Sustainable Environment Research

RESEARCH

Open Access

Improving air pollutant prediction in Henan Province, China, by enhancing the concentration prediction accuracy using autocorrelation errors and an Informer deep learning model

Kun Cai[1,2], Xusheng Zhang[1], Ming Zhang[2,3], Qiang Ge[1,4,5*], Shenshen Li[6], Baojun Qiao[1,4,5] and Yang Liu[1,5,7]

Abstract

Air pollution is an important issue affecting sustainable development in China, and accurate air quality prediction has become an important means of air pollution control. At present, traditional methods, such as deterministic and statistical approaches, have large prediction errors and cannot provide effective information to prevent the negative effects of air pollution. Therefore, few existing methods could obtain accurate air pollutant time series predictions. To this end, a deep learning-based air pollutant prediction method, namely, the autocorrelation error-Informer (AE-Informer) model, is proposed in this study. The model implements the AE based on the Informer model. The AE-Informer model is used to predict the hourly concentrations of multiple air pollutants, including PM_{10}, $PM_{2.5}$, NO_2, and O_3. The experimental results show that the mean absolute error (MAE) and root mean square error (RMSE) values of AE-Informer in multivariate prediction are 3% less than those of the Informer model; thus, the prediction error is effectively reduced. In addition, a stacking ensemble model is proposed to supplement the missing air pollutant time series data. This study uses Henan Province in China as an example to test the validity of the proposed methodology.

Keywords Air pollutant forecast, Missing time series data supplement, Deep learning, Informer, Autocorrelated Errors

*Correspondence:
Qiang Ge
gqhenu@126.com
[1] School of Computer and Information Engineering, Henan University, Kaifeng 475004, China
[2] Henan Key Laboratory of Spatial Information Application On Eco-Environmental Protection, Henan Environmental Monitoring Center, Zhengzhou 450007, China
[3] The 27Th Research Institute, China Electronics Technology Group Corporation, Zhengzhou 450007, China
[4] Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng 475004, China
[5] Henan Engineering Laboratory of Spatial Information Processing, Henan University, Kaifeng 475004, China
[6] State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China
[7] Shenzhen Research Institute, Henan University, Shenzhen 518000, China

1 Introduction

Air pollutants (PM_{10}, $PM_{2.5}$, O_3, NO_2, etc.) are important problems in ecological environments [1–3] that cause several issues, such as reduced air quality and human health risks [4]. The maximum 8-h 90th quantile concentration of ozone in cities such as Beijing, Tai'an, Zibo, Dezhou, Handan, and Kaifeng increased from 2015 to 2018, the annual concentration went up from 168 to 212 μg m^{-3} [5]. In recent years, public safety studies found that the levels of $PM_{2.5}$ and O_3 were closely related to cardiovascular, cerebrovascular, nervous system, and respiratory diseases [6], while long-term exposure to O_3 and NO_2 increased the risk of death [7]. In addition, air pollutants affect people's happiness, population

migration, and other livelihood issues [8]. Short-term air quality prediction can be used not only to reduce future high air pollution events but also to decrease resident exposure. Therefore, the real-time acquisition of air pollutant concentration information and the accurate prediction of future concentration information are of great significance for air pollution management and public health protection.

The most commonly used air pollutant concentration prediction methods are deterministic methods, statistical methods, and machine learning methods [9–11]. Deterministic methods predict the concentration of air pollutants by simulating atmospheric chemical diffusion and transport processes. Commonly used deterministic methods include chemical transport models [12], and operational street pollution models [13]. Although these methods can be used to generate pollutant predictions, they have considerable computational costs, and the prediction results may be inaccurate due to a lack of actual observation data [14, 15]. Statistical methods address the problem of limited data in deterministic methods. The most commonly used statistical methods include the autoregressive integrated moving average (ARIMA) method, geographically weighted regression method and generalized additive model [16–18]. These methods have been widely used for time series prediction of air pollutant levels. For example, Slini et al. [19] used the ARIMA model to predict ozone concentrations in Greece. However, most statistical methods assume linear relationships between variables and labels, which is inconsistent with real-world nonlinearities. To solve this problem, researchers applied nonlinear models in machine learning. For example, Ma et al. [20] used support vector machines to predict the concentrations of air pollutants such as PM_{10} and $PM_{2.5}$. Rubal et al. [21] used the random forest (RF) model to predict the future 1-h concentrations of seven pollutants, including NO_2. Although these models obtain improved prediction accuracy, they ignore time series trends in air pollutant concentrations.

With the rapid development of deep learning techniques, traditional machine learning and shallow neural network models no longer obtain state-of-the-art performance. Different kinds of deep learning models have been proposed to improve the air quality prediction performance. For example, Ma et al. [22] used a bidirectional long short-term memory neural network model (Bi-LSTM) based on the recurrent neural network (RNN) structure and transfer learning to predict future 1-h, 1-d and 1-wk concentrations of PM_{10}. Chauhan et al. [23] used the convolutional neural network (CNN) structure to predict the future 1-d concentrations of five pollutants, including PM_{10} and $PM_{2.5}$, in India. RNNs are limited in solving gradient problems. CNNs are limited in obtaining long-term historical information, and they cannot obtain accurate predictions of air pollutant concentrations. In the past two years, the transformer model [24] was introduced in the field of time series prediction, and its self-attention mechanism provides an effective method for obtaining long-term macroscopic information in time series. Many improved transformer based models have been proposed. For example, the LogTrans model [25] showed high accuracy in predicting future hourly electricity consumption and reduced the running cost of the model. The Star-Transformer model [26] improved the prediction performance of future hourly meteorological index. Additional models, such as the MetaFormer [27], AutoFormer [28], Transformer-XL [29], and Set Transformer [30] models, all exhibited considerable gains in time series prediction. The Informer model [31] is an improved transformer time series prediction model based on the Kullback–Leibler (KL) divergence that was proposed in 2021. The Informer model improves the time series prediction accuracy while reducing the running cost of the model, saving considerable time. This model showed improved performance for power consumption time series prediction and traffic flow time series prediction but has not been applied in air quality prediction. In this study, we apply the Informer model to air quality time series prediction and modify the method to further improve its prediction accuracy.

Due to factors such as human and machine failures, a large amount of data is lost in the acquired state control site data sets, resulting in discontinuous time states, which seriously affect subsequent data analyses. The issue of missing data in the time series needs to be addressed. In this paper, we use the stacking ensemble model to supplement the missing data in the air pollutant concentration time series and compare the effectiveness of the stacking ensemble model with existing approaches. Then, the Informer model is used to obtain air pollutant concentration time series predictions, and the performance of the proposed model is compared with that of other deep learning models. The AE-Informer model is proposed, which combines the AE strategy [32] with the Informer model to improve the prediction accuracy. To verify the framework of the proposed method, we model the levels of four major air pollutants, namely, PM_{10}, $PM_{2.5}$, NO_2, and O_3, in the study area in Henan Province.

2 Materials and methods

Figure 1 presents the methodological framework of the model proposed this paper. The framework has three parts: (1) air pollutant data collection and missing

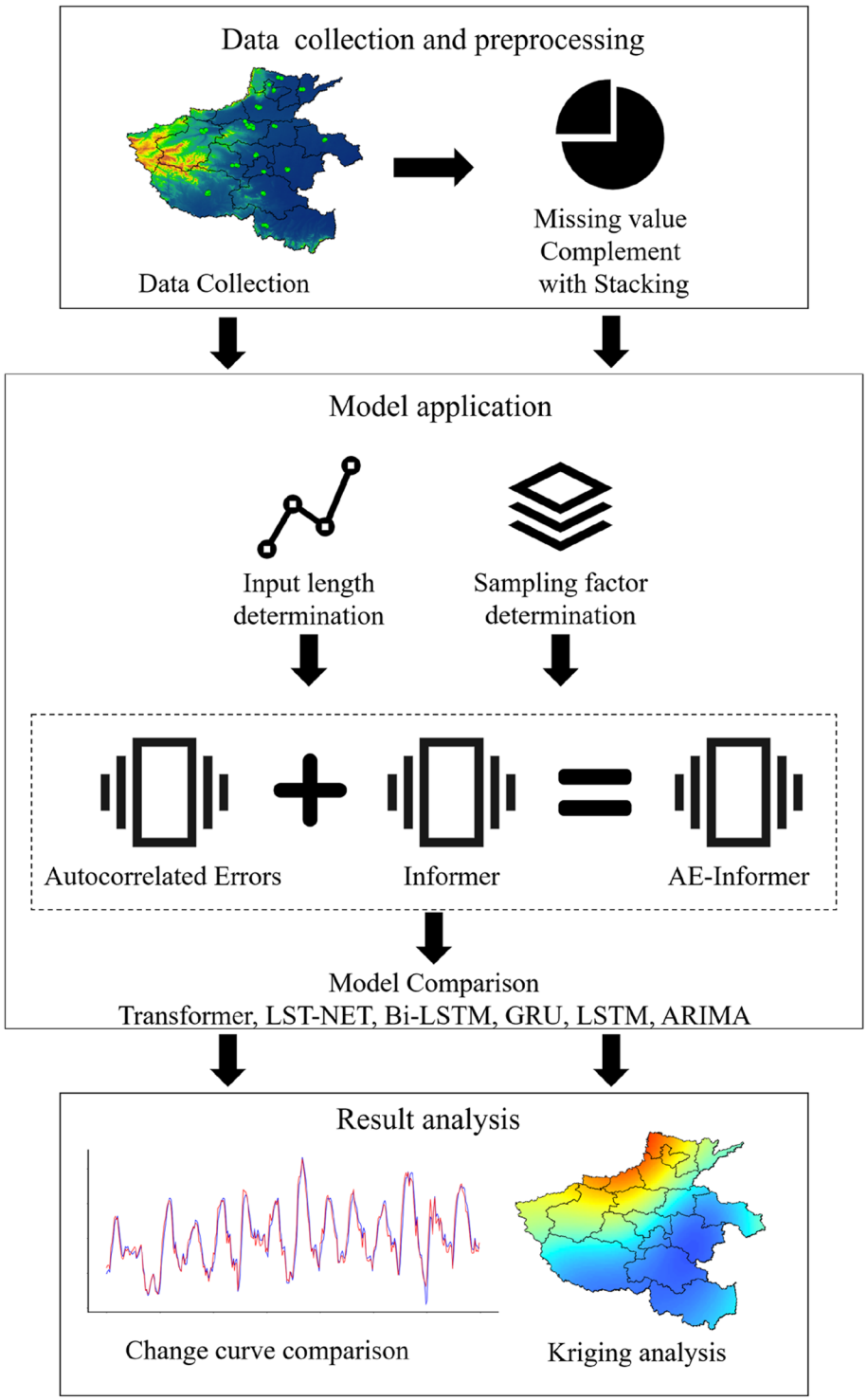

Fig. 1 Methodological framework (data, model application and result analysis)

value supplementation, (2) structural design of the AE-Informer model and the prediction of air pollutants, and (3) analysis of the prediction result and generalization tests.

2.1 Research area and data

As shown in Fig. 2, Henan Province is located at the junction of the coastal open areas and the central and western regions. It is the core area of China's economic and social

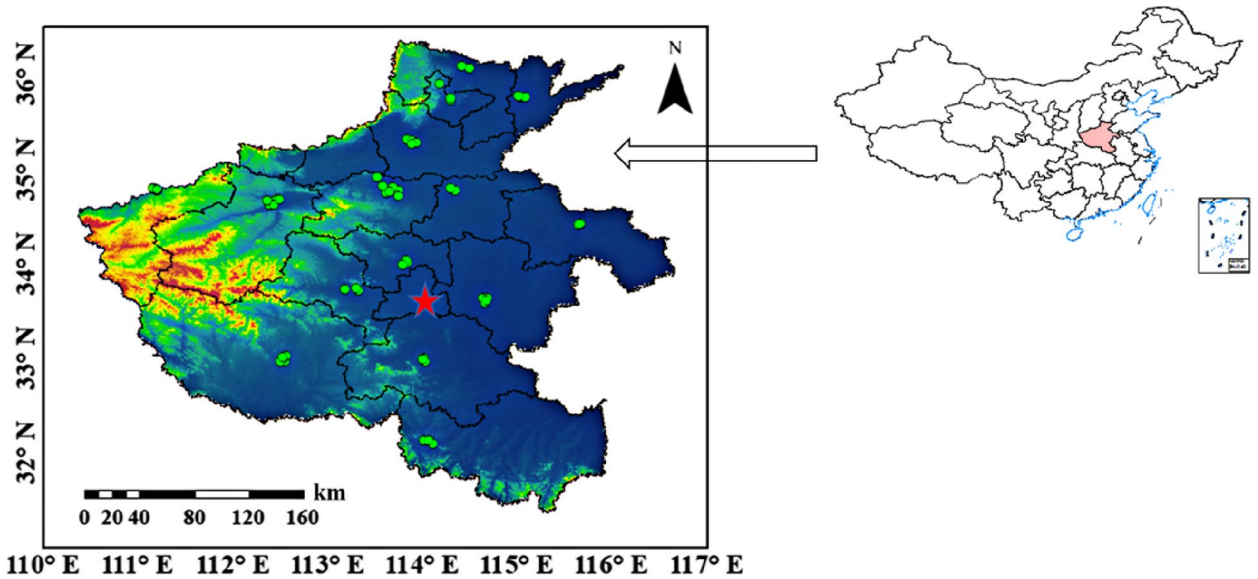

Fig. 2 Overview of Henan Province overlaid with ground-based stations

development, as well as one of the most densely populated and polluted areas in China [1, 33]. In this study, ground measurements of the PM_{10}, $PM_{2.5}$, NO_2 and O_3 mass concentrations were collected hourly from January 1, 2019, to December 31, 2020, at 60 stations in Henan Province by the China Environmental Monitoring Centre (CEMC). The green dots denote ground-based CEMC sites, and the red five-pointed star denotes the site used in the case study and experiments in this paper. We first removed invalid values and outliers due to instrument calibration issues. The collected data have missing values due to instrument damage, human error, and other factors. Therefore, we use the stacking ensemble learning model to fill in the missing values; more details are provided in Sect. 3.1.

2.2 Methods

2.2.1 Informer model

In air pollutant time series, the value at the current moment is correlated with the value at each moment in the previous period, and the air pollutant concentration at the current moment can be predicted based on the historical time series information. Informer [31] is an improved time series prediction model based on the Transformer. The Informer model has an encoder-decoder structure, and the core of this model is the self-attention mechanism. In contrast to models with RNN and CNN structures, models with self-attention mechanisms do not need to consider the position in the sequence when obtaining the historical time series information, and the cost of calculating the association between two positions in the time series does not increase with increasing distance. Therefore, historical information can be obtained more effectively to accurately predict the air pollutant concentration at the current moment. The calculation equation is shown in Eq. (1):

$$Attention(X, X, X) = softmax\left(XX^T\right)X \tag{1}$$

The calculation process is shown in Fig. 3. In this figure, q and k are sequences that are obtained by multiplying X by the weights W^q and W^k, respectively; these sequences are essentially the same as X. The inner product of q and k is equivalent to XX^T, which represents the inner product of the current moment and the value at each moment in the previous period. The result of the inner product is normalized by the softmax function to generate a new sequence $\widehat{\alpha}$. The larger the value at a certain position in the sequence is, the higher the correlation between the value at the moment to be predicted and the value at that position. When predicting the value at the current moment, more information about this moment is considered. Finally, the inner product of $\widehat{\alpha}$ and v is used to obtain the attention score, which is an internal representation of X in the model that represents various features of X.

In addition, the Informer model combines the self-attention mechanism with the KL divergence strategy to create ProbSparse self-attention. Since most of the historical information is provided by the values at a few positions in the time series, to reduce the computational costs, the positions that provide a large

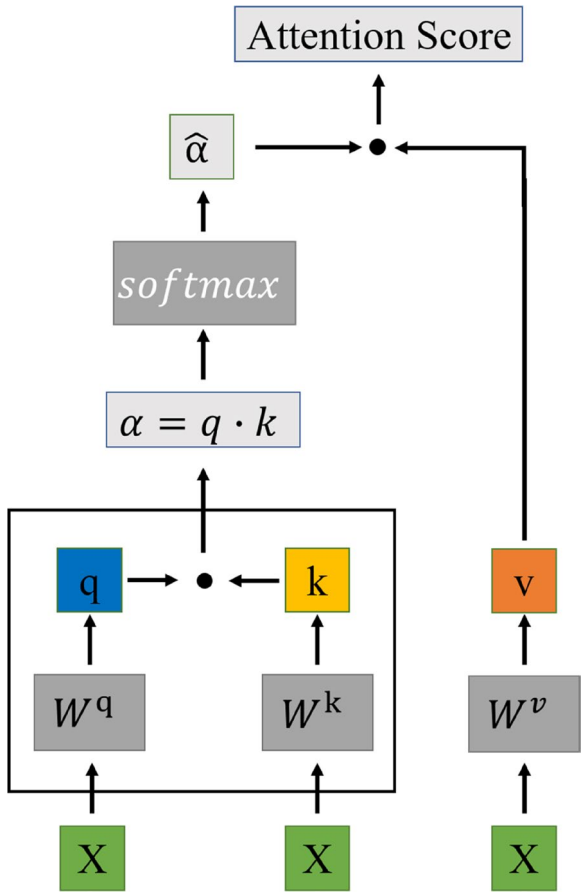

Fig. 3 Self-attention calculation process (X represents the input pollutant time series; q, k and v are sequences obtained by multiplying X by the weights W^q, W^k and W^v; α is the inner product of q and k; and $\widehat{\alpha}$ is α normalized by the softmax function)

amount of information are found according to the sparse scores at various positions, and dot product calculations are performed to obtain the available historical information. These dot product operations are not required at other locations. The calculation equation is shown in Eq. (2):

$$M(q_i, K) = \ln \sum_{j=1}^{L_k} e^{\frac{q_i k_j^T}{\sqrt{d}}} - \frac{1}{L_K} \sum_{j=1}^{L_k} \frac{q_i k_j^T}{\sqrt{d}} \tag{2}$$

where q_i is the value at the i-th position in the air pollutant time series X, K is the entire X sequence, and L_k is the length of X. Dividing by $\sqrt{d}$ ensures that the input to the softmax function is not too large, which would cause the partial derivative to approach 0. The larger the M value at the i-th position is, the more information this position carries, and the more important this position is in the self-attention operation.

2.2.2 AE-Informer model

Autocorrelated errors are introduced when insufficient covariates are added, data collection errors occur, and when the time series prediction model does not fully fit. To reduce the influence of these errors on the prediction results, AE [31] can be incorporated into the Informer model. Since the errors are autocorrelated, the current moment error can be represented by the errors at each moment in the previous period. The calculation equation is shown in Eq. (3):

$$e_t = \rho_1 e_{t-1} + \cdots + \rho_p e_{t-p} + \varepsilon_t, |\rho_i| < 1, \forall i \tag{3}$$

where ρ is the error parameter at each moment, e is the error at each moment, and ε_t is the error of the entire period. For the convenience of calculation, the equation is reduced to the first-order form, yielding $e_t = \rho_1 e_{t-1}$. Assuming that $\widehat{\varepsilon} = \widehat{X_t} - \widehat{\rho} X_{t-1}$, the new input and output of the model can be constructed by combining the two equations. The input changes from the observed value of the air pollutant concentration at each moment in the previous period to the error value at each moment, and the output changes from the predicted value at the current moment to the predicted value of the error at the current moment, where ρ is used as a parameter to train the model. Finally, by applying $\widehat{X_t} = \widehat{\varepsilon} + \widehat{\rho} X_{t-1}$, the predicted value of the error at the current moment is added to the observed value at the previous moment to obtain the predicted value at the current moment. This approach improves deep learning models such as LSTM; thus, this method was used in the Informer model to improve the accuracy of the air pollutant concentration predictions.

To improve the hourly prediction accuracy of the Informer model [31], in this study, we fuse the Informer model with AEs [32] (Fig. 4) and propose the AE-Informer model. Figure 4a shows the traditional Informer model, and 4b presents the modified AE-Informer model. When the air pollutant concentration $\widehat{X_t}$ is predicted at time t, the input to the Informer model is adjusted from the hourly pollutant concentration observations $\{X_{t-w}, \ldots, X_{t-1}\}$ to the hourly observations and the error values between these observations and those in the previous hour $\{X_{t-w} - X_{t-w}, X_{t-w-1}, \ldots, X_{t-1} - X_{t-2}\}$. The output changes from the predicted air pollutant concentration at the current time $\widehat{X_t}$ to the predicted value of the error between the current and previous time $\widehat{X_t} - \rho X_{t-1}$. Then, ρX_{t-1} is added to the prediction result to obtain the final prediction value $\widehat{X_t}$ at time t. ρ is a parameter of the error between each moment and the previous moment that is added to the Informer model. Finally, the model is trained and iterated.

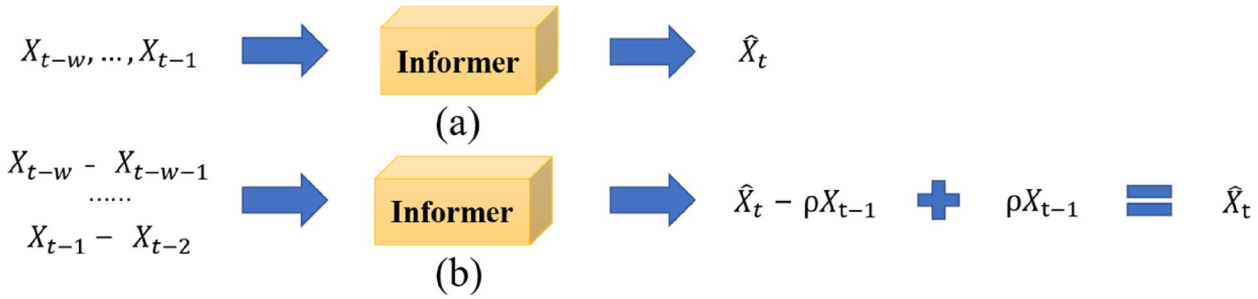

Fig. 4 Improvement of the Informer model based on the AE strategy: Informer (**a**) and AE-Informer (**b**)

2.2.3 Stacking ensemble learning

Before the air pollutant concentration can be predicted, the time series must be supplemented to address the missing values. In recent years, with the rapid development of machine learning, many studies have applied machine learning models to the field of missing data supplementation [34, 35], and ensemble methods can integrate these basic machine learning models to improve performance.

Stacking ensembles are ensemble learning techniques that fuse multiple regression models through a meta-regressor. Each base regression model uses the complete training set during training, and the output of each base regression model during the ensemble learning process is used as a meta-feature that becomes the input of the meta-regressor. The meta-regressor fits these meta-features to obtain multiple fused models. In this approach, a variety of meta-regressors can be used to effectively reduce the bias and variance of the prediction results. Therefore, in this study, we use the stacking ensemble method (Fig. 5) to fuse five basic models: extreme gradient boosting (XGBoost), light gradient boosting (LGBM), gradient boosting decision tree (GBDT), random forest (RF) and extra tree (ET). This approach improves the accuracy of the supplemented missing data in the air pollutant time series.

2.2.4 Model evaluation

To evaluate the collected time series data, the missing data supplementation experiments were first performed. Then, the data were input into the model to conduct several experiments to (1) determine the optimal input sequence length, (2) determine the optimal sampling factor size, and (3) generate the multivariate air pollutant concentration time series predictions. To evaluate the results of each experiment, three performance metrics were used in this study: the correlation coefficient (R^2), the root mean square error (RMSE), and the mean absolute error (MAE). These metrics are calculated as shown in Eqs. (4), (5) and (6):

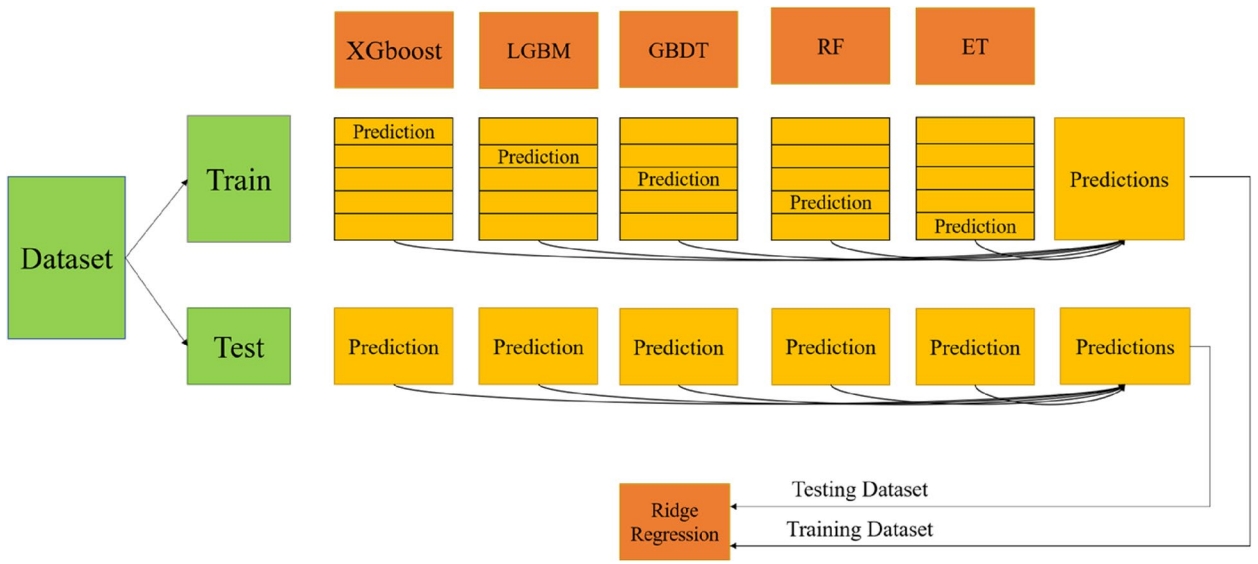

Fig. 5 Stacking ensemble model structure and workflow

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - u_{\widehat{y}})(y_i - u_y)}{\sqrt{\sum_{i=1}^{n}(\widehat{y}_i - u_{\widehat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - u_y)^2}} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2} \quad (6)$$

where n is the predicted sequence length, y_i is the observed value at the i-th position in the air pollutant time series, and $\widehat{y}_i$ is the predicted value at the i-th position in the sequence. $u_{\widehat{y}}$ and u_y are the average forecasted and average observed air pollutant concentrations, respectively.

3 Results and discussion

3.1 Missing data supplementation

The stacking ensemble model and five machine learning models were used to conduct missing data supplementation experiments to verify the effectiveness of the stacking ensemble model in improving the accuracy of the supplemented data. First, the five models, namely, ET, RF, GBDT, XGBoost, and LGBM, were adjusted by Bayesian optimization to achieve the best model effect. These five models were then used as the first layer in the stacking ensemble model. The prediction data of each model's cross operation were fused to form a new training set, and the prediction results of each model's test set were fused to form a new test set. The new training and test sets were passed to the second-layer ridge regression model to train the model, achieving more accurate supplemented data.

As shown in Table 1, compared with other machine learning models, the R^2 value of the four pollutants was the highest in the stacking ensemble model. Among them, the missing value of $PM_{2.5}$ had the highest R^2 value (0.979), and the R^2 values of the other three air pollutants were all greater than 0.87. Compared with the XGBoost machine learning model, the MAE and RMSE of the proposed model were generally reduced by 1–6%. Except for the MAE of NO_2 in the of XGBoost model, the stacking ensemble method yielded better results in terms of all other metrics. The stacking ensemble model improved the accuracy of the supplemented data based on its composition model and exhibited a wide range of applicability to four kinds of pollution (PM_{10}, $PM_{2.5}$, NO_2, and O_3).

In addition, to more intuitively display the effects after supplementing the missing data, Fig. 6 shows the verification results of the four predicted pollutants PM_{10} (a), $PM_{2.5}$ (b), NO_2 (c), and O_3 (d) versus the ground measurements. In the scatter plots of the four pollutants, the scatter points are distributed near the diagonal, which indicates that the prediction values are close to the observed values with low error. Therefore, the stacking ensemble model reliably supplements the missing CEMC data. Among the four pollutants, NO_2 has more scattered points than $PM_{2.5}$, PM_{10} and O_3. This result occurred because NO_2 has more missing values, leading to fewer training samples, which affects the prediction performance of the model. Compared with NO_2 and PM_{10}, the concentration values of O_3 and $PM_{2.5}$ are more evenly distributed, which leads to a higher prediction accuracy.

Table 1 Comparison of the supplemented missing data with the stacking ensemble models

Model	Metric	$PM_{2.5}$	PM_{10}	NO_2	O_3
Stacking	R^2	0.979	0.944	0.871	0.948
	MAE	4.54	7.97	4.23	7.90
	RMSE	6.96	13.9	6.53	10.8
XGBoost	R^2	0.966	0.916	0.870	0.942
	MAE	5.66	9.79	4.22	8.33
	RMSE	8.98	16.9	6.53	11.4
LGBM	R^2	0.951	0.888	0.836	0.938
	MAE	7.13	12.4	5.01	8.85
	RMSE	10.8	19.6	7.34	11.8
GBDT	R^2	0.972	0.932	0.865	0.945
	MAE	5.16	8.85	4.30	8.20
	RMSE	8.08	15.2	6.66	11.1
RF	R^2	0.946	0.917	0.806	0.906
	MAE	7.38	9.80	5.34	10.5
	RMSE	11.2	16.9	7.98	14.6
ET	R^2	0.978	0.937	0.803	0.862
	MAE	4.64	8.52	5.35	12.9
	RMSE	7.15	14.6	8.04	17.6

3.2 Determining the input sequence length

In the Informer model, the input sequence length represents how many hours or days of pollutant concentration data the model needs to use to predict the pollutant concentration in the next hour or day, shorter input sequences cannot ensure that the model has sufficient historical air pollutant data, while long input sequences increase irrelevant inputs and the computational complexity. Therefore, it is necessary to determine the optimal input sequence length to achieve the best model prediction performance. To determine the most appropriate input sequence length, 10 prediction experiments were conducted with different input sequence lengths for 60 state-controlled stations in the study area, and the results of each experiment were averaged to obtain the final RMSE and MAE performance indicators.

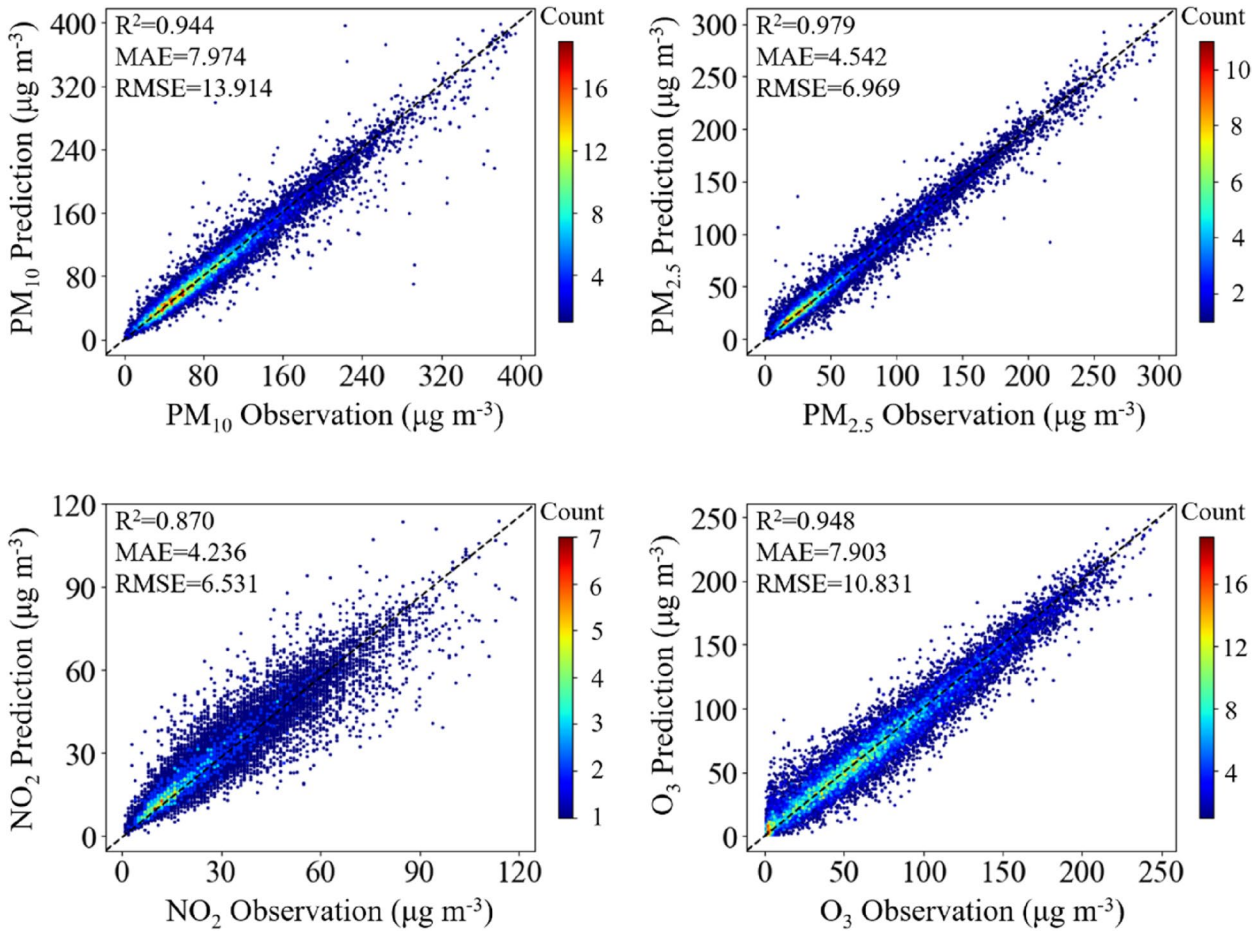

Fig. 6 Scatter plots of the predicted and observation results of PM_{10}, $PM_{2.5}$, NO_2 and O_3 based on the stacking ensemble model

Table 2 Prediction performance using different input sequence lengths

Encoder Length	MAE	RMSE	Encoder Length	MAE	RMSE
6	5.66	9.63	28	5.75	9.69
8	5.57	9.47	30	5.73	9.70
10	5.67	9.63	32	5.76	9.73
12	5.65	9.67	34	5.76	9.74
14	5.67	9.62	36	5.73	9.65
16	5.66	9.59	38	5.75	9.67
18	5.63	9.58	40	5.74	9.63
20	5.61	9.54	42	5.70	9.63
22	5.68	9.66	44	5.71	9.63
24	5.71	9.70	46	5.70	9.66
26	5.70	9.64	48	5.75	9.69

Table 2 shows the MAE and RMSE values obtained with different encoder lengths at the Luohe University site. When the input sequence length was 6–48, the MAE and RMSE ranged from 5.5–5.8 and 9.4–9.8, respectively. When the encoder length was 8, the MAE and RMSE reached their lowest values, which indicates that the AE-Informer model has the lowest prediction error and highest model accuracy. Other national control stations also obtain better prediction effects when the encoder length is 8. Therefore, the length of the input sequence is set to 8 in this study.

3.3 Optimal sampling factor size

By evaluating the KL divergence, we found that there was a large difference between the attention distribution and the uniform distribution, demonstrating that the self-attention mechanism was sparse, and only a small amount of data in the air pollutant sequence contributed important historical information. Selecting data at fewer positions in the input sequence yields less historical information, resulting in insufficient information to make accurate predictions. However, selecting data at too many locations leads to considerably complex historical information, which increases the noise in the prediction

results and the computational costs. Therefore, it is very important to select data at an appropriate number of positions in the sequence. According to the size of the sampling factor (Factor), the first few positions with the highest sparsity scores are selected by the model to obtain historical information. To determine the optimal sampling factor size, we conducted 10 prediction experiments using different sampling factors, and the default input sequence length was the optimal input sequence length determined in Sect. 3.2. The results of each experiment were averaged to obtain the final RMSE and MAE to determine the optimal sampling factor.

The experimental results from the Luohe University site are shown in Table 3. When the sampling factor was 5, the MAE and RMSE of the air pollutant prediction results reached 5.57 and 9.4, respectively, which proves that the model achieves the best prediction effect with this sampling factor. In the experiments at other national control stations, high prediction accuracy was also achieved when the factor was 5. Therefore, the sampling factor is set to 5 in this work.

3.4 Comparison of experimental results

To demonstrate the effectiveness of the AE-Informer model, multivariate prediction experiments were conducted on the AE-Informer model and other commonly used models, and the experimental results were compared. The comparison models included ARIMA [17], Informer [31], Transformer [24], Bi-LSTM [22], the gated recurrent unit (GRU) [36], long short-term memory (LSTM) model [37] and long short-term network (LST-Net) [38]. The ARIMA model was divided into three components: the autoregressive (AR) term, the differential term, and the moving average (MA) term. The AR term refers to the past value used to predict the next value, the MA term defines the number of past prediction errors when predicting future values, and the difference term specifies the number of times that the difference operation is performed on the sequence. The difference operation ensures that the data remain balanced. The traditional Transformer and Informer models are typical models that use attention mechanisms for time series prediction. The Bi-LSTM, GRU, and LSTM models are typical models that use RNN structures to solve time series prediction problems. LST-Net applies the CNN structure to the field of time series prediction.

The results show that the AE-Informer model proposed in this paper outperforms the traditional Informer model and the other comparative models in terms of air pollutant prediction (Table 4). The R^2, MAE, and RMSE of the AE-Informer model reached 0.976, 5.42, and 9.41, respectively, and the error was reduced by 3–7% compared with the other models. The MAE and RMSE of the Informer 13% less than those of the ARIMA model, and the prediction accuracy was significantly improved. The experiment proved that the Transformer and Informer models based on the self-attention mechanism outperform the RNN-based Bi-LSTM, GRU, and LSTM models and the CNN-based LST-Net model. The traditional ARIMA prediction model appears to have inadequate time series prediction performance.

Table 5 shows the evaluation indicators for the individual prediction results of the four pollutants. The simultaneous prediction of multiple pollutants does not

Table 3 Prediction performance using different sampling factors on site data of Luohe University

Factor	MAE	RMSE	Factor	MAE	RMSE
1	5.63	9.52	6	5.60	9.05
2	5.66	9.56	7	5.56	9.46
3	5.60	9.52	8	5.56	9.47
4	5.57	9.49	9	5.59	9.49

Table 5 Evaluation indicators for the prediction results of $PM_{2.5}$, PM_{10}, O_3, and NO_2

Air	R	MAE	RMSE
$PM_{2.5}$	0.883	3.09	5.16
PM_{10}	0.931	8.24	14.4
NO_2	0.923	3.29	4.85
O_3	0.972	7.08	9.78

Table 4 Performance comparison of the AE-Informer model and other models for multivariate time series prediction

Evaluation indicators	AE-Informer	Informer	Transformer	LST-Net	Bi-LSTM	GRU	LSTM	ARIMA
R^2	0.976	0.975	0.974	0.967	0.957	0.956	0.960	0.894
MAE	5.42	5.58	5.66	5.96	6.58	6.84	6.58	7.84
RMSE	9.41	9.53	9.73	11.1	10.6	10.7	10.4	10.5

affect the prediction effects of single pollutants, and the correlation between the predicted and true values of each pollutant is approximately 0.85.

3.5 Prediction performance of the AE-Informer model

After improving the Informer model by introducing AEs, AE-Informer was used to predict the hourly concentrations of four common pollutants. To compare the prediction performance before and after the model was improved, the change curve of the predicted and actual value was generated. The change curve demonstrates the effectiveness of the AE-Informer model by showing the consistency between the predicted and actual results. Figure 7 depicts a comparison of the observed and predicted change curves of the AE-Informer model ((a) is PM_{10}, (b) is $PM_{2.5}$, (c) is NO_2, and (d) is O_3). The blue line represents the actual air pollutant value, and the red line represents the predicted value. In the change curve of the four pollutants, for some extreme value predictions, the error between the AE-Informer prediction value and the observation value is small. In addition, at some time series steps, the predicted value of the AE-Informer model is consistent with the observed value.

The above experiments were based on the data of one site as an example. To evaluate the broad applicability of the proposed AE-Informer model, the model was applied to all other state control sites within the study area. The Informer model and the AE-Informer model were applied to predict the hourly pollutant concentrations at each monitoring station in Henan Province. Then, kriging interpolation was used to interpolate the RMSE and MAE values in the study area to more intuitively demonstrate the improvement in the prediction performance.

The AE-Informer and Informer prediction evaluation indicators after kriging interpolation are shown in Fig. 8, with (a) showing the AE-Informer multivariate

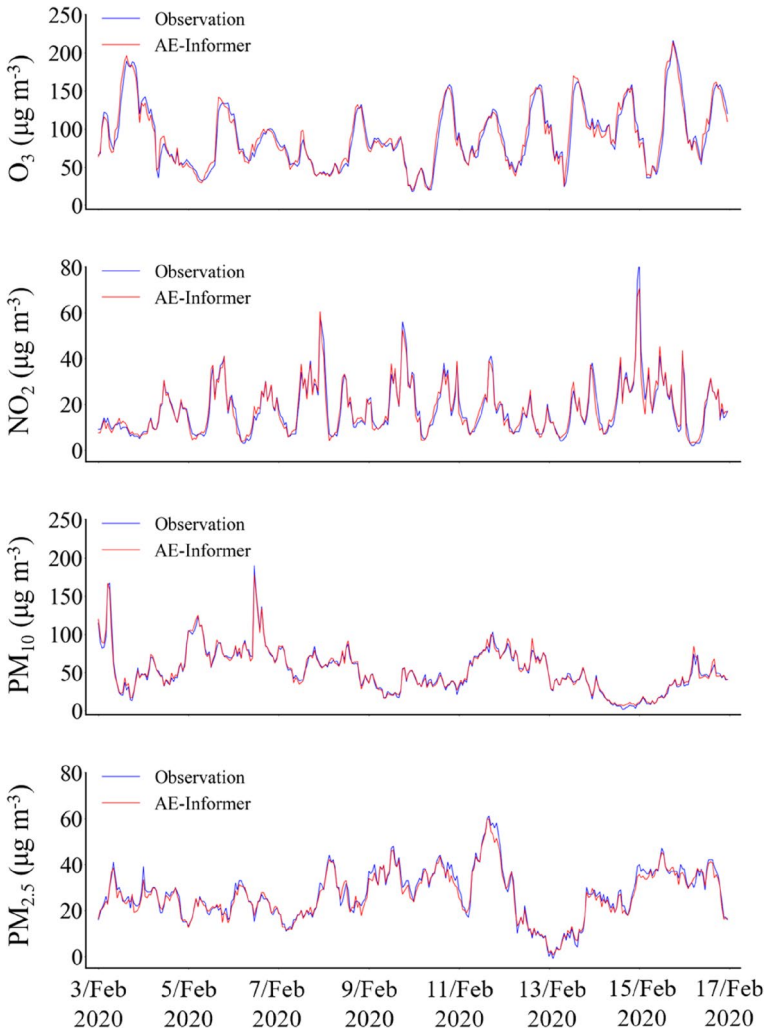

Fig. 7 Comparison between the change curves of the predicted and observed values based on the AE-Informer model

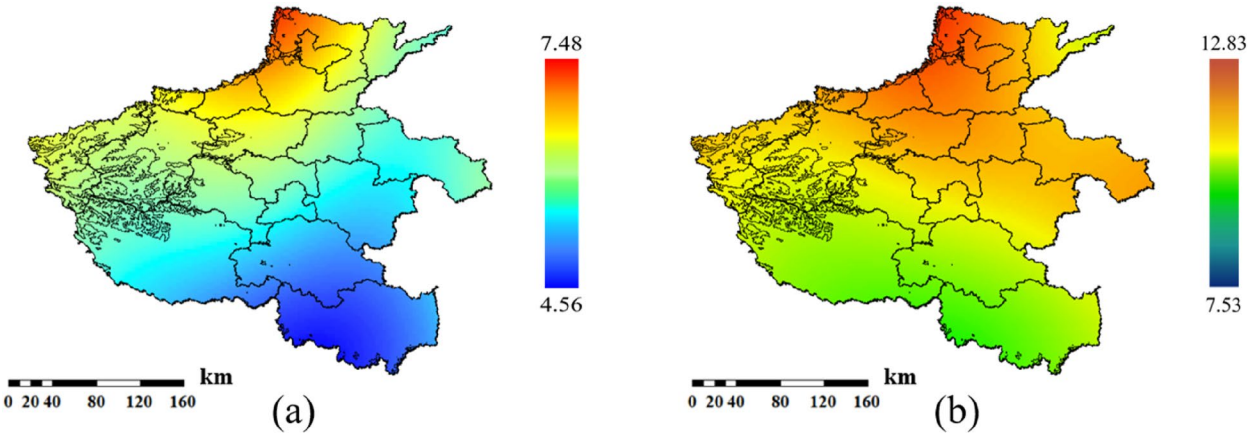

Fig. 8 Spatial distribution of the performance indicators for air pollutant time series prediction: **a** MAE and **b** RMSE

prediction MAE and (b) showing the AE-Informer multivariate prediction RMSE. The bottom color in the color bar indicates a higher prediction level and a smaller prediction error, while the top color indicates a lower prediction level and a larger prediction error. The MAE value of the AE-Informer model ranges from 4.56–7.48, and the RMSE value ranges from 7.53–12.8. The prediction performance in the whole study area is significantly improved in terms of both the MAE and RMSE. Thus, adding the AE effectively reduces the prediction error of the Informer model and improves the hourly prediction accuracy. Moreover, the model is generally applicable in the whole research area, proving the effectiveness of the method proposed in this paper.

4 Conclusions

In conclusion, in this paper, we proposed a methodological framework for studying the effectiveness of the Informer model and AE in improving the prediction accuracy of air pollutant concentrations and compared the prediction performance of various models. Introducing the AE improved the air pollutant concentration time series prediction accuracy. The hourly air pollutant concentration data from all available monitoring stations in Henan Province from 2019 to 2020 were obtained to test the validity of our method. The main contributions of this study can be summarized as follows:

(1) The stacking ensemble method was introduced to supplement missing time series data. Five basic meta-regressors, XGBoost, LGBM, GBDT, RF, and ET, were integrated, and their performance was compared. The experimental results showed that stacking improved the accuracy of missing time series data supplementation; compared with the XGBoost model, the MAE and RMSE of $PM_{2.5}$ were reduced by up to 6% when the proposed model was applied.

(2) For the first time, the Informer model was applied in the field of air pollutant time series prediction. The self-attention mechanism in the Informer model efficiently obtained historical time series information. The experimental results showed that the MAE and RMSE of the proposed model were 13% less than those of the ARIMA model, and the prediction accuracy was significantly improved.

(3) This paper is one of the few pioneering studies that fuses deep learning with the AE strategy to predict air pollutant concentration. This model can help governments and researchers assess trends more accurately in long-term air quality analyses, especially for multivariate time series forecasting.

(4) The results showed that the AE-Informer model proposed in this paper effectively improved the prediction of air pollutant concentrations in multivariate time series. Compared with the Informer model, the MAE and RMSE values of the proposed model were reduced by 3%, and the errors of the predicted values were also reduced.

This research can be extended to explore higher resolution data. Moreover, transfer learning can be introduced to achieve daily time series prediction, and data from more state controlled sites can be applied to assess areas with fewer state controlled sites.

Acknowledgements

We would like to acknowledge the use of the mass PM_{10}, $PM_{2.5}$, NO_2, and O_3 concentration data from http://113.108.142.147:20035/emcpublish.

Authors' contributions

Kun Cai designed the experiments, downloaded the data, processed the data, and wrote the paper. Xusheng Zhang and Ming Zhang edited the paper and adjusted unreasonable sentences. Qiang Ge provided visualizations, participated in the investigation and developed the methodology. Shenshen

Li revised the full text, provided useful suggestions about the validation experiments and provided funding support. Baojun Qiao supervised the experiments and provided resources. Yang Liu provided funding support. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (Grant No. 42071409), the Open Foundation of Key Laboratory of Ecological Environment Protection of Space Information Application of Henan (22FW070108), the Key Research Projects of Henan Higher Education Institutions (23A520024), the Shenzhen Special Foundation of Central Government to Guide Local Science & Technology Development (2021Szvup032) and the Major Project of Science and Technology of Henan Province (201400210300, 201300311400).

Availability of data and materials
All data that were generated or analyzed during this study are available upon request.

Declarations

Competing interests
The authors declare no conflict of interest.

Received: 22 November 2022 Accepted: 23 March 2023
Published online: 14 April 2023

References

1. Cai K, Li SS, Zheng FB, Yu C, Zhang XY, Liu Y, et al. Spatio-temporal variations in NO_2 and $PM_{2.5}$ over the Central Plains Economic Region of China during 2005–2015 based on satellite observations. Aerosol Air Qual Res. 2018;18:1221–35.
2. Li SS, Ma ZW, Xiong XZ, Christiani DC, Wang ZX, Liu Y. Satellite and ground observations of severe air pollution episodes in the winter of 2013 in Beijing, China. Aerosol Air Qual Res. 2016;16:977–89.
3. Li SS, Chen LF, Xiong XZ, Tao JH, Su L, Han D, et al. Retrieval of thE HAZE OPTICAL THICKNess in North China Plain using MODIS Data. IEEE T Geosci Remote. 2013;51:2528–40.
4. Li G, Zeng Q, Pan X. Disease burden of ischaemic heart disease from short-term outdoor air pollution exposure in Tianjin, 2002–2006. Eur J Prev Cardiol. 2016;23:1774–82.
5. Xiao CC, Chang M, Guo PK, Gu MF, Li Y. Analysis of air quality characteristics of Beijing–Tianjin–Hebei and its surrounding air pollution transport channel cities in China. J Environ Sci-China. 2020;87:213–27.
6. Zhou CJ, Wei G, Zheng HP, Russo A, Li CC, Du HD, et al. Effects of potential recirculation on air quality in coastal cities in the Yangtze River Delta. Sci Total Environ. 2019;651:12–23.
7. Chen ZJ, Cui LL, Cui XX, Li XW, Yu KK, Yue KS, et al. The association between high ambient air pollution exposure and respiratory health of young children: A cross sectional study in Jinan, China. Sci Total Environ. 2019;656:740–9.
8. Song Y, Zhou AN, Zhang M. Exploring the effect of subjective air pollution on happiness in China. Environ Sci Pollut R. 2020;27:43299–311.
9. Kang Z, Qu ZY. Application of BP neural network optimized by genetic simulated annealing algorithm to prediction of air quality index in Lanzhou. In: 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). Beijing: IEEE; 2017.
10. Li X, Peng L, Yao XJ, Cui SL, Hu Y, You CZ, et al. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environ Pollut. 2017;231:997–1004.
11. Singh KP, Gupta S, Kumar A, Shukla SP. Linear and nonlinear modeling approaches for urban air quality prediction. Sci Total Environ. 2012;426:244–55.
12. Stern R, Builtjes P, Schaap M, Timmermans R, Vautard R, Hodzic A, et al. A model inter-comparison study focussing on episodes with elevated PM_{10} concentrations. Atmos Environ. 2008;42:4567–88.
13. Berkowicz R. OSPM - A parameterised street pollution model. Environ Monit Assess. 2000;65:323–31.
14. Catalano M, Galatioto F. Enhanced transport-related air pollution prediction through a novel metamodel approach. Transport Res D-Tr E. 2017;55:262–76.
15. Kukkonen J, Partanen L, Karppinen A, Ruuskanen J, Junninen H, Kolehmainen M, et al. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmos Environ. 2003;37:4539–50.
16. Jian L, Zhao Y, Zhu YP, Zhang MB, Bertolatti D. An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. Sci Total Environ. 2012;426:336–45.
17. Shukur OB, Lee MH. Daily wind speed forecasting through hybrid KF-ANN model based on ARIMA. Renew Energ. 2015;76:637–47.
18. Davis JM, Speckman P. A model for predicting maximum and 8 h average ozone in Houston. Atmos Environ. 1999;33:2487–500.
19. Slini T, Karatzas K, Moussiopoulos N. Statistical analysis of environmental data as the basis of forecasting: an air quality application. Sci Total Environ. 2002;288:227–37.
20. Ma J, Cheng JCP. Data-driven study on the achievement of LEED credits using percentage of average score and association rule analysis. Build Environ. 2016;98:121–32.
21. Rubal, Kumar D. Evolving differential evolution method with random forest for prediction of air pollution. Procedia Comput Sci. 2018;132:824–33.
22. Ma J, Cheng JCP, Lin CQ, Tan Y, Zhang JC. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmos Environ. 2019;214:116885.
23. Chauhan R, Kaur H, Alankar B. Air quality forecast using convolutional neural network for sustainable development in urban environments. Sustain Cities Soc. 2021;75:103239.
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: 31st Annual Conference on Neural Information Processing Systems (NIPS). Long Beach: MIT Press; 2017.
25. Li SY, Jin XY, Xuan Y, Zhou XY, Chen WH, Wang YX, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In: 33rd Conference on Neural Information Processing Systems (NIPS). Vancouver: MIT Press; 2019.
26. Guo QP, Qiu XP, Liu PF, Shao YF, Xue XY, Zhang Z. Star-Transformer. In: Conference of the North-American-Chapter of the Association-for-Computational-Linguistics - Human Language Technologies (NAACL-HLT). Minneapolis: ACL; 2019.
27. Yu WH, Luo M, Zhou P, Si CY, Zhou YC, Wang XC, et al. MetaFormer is actually what you need for vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE; 2022.
28. Chen MH, Peng HW, Fu JL, Ling HB. AutoFormer: searching transformers for visual recognition. In: EEE/CVF International Conference on Computer Vision (ICCV). Virtual: IEEE; 2021.
29. Dai ZH, Yang ZL, Yang YM, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. In: Association for Computational Linguistics (ACL) 2019. Florence: ACL; 2019.
30. Lee J, Lee Y, Kim J, Kosiorek AR, Choi S, The YW. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In: 36th International Conference on Machine Learning (ICML). California: ACM; 2019.
31. Zhou HY, Zhang SH, Peng JQ, Zhang S, Li JX, Xiong H, et al. Informer: beyond efficient transformer for long sequence time-series forecasting. In: 35th AAAI Conference on Artificial Intelligence (AAAI-21). Virtual: AAAI; 2021.
32. Sun FK, Lang CI, Boning DS. Adjusting for autocorrelated errors in neural networks for time series. In: 35th Conference on Neural Information Processing Systems (NIPS). Virtual; 2021 Dec 6–14.
33. Liu SH, Hua SB, Wang K, Qiu PP, Liu HJ, Wu BB, et al. Spatial-temporal variation characteristics of air pollution in Henan of China: Localized emission inventory, WRF/Chem simulations and potential source contribution analysis. Sci Total Environ. 2018;624:396–406.
34. Arriagada P, Karelovic B, Link O. Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. J Hydrol. 2021;598:126454.

35. Hurtado JC Chacon, Alfonso L, Solomatine D. Comparison of machine learning methods for data infilling in hydrological forecasting. In: EGU General Assembly 2014. Vienna; 2014 Apr 27–May 2.
36. Dey R, Salem FM. Gate-variants of Gated Recurrent Unit (GRU) neural networks. In: 60th IEEE International Midwest Symposium on Circuits and Systems (MWSCAS). Boston; 2017 Aug 6–9.
37. Yu Y, Si XS, Hu CH, Zhang JX. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 2019;31:1235–70.
38. Li LD, Wang K, Li S, Feng XC, Zhang L. LST-Net: Learning a convolutional neural network with a learnable sparse transform. In: 16th European Conference on Computer Vision (ECCV) 2020. Glasgow; 2020 Aug 23–28.

Publisher's Note