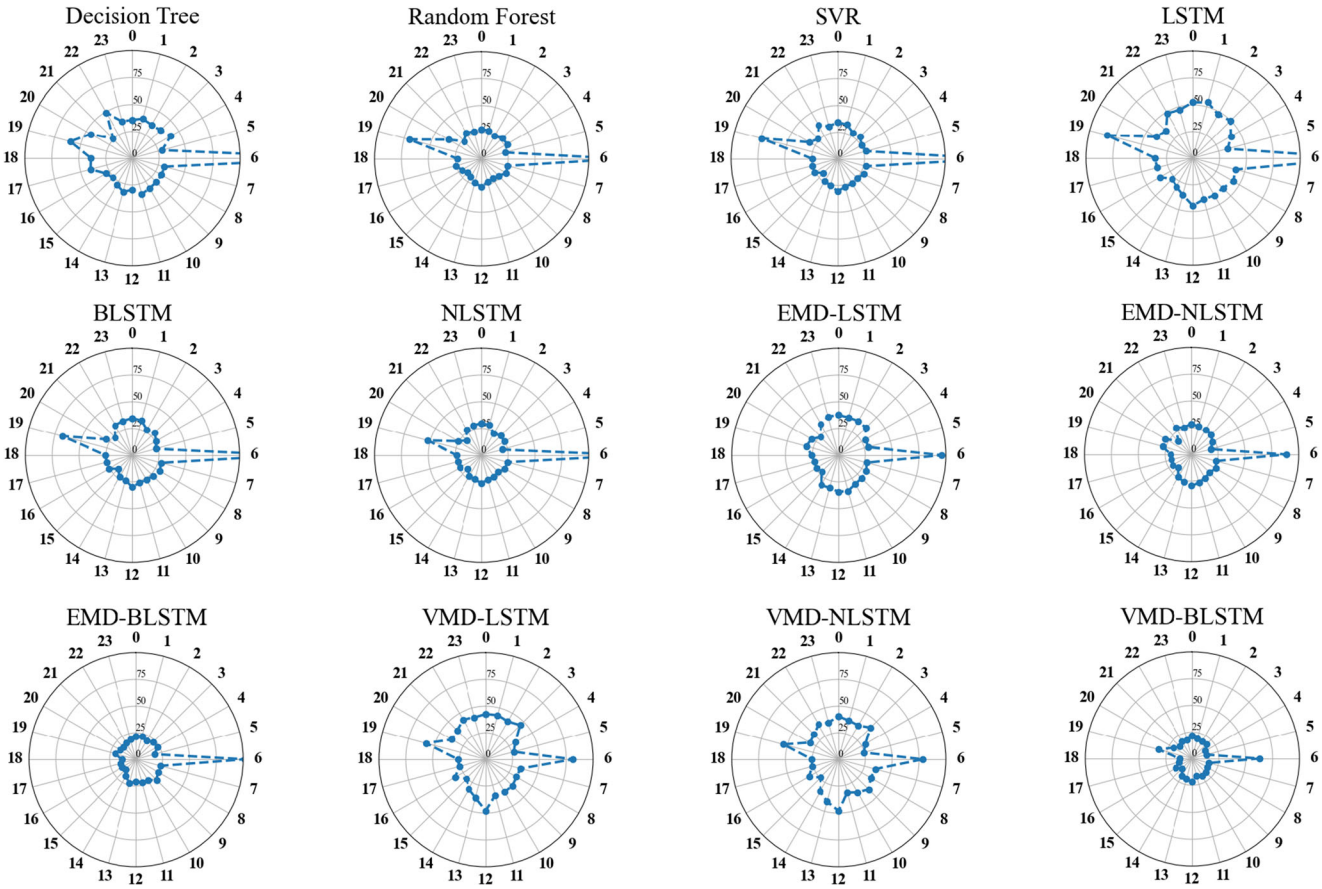

Fig. 14 The relative error (%) of the PM2.5 predicted value between VMD-BiLSTM and other models, where each circle from the origin represents a relative prediction error of 25%

error polar coordinate plot to comprehensively evaluate the experimental result

Figure 12 scatter plot can be more intuitive to see the prediction of the model. The dotted line represents the true value, and the dashed line represents the straight line fitted according to the predicted point. The closer to the $y = x$ axis, the better the prediction effect of the model. Among them, $y = wx + b$ represents linear fitting (w represents straight line gradient, b represents intercept) to summarize the accuracy of the model.

It is seen from the Fig. 12 that the training results of the traditional machine learning methods are not satisfactory with large deviations; the prediction results using time series model are improved and further improved by combining with EMD decomposition. The most stable and optimal results are obtained by combining the LSTM extensions with VMD

Figure 13 shows a box plot of the absolute prediction error between the predicted value and the true value. The upper and lower two short solid lines in the figure represent the extreme absolute prediction errors (0–100%) of each model. The red line in the middle represents 50% of the absolute prediction error of the model. The bottom of the box chart represents 25% absolute error, and the upper bottom represents 75% absolute error. Compared with other models, VMD-BiLSTM has the smallest box plot distribution, which means that the model has the highest prediction accuracy and the smallest absolute prediction error.

To further confirm the prediction effect of the VMD-BiLSTM model, this paper draws a polar coordinate diagram of the relative prediction error for the above experiment (Fig. 14). Theoretically, the closer the prediction error curve is to

Table 3 Additional PM2.5 data sets collected by selected three observing stations

Index	Dataset	Total number	Training set	Training number	Test set	Test number
Dingling	2013/3/1–2014/4/9	9720	2013/3/1–2014/4/4	9600	2014/4/5–2014/4/9	120
Dongsi	2013/3/1–2014/4/14	9840	2013/3/1–2014/4/9	9720	2014/4/10–2014/4/14	120
Gucheng	2013/3/1–2014/4/9	9720	2013/3/1–2014/4/4	9600	2014/4/5–2014/4/9	120

Table 4 Results and evaluation of Dingling Station PM2.5 data by various models.

Algorithms	MAE	RMSE%	MAPE	R^2	ACC
Decision tree	17.082	25.742	21.608	0.810	47.826
Random forest	11.353	15.320	16.759	0.933	46.957
SVR	8.784	12.773	13.169	0.953	53.913
MLP	9.030	13.335	13.435	0.949	50.435
LSTM	8.772	12.688	15.206	0.954	53.913
NLSTM	8.915	13.162	15.647	0.950	53.913
BiLSTM	8.704	12.811	14.701	0.953	53.913
EMD+LSTM	9.378	12.220	14.634	0.957	68.696
EMD+NLSTM	11.896	13.585	21.742	0.947	73.043
EMD+ BiLSTM	4.987	7.852	7.181	0.982	67.826
VMD+LSTM	4.138	6.693	5.389	0.991	71.304
VMD+NLSTM	4.340	5.893	5.674	0.990	73.913
Proposed	3.562	5.232	5.427	0.992	77.391

Table 6 Results and evaluation of Gucheng Station PM2.5 data by various models

Algorithms	MAE	RMSE%	MAPE	R^2	ACC
Decision tree	16.155	22.937	21.406	0.846	44.348
Random forest	10.492	15.489	16.184	0.930	47.826
SVR	9.059	13.892	14.881	0.943	47.826
MLP	8.989	13.689	13.179	0.945	49.565
LSTM	9.652	14.029	15.698	0.942	47.826
NLSTM	9.413	14.266	13.578	0.940	48.696
BiLSTM	9.716	14.204	15.174	0.941	47.826
EMD+LSTM	5.892	7.692	10.319	0.983	60.000
EMD+NLSTM	4.877	6.393	8.080	0.988	66.957
EMD+ BiLSTM	4.768	6.497	7.581	0.988	66.957
VMD+LSTM	4.453	5.947	10.185	0.990	68.696
VMD+NLSTM	4.029	5.697	7.523	0.990	69.565
Proposed	3.481	5.121	6.185	0.992	73.043

the origin, the better the prediction effect of the model. The above figure compares the predicted value with the real value (00.00–23.59) based on the day's 24 h. It is seen that the prediction effect of each model is the worst for the two-time points at 6 o'clock and 19 o'clock, but the prediction curve of VMD-BiLSTM is most concentrated at the origin, which means that the prediction effect of the model is the best.

To better demonstrate the model's generalization and avoid randomness, the data of three more observatories are utilized to conduct a further test on the proposed model. The basic information of the data is shown in Table 3. The comparative forecasting results are illustrated in Figs. 15, 16 and 17 and evaluated in Tables 4, 5 and 6 using the five evaluation metrics. The results show that the proposed VMD-BiLSTM is superior to other compared forecasting methods, specifically manifested in the small prediction errors, high fit with real data, high trend prediction accuracy, tall model stability and strong generalization.

Table 5 Results and evaluation of Dongsi Station PM2.5 data by various models

Algorithms	MAE	RMSE%	MAPE	R^2	ACC
Decision tree	28.159	38.852	27.745	0.713	41.739
Random forest	18.678	27.243	17.226	0.859	45.217
SVR	16.773	25.302	15.440	0.878	47.826
MLP	16.853	25.349	15.279	0.878	48.696
LSTM	16.679	25.127	15.304	0.880	46.957
NLSTM	17.038	25.528	15.537	0.876	45.217
BiLSTM	17.992	26.462	15.472	0.867	46.957
EMD+LSTM	16.456	36.092	20.423	0.752	64.348
EMD+NLSTM	12.464	25.699	15.268	0.875	69.565
EMD+ BiLSTM	15.279	33.071	19.355	0.792	66.957
VMD+LSTM	7.879	10.222	8.625	0.980	69.565
VMD+NLSTM	7.812	9.263	6.955	0.984	70.235
Proposed	7.124	9.295	6.368	0.981	70.435

Conclusion

Considering the complex characteristics of non-linear, non-periodic and non-stationary $PM_{2.5}$ concentration data, a hybrid neural network VMD-BiLSTM model is proposed to perform accurate hour-ahead time series forecasting.

First, a cutting-edge data decomposition method VMD is employed to decompose the original $PM_{2.5}$ time series data into a sequence of IMFs according to the frequency. Secondly, BiLSTM neural network is implemented to construct training and prediction models for each IMF component. Finally, the results of the prediction sub-sequences are combined to obtain the final forecasting result.

Compared with other time series forecasting models, the proposed VMD-BiLSTM method outputs quality forecasting results by combining forward and backward data features in the LSTM neural network. Compared with EMD decomposition, VMD effectively avoids the delay phenomenon of prediction.

In this article, $PM_{2.5}$ data from four observation stations in China are adopted for verification purposes. Traditional machine learning methods, LSTM models and hybrid models combining LSTMs and decomposition methods are employed in a comprehensive comparative study and justify the

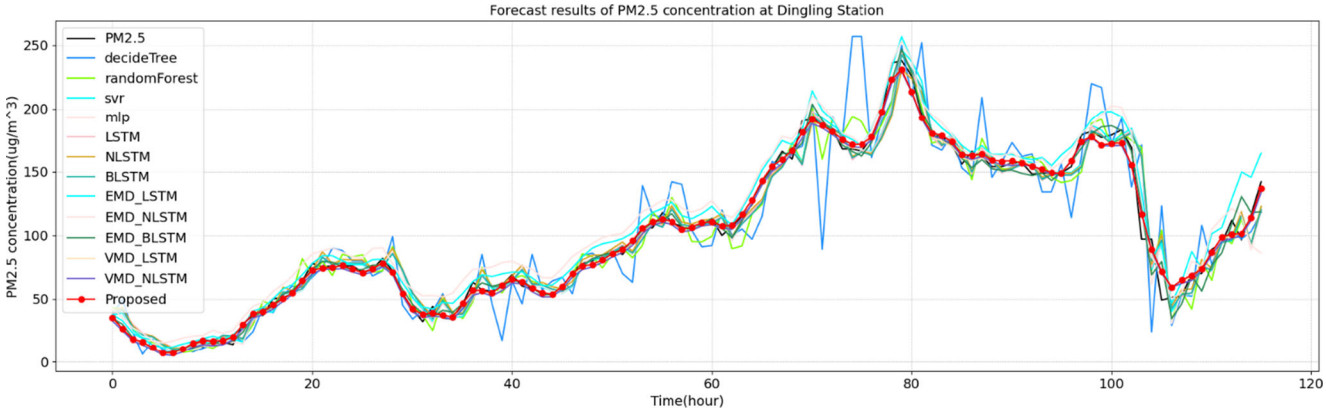

Fig. 15 Prediction results of Dingling Station PM2.5 data by various models

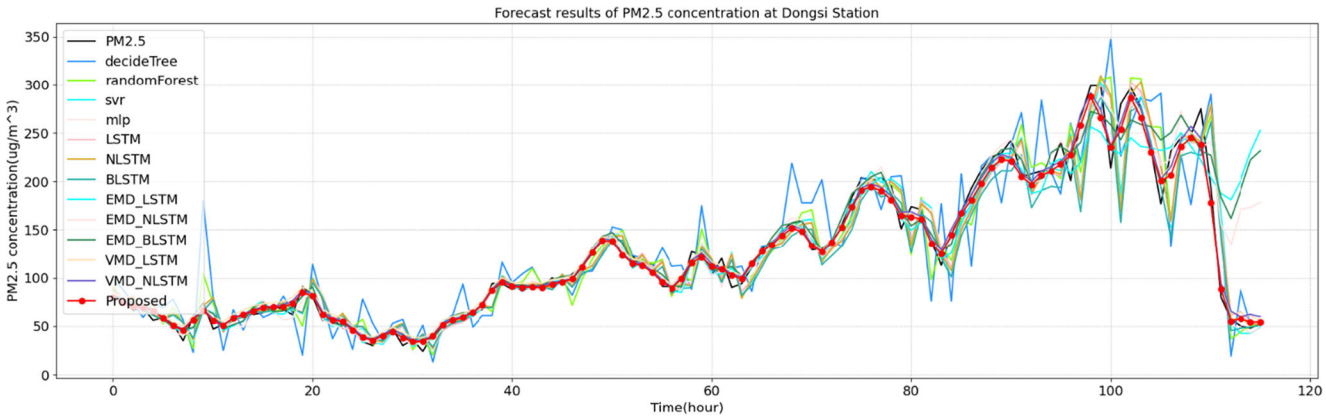

Fig. 16 Prediction results of Dongsi Station PM2.5 data by various models

forecasting quality of the proposed model. Experimental results show that BiLSTM is more suitable to be combined with VMD decomposition compared with other decomposition methods. The proposed model outperforms the existing models in terms of prediction error and trend prediction accuracy.

The next step of the research is to extend the proposed model to other forecasting fields, including photovoltaic forecasting and household electricity forecasting, which also show strong nonlinearity and non-stationarity. Additionally, further study of the multi-channel model for air quality prediction will be conducted by comprehensively considering air quality data series volatility.

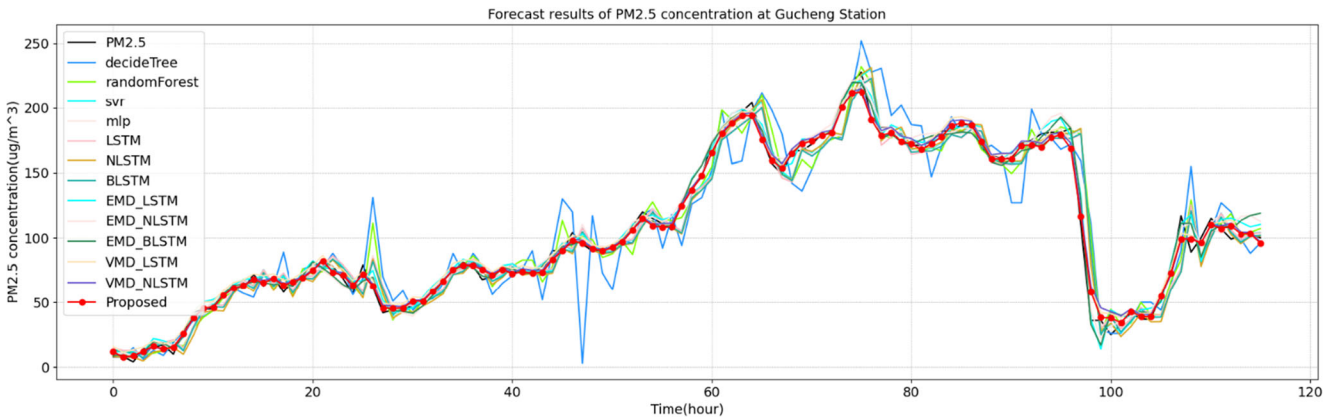

Fig. 17 Prediction results of Gucheng Station PM2.5 data by various models

Acknowledgements Authors from the original publication of the dataset (Zhang et al. 2017) are appreciated.

Author contribution Conceptualization, K.Y.; methodology, K.Y.; software, Z.Z.; validation, Y.Z.; formal analysis, Z.Z.; investigation, K.Y.; resources, K.Y.; original draft preparation, K.Y.; writing, review and editing, K.Y.; visualization, Y.Z.; supervision, K.Y.; project administration, K.Y.; funding acquisition, K.Y.

Funding This work was supported by the Ministry of Education (MOE) Singapore, Tier 1 Grant for National University of Singapore (NUS) under grant number R296000208133.

Data availability The data used in this paper is publicly available at UCI machine learning knowledge base. URL: http://archive.ics.uci.edu/ml/

Declarations

Ethics approval and consent to participate Ethics Committee approval was received from the research ethics committees in the College of Information Engineering of China Jiliang University and National University of Singapore.

Consent for publication All authors declare that they have the consent regarding the publication of this manuscript

Competing interests The authors declare that they have no competing interest.

References

Chan CK, Yao X (2008) Air pollution in mega cities in China. Atmos Environ 42(1):1–42

Chang Z, Zhang Y, Chen W (2019) Electricity price prediction based on hybrid model of Adam optimized LSTM neural network and wavelet transform. Energy 187:115804

Chang YS, Chiao HT, Abimannan S, Huang YP, Tsai YT, Lin KM (2020) An LSTM-based aggregated model for air pollution forecasting. Atmos Pollut Res 11(8):1451–1463

Ding G, Qin L (2019) Study on the prediction of stock price based on the associated network model of LSTM. Int J Mach Learn Cybern 11: 1307–1317

Ding Y, Zhang M, Chen S, Wang W, Nie R (2019) The environmental Kuznets curve for PM2.5 pollution in Beijing-Tianjin-Hebei region of China: a spatial panel data approach. J Clean Prod 220:984–994

Gendeel M, Yuxian Z, Aoqi H (2018) Performance comparison of ANNs model with VMD for short-term wind speed forecasting. IET Renew Power Gener 12:1424–1430

Gers FA, Schmidhuber J, Cummins F (2000) Learning to forget: continual prediction with lstm. Neural Comput 12(10):2451–2471

Han L, Zhang R, Wang X, Bao A, Jing H (2019) Multi-step wind power forecast based on VMD-LSTM. IET Renew Power Gener 13:1690–1700

Hao Y, Luo B, Simayi M, Zhang W, Jiang Y, He J, Xie S (2020) Spatiotemporal patterns of PM2.5 elemental composition over China and associated health risks. Environ Pollut 265:114910

Hochreiter S, Schmidhuber J (1997) Long short term memory. Neural Comput 9(8):1735–1780

Jin N, Wu J, Ma X, Yan K, Mo Y (2020) Multi-task learning model based on multi-scale CNN and LSTM for sentiment classification. IEEE Access 8:1-1. https://doi.org/10.1109/access.2020.2989428

Li X, Jin L, Kan H (2019a) Air pollution: a global problem needs local fixes. Nature 570(7762):437–439

Li J, Zhu Y, Kelly JT, Jang CJ, Wang S, Hanna A et al (2019b) Health benefit assessment of PM2.5 reduction in Pearl River Delta region of China using a model-monitor data fusion approach. J Environ Manag 233:489–498

Li T, Hua M, Wu X (2020) A hybrid CNN-LSTM model for forecasting particulate matter (PM2.5). IEEE Access 8:26933–26940

Liang D, Xu J, Li S, Sun C (2020) Short-term passenger flow prediction of rail transit based on VMD-LSTM neural network combination model. In: 2020 Chinese Control And Decision Conference (CCDC). IEEE, Hefei, pp 5131–5136

Moniz JRA, Krueger D (2018) Nested lstms. arXiv preprint arXiv: 1801.10308.

Morillas H, Marcaida I, Maguregui M, Upasen S, Gallego-Cartagena E, Madariaga JM (2019) Identification of metals and metalloids as hazardous elements in PM2.5 and PM10 collected in a coastal environment affected by diffuse contamination. J Clean Prod 226:369–378

O'Donnell MJ, Fang J, Mittleman MA, Kapral MK, Wellenius GA (2011) Fine particulate air pollution (pm2.5) and the risk of acute ischemic stroke. Epidemiology 22(3):422–431

Song X, Huang J, Song D (2019) Air quality prediction based on LSTM-Kalman model. In: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, Chongqing, pp 695–699

Wang H, Zhang S (2020) Prediction of daily PM2.5 concentration in China using data-driven ordinary differential equations. Appl Math Comput 375:125088

Wang J, Li J, Wang X, Wang J, Huang M (2020) Air quality prediction using CT-LSTM. Neural Comput & Applic:1–14

Wu Q, Lin H (2019a) Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. Sustain Cities Soc 50:101657

Wu Q, Lin H (2019b) A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Sci Total Environ 683:808–821

Wu L, Luo XS, Li H, Cang L, Yang J, Yang J et al (2019) Seasonal levels, sources, and health risks of heavy metals in atmospheric PM2.5 from four functional areas of Nanjing City, Eastern China. Atmosphere 10(7)

Wu X, Li J, Jin Y, Zheng S (2020) Modeling and analysis of tool wear prediction based on SVD and BiLSTM. Int J Adv Manuf Technol 106(9):4391–4399

Xie Y, Liang R, Liang Z, Huang C, Zou C, Schuller B (2019) Speech emotion classification using attention-based lstm. IEEE/ACM IEEE Trans Audio Speech Lang Process 27(11):1675–1685

Xu X, Yoneda M (2019) Multitask air-quality prediction based on LSTM-autoencoder model. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2019.2945999

Yan K, Wang X, Du Y, Jin N, Huang H, Zhou H (2018) Multi-step short-term power consumption forecasting with a hybrid deep learning strategy. Energies 11(11):3089

Yan K, Li W, Ji Z, Qi M, Du Y (2019) A hybrid LSTM neural network for energy consumption forecasting of individual households. IEEE Access 7:157633–157642

Zhang S, Guo B, Dong A, He J, Xu Z, Chen SX (2017) Cautionary tales on air-quality improvement in Beijing. Proc R Soc A Math Phys Eng Sci 473(2205):20170457

Zhao CN, Xu Z, Wu GC, Mao YM, Liu LN, Qian W et al (2019) Emerging role of air pollution in autoimmune diseases. Autoimmun Rev 18:607–614

Zheng H, Yuan J, Chen L (2017) Short-term load forecasting using emd-lstm neural networks with a xgboost algorithm for feature importance evaluation. Energies 10(8):1168–1180

Zhou H, Zhang Y, Yang L, Liu Q, Yan K, Du Y (2019) Short-term photovoltaic power forecasting based on long short term memory neural network and attention mechanism. IEEE Access 7:78063–78074