scientific reports

OPEN

Novel hybrid data-driven modeling based on feature space reconstruction and multihead self-attention gated recurrent unit: applied to PM2.5 concentrations prediction

Xiaoxin Yue, Yulong Bai✉, Qinghe Yu, Lin Ding, Wei Song, Wenhui Liu, Huhu Ren & Qi Song

In response to the problem of neglecting the periodic and global characteristics of sequence data when predicting PM2.5 concentrations via machine learning models, a PM2.5 concentrations prediction model based on feature space reconstruction and multihead self-attention gated recurrent unit (FSR-MSAGRU) is proposed in this study. First, the raw sequence data are subjected to frequency spectrum analysis to determine the period value of the PM2.5 sequence data. Subsequently, the seasonal trend decomposition procedure based on loess (STL) is employed to capture the periodicity and trend information in the PM2.5 sequence data. Then, the feature space of the PM2.5 sequence data is reconstructed using the raw PM2.5 sequence data, decomposed seasonal components, trend components, and residual components. Finally, the reconstructed feature data are input into multihead self-attention gated recurrent unit (MSAGRU) with the ability to capture global feature information to predict PM2.5 concentrations. Favorable prediction results were attained by the proposed FSR-MSAGRU model across 6 distinct experimental datasets, with a PCC exceeding 0.98 and a decrease in the prediction accuracy metric SMAPE of at least 68% compared to that of the GRU model. Comparative experimental results with 13 reference models demonstrate that the proposed model exhibits better prediction performances and stronger generalization abilities.

Keywords Machine learning, Feature space reconstruction, Multihead Self-attention, Gated recurrent unit, PM2.5 concentration prediction

Background

Accurate and effective predictions of PM2.5 concentrations are important for air quality management[1]. PM2.5, characterized by particles with a diameter less than 2.5 micrometers, is regarded as one of the most detrimental components of air pollution[2,3]. Studies have indicated that prolonged exposure to high concentrations of PM2.5 not only increases the risk of lung cancer[4] but also leads to severe damage to the heart and other organs[2,5,6]. Due to the hazards caused by excessive PM2.5 concentrations, there has been a growing focus on predictions. However, the performance of the current methods for predicting PM2.5 concentrations is still insufficient[7]. Hence, achieving greater precision in forecasting PM2.5 concentrations is instrumental in shaping and executing alert decision-making procedures, which are indispensable for safeguarding public health and tackling environmental issues[1,8,9].

Research status

There are three types for PM2.5 concentrations predictions, the mechanistic models, statistical models, and machine learning models. Mechanistic models primarily focused on investigating the PM2.5 formation mechanisms[10], taking into account the interactions among various pollutants and employing mathematical methods to describe the diffusion and deposition processes of PM2.5[10,11]. Prominent examples of such

College of Physics and Electrical Engineering, Northwest Normal University, Lanzhou 730070, Gansu, China. ✉email: baiyulong@nwnu.edu.cn

models include the weather research and forecasting model coupled with chemistry (WRF-Chem)[12], the community multiscale air quality model (CMAQ)[10], and the comprehensive air quality model with extensions (CAMx)[13], among others. However, due to limited knowledge of air pollution sources and the influence such as geography and meteorology, mainstream mechanistic models[14,15] exhibit significant biases in predicting PM2.5 concentration, especially during heavy pollution periods in autumn and winter, with prediction biases reaching as high as 30–50%[16–18]. Compared to mechanistic models, statistical models based on historical data are simpler and often exhibit higher prediction accuracy[19,20]. These models include the autoregressive integrated moving average model (ARIMA)[21], the Kalman filter[22], and the multiple linear regression (MLR)[23], among others. Given the nonlinear and non-stationary nature of PM2.5, these statistical models often face challenges in achieving the desired level of predictive accuracy[24,25].

PM2.5 concentration is predicted using machine learning models, including the extreme learning machines (ELM)[26], the support vector regression (SVR)[27], and other artificial neural networks (ANN)[28]. These models leverage neural networks and other machine learning methods to extract patterns in historical PM2.5 and predict future trends. Compared to traditional mechanistic and statistical models, machine learning models are proficient in capturing the nonlinear features of PM2.5[29]. In recent years, recurrent neural network models, proficient in handling sequence data, have demonstrated advantages in PM2.5 prediction. For example, gated recurrent units (GRU) networks[30–32] and long short-term memory (LSTM) networks[33,34] have been used for PM2.5 concentration prediction. Despite their effectiveness, basic single recurrent neural networks (RNNs) models often struggle to comprehensively explore data variation patterns, potentially leading prediction instabilities[7]. Currently, researchers are actively investigating strategies to enhance the predictive performance of single models from various perspectives.

Researchers have sought to enhance the prediction performance through model improvements or the hybridization of different models. Advanced models such as directional long short-term memory (BiLSTM)[35], built upon improved versions of LSTM, effectively reduce prediction errors by capturing features from both forward and backward directions. Combining different neural networks, such as convolutional and recurrent networks[36–38], leverages their complementary advantages for improved prediction effectiveness. The attention mechanism, widely applied in various fields[39,40], enhances prediction accuracy by extracting long-term dependencies and minimizing information loss through probabilistic weight allocations. Recent studies[19,41,42] have demonstrated that PM2.5 prediction models that integrate attention mechanisms can enhance the precision and generalizability of predictions. The integration of attention mechanisms into PM2.5 prediction models enhances performance but requires attention weights to be computed at each time step[35,42,43].

The integration of data preprocessing into machine learning models effectively enhances the PM2.5 prediction performance. By integrating data preprocessing techniques, the predictive accuracy of PM2.5 models has improved through simplifying and decomposing the original sequence data into subsequences with distinct informational content[7,41]. For instance, hybrid models that combine empirical mode decomposition (EMD) and its variants, or variational mode decomposition (VMD) with neural networks, have demonstrated substantial improvements in PM2.5 prediction[41,42,44,45]. Further refinements using hybrid data decomposition techniques and secondary decomposition methods have enabled more detailed feature extractions, boosting model performance[46]. In order to capture the periodicity characteristics of PM2.5 sequence data, the method of seasonal trend decomposition based on loess (STL) was attempted by researchers[47–50] to separate the trend factors and periodic variation characteristics of sequence data.

Research motivation and innovation

Summary of the research status and research gap

The aforementioned studies indicate that employing various model hybridizations and integrating data preprocessing into machine learning models can effectively enhance the prediction performance of PM2.5 concentrations. Various methods for improving the performance of the PM2.5 prediction models are summarized in Table 1. According to Table 1, the following research gap can be observed:

Reference	Data Preprocessing	Model Improvement Methods	Reference	Data Preprocessing	Model Improvement Methods
Reference[35]	–	Bidirectional LSTM	Reference[9]	EMD	–
Reference[51]	–	FL-DPLA + LSTM	Reference[52]	EMD	BiLSTM
Reference[37,53,54]	–	CNN + LSTM	Reference[55]	EEMD	–
Reference[56]	–	CNN + GRU	Reference[57]	CEEMDAN	CNN + LSTM
Reference[36]	–	3D CNN + GRU	Reference[44]	BVMD	CNN + ELM
Reference[38]	–	1D-CNN + biLSTM	Reference[58]	MVMD	KRidge
Reference[59]	–	Auto-Encoder + Bi-LSTM	Reference[46]	CEEMDAN-VMD	–
Reference[19]	–	GAT + GNN + GRU	Reference[48]	STL	–
Reference[60]	–	CNN + LSTM + Attention	Reference[47]	STL	GRU + ARIMA
Reference[61]	–	CNN + BiLSTM + Attention	Reference[62]	CEEMDAN	Attention + LSTM
Reference[63]	–	CNN + BiLSTM + Attention	Reference[41]	CEEMDAN	BiLSTM + Attention
Reference[64]	CEEMD	AE + BiLSTM	Reference[42]	CEEMD	ProbSparse + Self-Attention

Table 1. PM2.5 concentrations prediction models with different hybridization methods.

(1) The data preprocessing method utilized is EMD or its variants. A certain level of effectiveness can be achieved in predicting PM2.5 concentrations when utilizing EMD and its variants within integrated data processing machine learning models. However, it is worth noting that these data preprocessing methods may neglect the periodicity of PM2.5 [7,48].

(2) Conventional attention mechanisms may decrease modeling efficiency. Conventional attention mechanisms necessitate the calculation of attention weights at each time step or position. While the integration of conventional attention mechanisms with neural networks can enhance the predictive performance of the model, it also results in decreased model efficiency[42,64].

(3) Predicting subsequences independently could disrupt their relationship with the raw sequence. During training, each subsequence is predicted separately, and their predictions are aggregated to determine the final result. This approach, concentrating solely on particular segments of the input sequence, overlooks intersegment relationships, potentially impeding the model's capacity to capture complex dependencies and comprehensive information across the sequence[43,65,66].

To address these issues, a novel hybrid data-driven model for predicting PM2.5 concentrations, called FSR-MSAGRU, is proposed. This model leverages feature space reconstruction and a multi-head self-attention gated recurrent unit. It not only captures the periodic characteristics of the sequence data but also effectively captures the global feature information within the data.

Contributions and innovations

The innovations and contributions of this study are summarized as follows:

(1) Based on gated recurrent unit, a multihead self-attention gated recurrent unit (MSAGRU) model that incorporates perception global feature information is proposed. The dynamic characteristics of the sequence data are captured by the model using a GRU, and multiple key information units within the GRU is concurrently processed through a multihead self-attention mechanism, facilitating the capture of the global feature information.

(2) By employing the STL, the periodic and trend variations within the PM2.5 sequence data can be accurately captured, mitigating noise interference in the prediction results. To determine the parameter values for the STL, Fourier transform analysis is utilized in this study to analyze the sequence data and identify the period value from the spectrum plot, reducing the influence of human factors on the decomposition results.

(3) A feature space reconstruction method for data preprocessing is proposed. It reconstructs the feature space from the raw PM2.5 sequence data and the STL decomposition results, and encompasses the seasonal, trend, and residual components. This method retains vital raw data information while incorporating sequence change trends and periodicity features, enhancing the feature expression capacity of the sequence data.

(4) A new PM2.5 prediction model is proposed that integrates model enhancements and data preprocessing. Compared with 13 reference models on 6 different experimental datasets, the proposed feature space reconstruction and multihead self-attention gated recurrent unit (FSR-MSAGRU) model achieves more favorable prediction performances and more robust generalizability.

Structure of the paper

The remaining sections of this paper are organized as follows. In Sect. 2, the process and results of reconstructing the feature space of PM2.5 sequence data are described. In Sect. 3, the framework and basic theoretical methods of the FSR-MSAGRU model proposed in this study are described, and the metrics used to evaluate the performance of each model are outlined. The results of the ablation and comparative experiments for the proposed model are presented in Sect. 4. In Sect. 5, the performance of the proposed model is verified through hypothesis testing and a comprehensive evaluation of the model's prediction performance. The influence of individual model enhancements and data preprocessing on improving model performance is thoroughly discussed, as well as the advantages and limitations of the FSR-MSAGRU model. Finally, the conclusions of this study are stated, and suggestions for future research are proposed in Sect. 6.

Data and data processing

This section primarily presents PM2.5 data and describes how its feature spaces is reconstructed through data processing.

Data description

In this study, the PM2.5 data observed at the air quality monitoring stations located in different regions and altitudes were selected. The data were obtained from the "PM2.5 Historical Data Network" (https://www.aqistudy.cn/historydata/). Data from 6 cities (the cities where data were collected are shown in Fig. 1), including Lanzhou, Xi'an, Beijing, Shijiazhuang, Chengdu, and Lhasa, covering the period from January 1, 2015, to December 31, 2022, totaling eight years of daily average PM2.5 data, were chosen. The data distributions of the 6 experimental datasets is illustrated in Fig. 2.

Figure 2 shows that PM2.5 exhibits substantial periodic variations across all 6 experimental datasets. This study provides empirical support for studying the seasonal/periodic patterns of PM2.5. Additionally, Fig. 2 also indicates that the amplitude of the PM2.5 fluctuations gradually decreases over time, showing an overall declining trend. This trend reflects positive strides in air quality improvement in China, and underscores the effectiveness and necessity of environmental management measures. In Fig. 2, Shijiazhuang and Lanzhou exhibited larger fluctuations in PM2.5, indicating significant variations in air quality over the different time

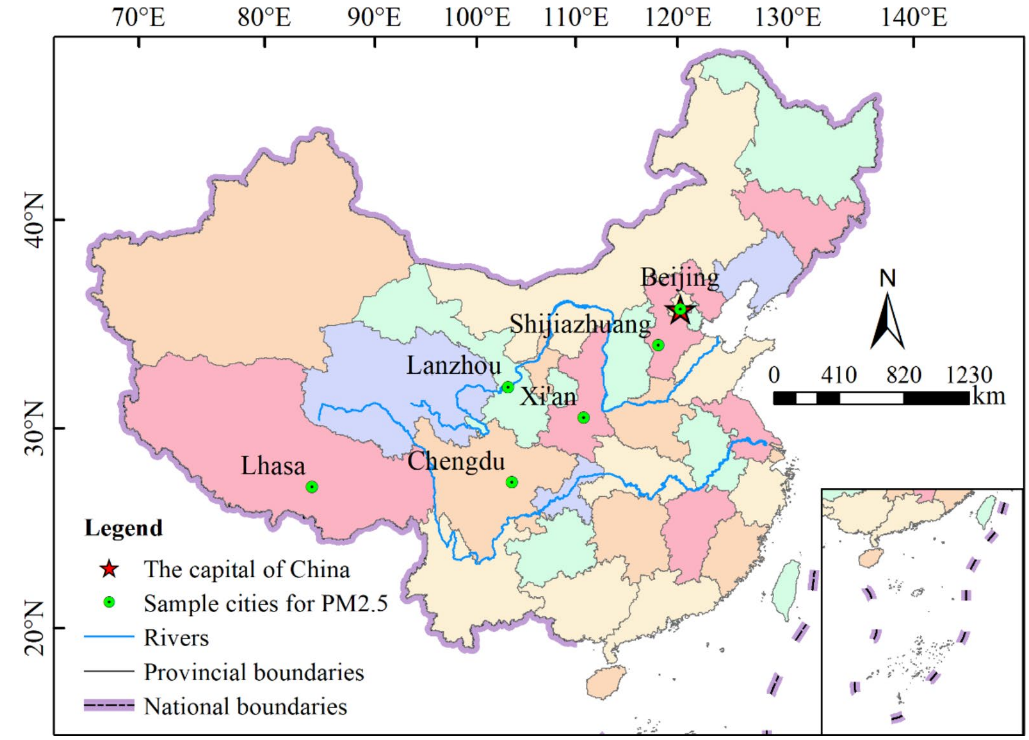


Fig. 1. Locations of cities for data collection.

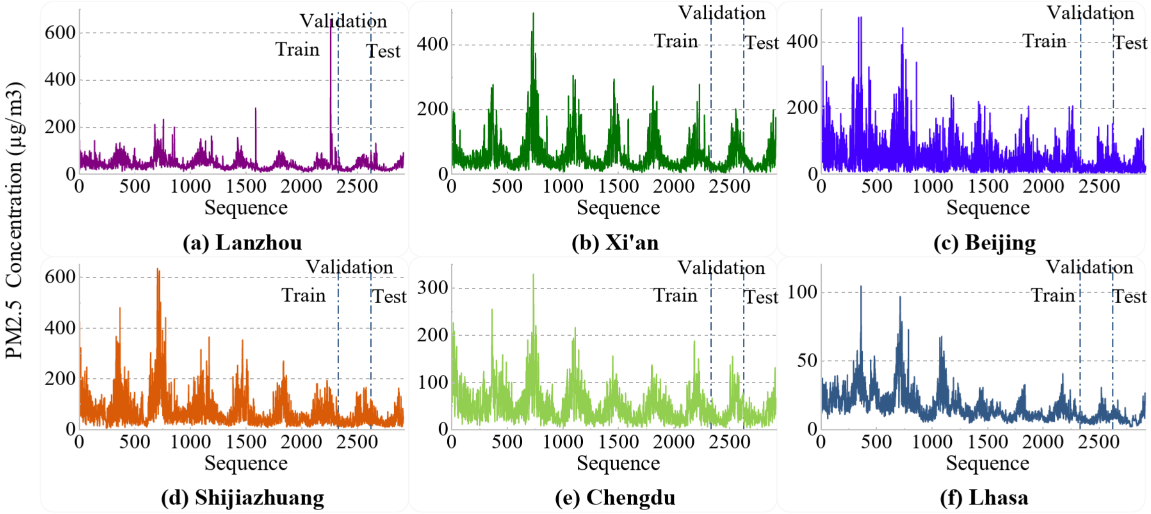


Fig. 2. Distribution of the dataset samples.

periods in these areas. In contrast, the data from Lhasa exhibited smaller fluctuations, suggesting relatively stable air quality conditions.

The sample statistics are presented in Table 2, including the minimum value (Min), maximum value (Max), mean value, standard deviation (Std), skewness, and kurtosis.

Table 2 shows that Lanzhou and Shijiazhuang exhibit extremely high maximum PM2.5 concentrations (657.25 μg/m^3 and 635.5 μg/m^3, respectively), indicating very unhealthy air quality under extreme pollution conditions. Moreover, Shijiazhuang has the highest mean PM2.5 concentration (70.3859 μg/m^3), suggesting generally poor air quality in the region. In contrast, Lhasa has a relatively low PM2.5 maximum concentration (104.5 μg/m^3) and the lowest mean concentration (16.3143 μg/m^3), indicating better air quality with fewer pollution events. The standard deviation of PM2.5 varies significantly across the 6 experimental datasets, indicating notable differences in the PM2.5 levels among the different cities. Among them, Shijiazhuang has the largest standard deviation, implying more significant fluctuations in PM2.5 compared to the other five

Dataset	Min/(μg/m^3)	Max/(μg/m^3)	Mean/(μg/m^3)	Std/(μg/m^3)	Skewness	Kurtosis
Lanzhou	8.6250	657.2500	44.1694	28.2141	5.8781	93.5463
Xi'an	5.1111	498.7917	58.6244	50.2718	2.5154	9.6307
Beijing	2.0417	476.6522	50.9391	50.6265	2.6545	10.8252
Shijiazhuang	6.4348	635.5000	70.3859	63.1089	2.8407	12.7860
Chengdu	3.9167	329.6250	49.4943	35.0086	1.8909	5.4761
Lhasa	2.2105	104.5000	16.3143	10.9104	2.2370	7.9797

Table 2. Statistical information for the dataset samples.

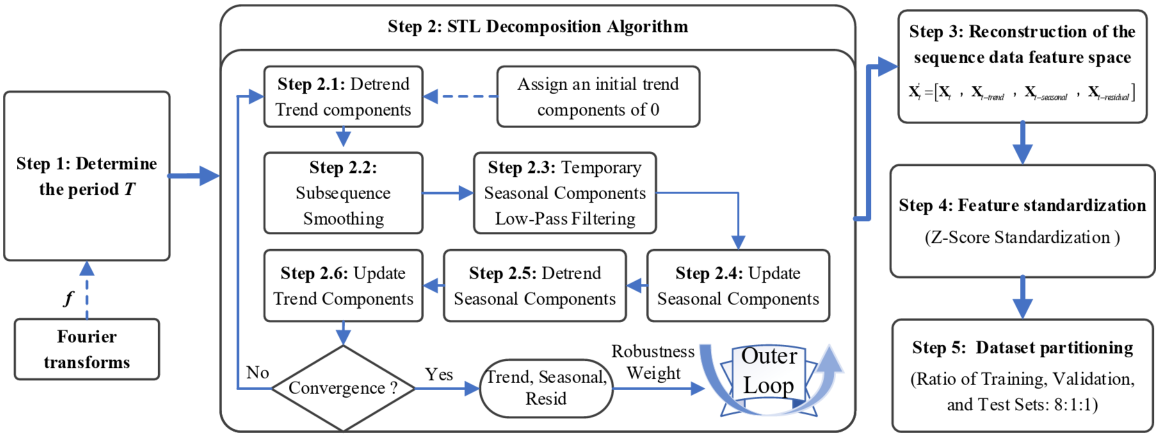

Step 2: STL Decomposition Algorithm
Step 1: Determine the period T
f
Fourier transforms
Step 2.1: Detrend Trend components
Assign an initial trend components of 0
Step 2.2: Subsequence Smoothing
Step 2.3: Temporary Seasonal Components Low-Pass Filtering
Step 2.6: Update Trend Components
Step 2.5: Detrend Seasonal Components
Step 2.4: Update Seasonal Components
No
Convergence ?
Yes
Trend, Seasonal, Resid
Robustness Weight
Outer Loop
Step 3: Reconstruction of the sequence data feature space
$X_t' = [X_t , X_{t-trend} , X_{t-seasonal} , X_{t-residual}]$
Step 4: Feature standardization
(Z-Score Standardization)
Step 5: Dataset partitioning
(Ratio of Training, Validation, and Test Sets: 8:1:1)

Fig. 3. Flowchart of the data processing.

cities. Furthermore, PM2.5 in Lanzhou exhibited very high skewness and kurtosis values (5.8781 and 93.5463, respectively), indicating a highly skewed data distribution with the presence of extreme high values. Conversely, PM2.5 in Chengdu has the lowest skewness and kurtosis values (1.8909 and 5.4761, respectively), suggesting a distribution closer to normal and fewer extreme values.

Data processing

Data processing methods

In this study, the data processing procedure follows a structured workflow to ensure the accuracy and reliability of the PM2.5 data. The detailed flowchart illustrating the steps involved in data preprocessing, feature extraction, and reconstruction is presented in Fig. 3.

The algorithm for the data processing procedure entails the following specific steps:

Step 1 Determine the period T. Perform Fourier transforms on the PM2.5 sequence data collected from different regions, plot the frequency domain spectrum of the sequence, and identify the frequency value f corresponding to the peak frequency spectrum. Calculate the period T of the sequence using Eq. (1).

$$T = 1/f \tag{1}$$

where f represents the peak value of the frequency spectrum of the sequence.

Step 2 Decomposition of sequence data. Using the period T determined in **Step 1** as the parameter value for the STL, the raw PM2.5 sequence data $\mathbf{X}_t$=[x_0, x_1, x_2, ……, x_t] are decomposed into the trend components $\mathbf{X}_{t\text{-}trend}$=[$x_{0\text{-}trend}$, $x_{1\text{-}trend}$, $x_{2\text{-}trend}$, ……, $x_{t\text{-}trend}$], the seasonal components $\mathbf{X}_{t\text{-}seasonal}$=[$x_{0\text{-}seasonal}$, $x_{1\text{-}seasonal}$, $x_{2\text{-}seasonal}$, ……, $x_{t\text{-}seasonal}$], and the residual components $\mathbf{X}_{t\text{-}residual}$=[$x_{0\text{-}residual}$, $x_{1\text{-}residual}$, $x_{2\text{-}residual}$, ……, $x_{t\text{-}residual}$].

Step 3 Reconstruction of the sequence data feature space. The raw PM2.5 sequence data $\mathbf{X}_t$, along with the trend components $\mathbf{X}_{t\text{-}trend}$, seasonal components $\mathbf{X}_{t\text{-}seasonal}$, and residual components $\mathbf{X}_{t\text{-}residual}$ obtained from the STL, are reconstructed into an elevated-dimensional feature space of the PM2.5 sequence data. The reconstructed feature space is represented as$\mathbf{X}'_t$

$$\mathbf{X}'_t = [\mathbf{X}_t, \mathbf{X}_{t-trend}, \mathbf{X}_{t-seasonal}, \mathbf{X}_{t-residual}] \tag{2}$$

where X_t, $\mathbf{X}_{t\text{-}trend,}$ $\mathbf{X}_{t\text{-}seasonal}$, and $\mathbf{X}_{t\text{-}residual}$ respectively represent the raw PM2.5 sequence data, the trend components, the seasonal components, and the residual components.

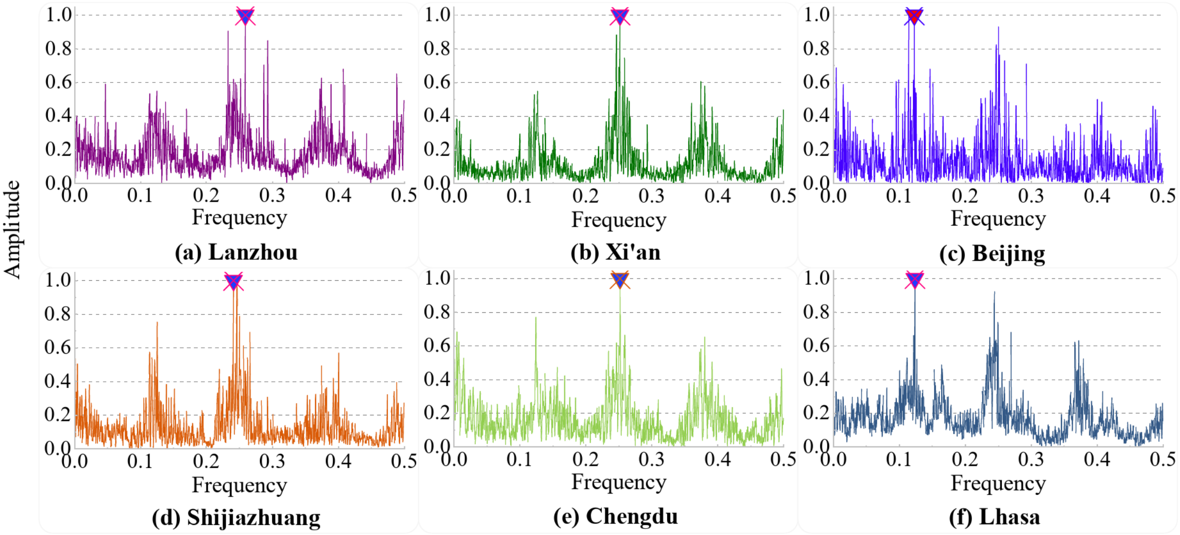


Fig. 4. Frequency spectrum of PM2.5 in different datasets.

	Lanzhou	Xi'an	Beijing	Shijiazhuang	Chengdu	Lhasa
Period *T* of PM2.5 (day)	3.86	3.97	8.17	4.15	3.97	8.10

Table 3. Period *T* of PM2.5 in the different datasets.

Step 4 Feature standardization. Z score standardization is applied to eliminate the influence of the different scales and dimensions in the feature space, ensuring that all the features have the same importance and proportion during model training.

Step 5 Dataset partitioning. the standardized feature space is partitioned into training, validation, and test sets at a ratio of 8:1:1. This partitioning ensures that the model is trained under a unified standard and validated for its generalizability to unknown data.

Data processing results

Variation period *T* of PM2.5 As indicated in Sect. 2.1, PM2.5 exhibits significant periodic variations over time. To further investigate and quantitatively analyze the periodicity of the PM2.5 variations, Fourier transforms were employed to perform frequency domain analysis on the PM2.5 sequence data from the 6 experimental datasets. The frequency domain spectra of PM2.5 in the different experimental datasets are provided in Fig. 4.

The PM2.5 variation period T on the different experimental datasets is presented in Table 3.

Figure 4 and Table 3 show that the variation period of PM2.5 varies significantly across the different experimental datasets, which is associated with the differences in the geographical location, environmental policies, and local meteorological conditions of the cities.

Reconstructed feature spaces Based on the specific feature space reconstruction process in Sect. 2.2.1, the additive model of the STL is used to decompose the raw PM2.5 sequence data, obtaining the trend components $\mathbf{X}_{t\text{-}trend}$, seasonal components $\mathbf{X}_{t\text{-}seasonal}$, and residual components $\mathbf{X}_{t\text{-}residual}$. The above decomposed data are dimensionally expanded according to Eq. (2) to reconstruct the feature space of PM2.5. The reconstructed feature spaces of the different datasets are depicted in Fig. 5.

Figure 5 shows that the reconstructed feature space, transformed from one-dimensional feature vectors to a four-dimensional feature space, integrates features such as sequence trends and periodic variations while retaining all the important information from the original data, enriching the feature representation of the sequence.

Models

This section provides a detailed introduction to the working principles and processes of the proposed model. Subsequently, a theoretical description of the proposed model is provided. Finally, the metrics utilized to evaluate the performance of the model predictions in this study are introduced.

Construction of the proposed model

This section provides a comprehensive overview of the proposed feature space reconstruction and multihead self-attention gated recurrent unit (FSR-MSAGRU) model, outlining its mathematical formulation, framework,

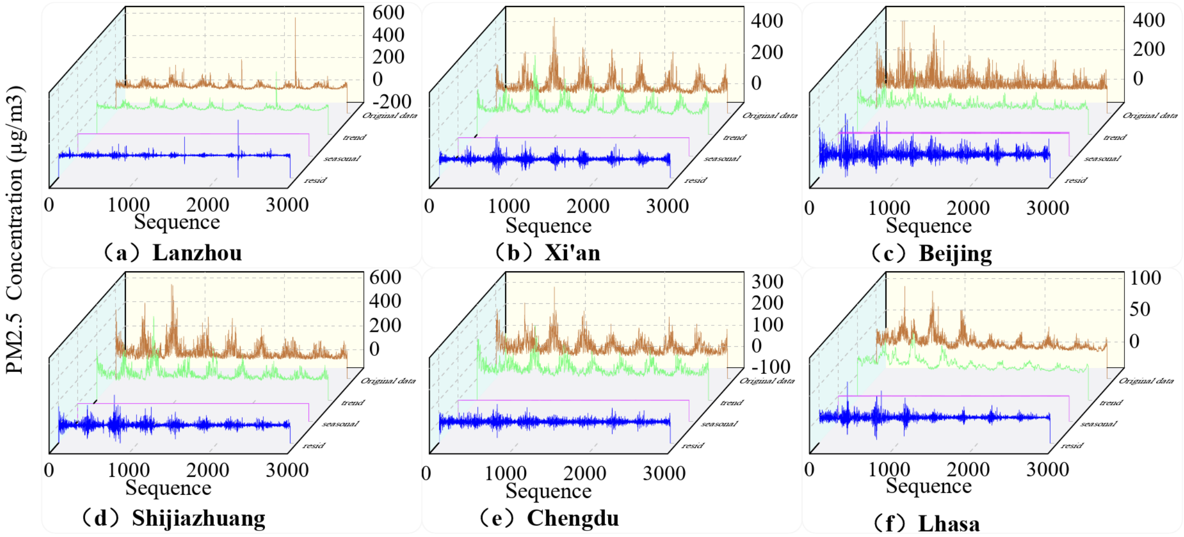


Fig. 5. Reconstructed feature spaces in the different datasets.

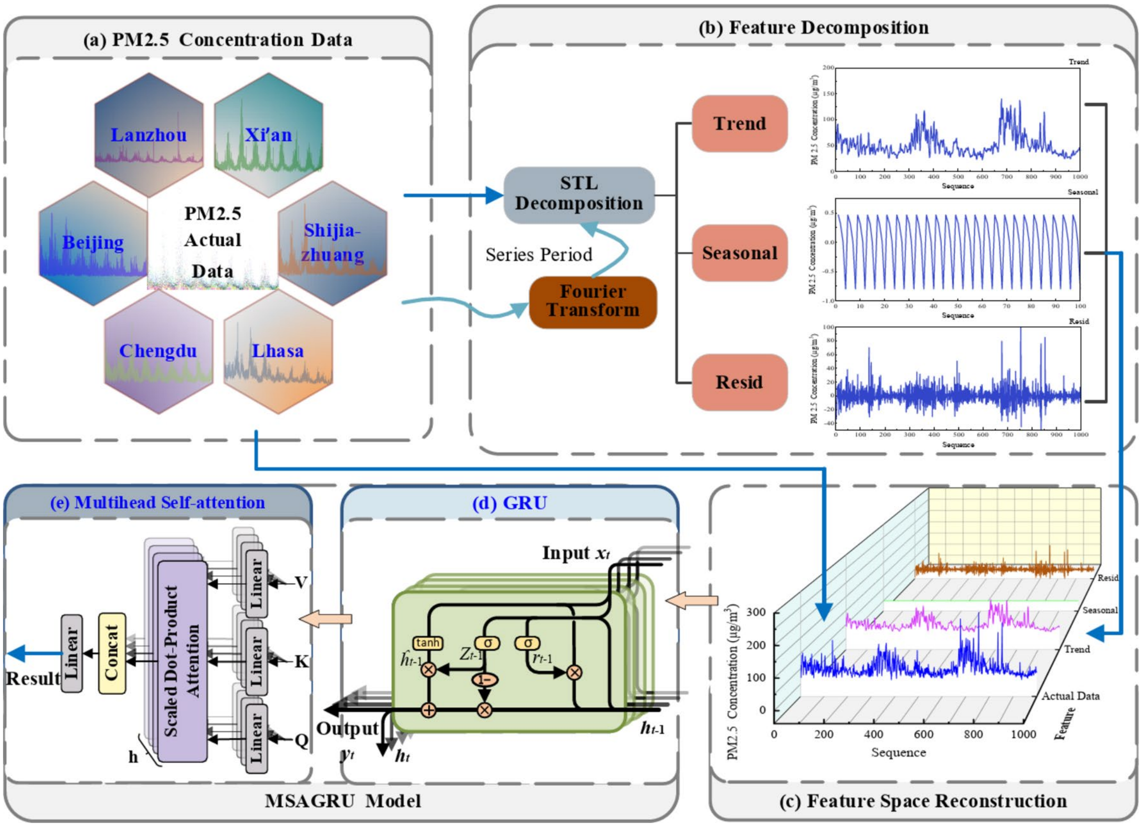


Fig. 6. Framework of the FSR-MSAGRU model.

and algorithmic steps. The FSR-MSAGRU model is designed to predict future PM2.5 concentration levels by effectively leveraging historical PM2.5 data while capturing both short-term temporal dependencies and long-range correlations. Its mathematical formulation is presented in Eq. (3), offering a formal representation of the underlying predictive mechanism. Furthermore, the overall framework of the FSR-MSAGRU model is illustrated in Fig. 6, demonstrating the key components and the flow of information within the proposed architecture.

$$\begin{aligned}\mathbf{H}_t &= f_{GRU}(\mathbf{X}'_{t-n}, \mathbf{X}'_{t-n+1}, \cdots, \mathbf{X}'_{t-1}) \\ \mathbf{Z}_t &= f_{MSA}(\mathbf{H}_{t-m}, \mathbf{H}_{t-m+1}, \cdots, \mathbf{H}_t) \\ \mathbf{X}'_{t+\mathrm{h}} &= f(\mathbf{Z}_t)\end{aligned} \tag{3}$$

where $\mathbf{X}'_{t-n}, \mathbf{X}'_{t-n+1}, \cdots, \mathbf{X}'_{t-1}$ represent the PM2.5 data reconstructed in the feature space over the past time steps, $\mathbf{H}_{t-m}, \mathbf{H}_{t-m+1}, \cdots, \mathbf{H}_t$denote the hidden states extracted by the GRU model, capturing short-term temporal features over time steps. $\mathbf{Z}_t$ represents the features processed by the multihead self-attention mechanism, which captures long-range dependencies. The predicted PM2.5 concentration for h time steps ahead is denoted as $\mathbf{X}'_{t+\mathrm{h}}$. The functions $f_{GRU}(\cdot)$ and $f_{MSA}(\cdot)$ respectively correspond to the nonlinear mapping functions of the GRU model and the multihead self-attention mechanism, respectively, while $f(\cdot)$ represents the final fully connected layer mapping function.

The FSR-MSAGRU model comprises two primary components, feature space reconstruction and multihead self-attention GRU. During feature space reconstruction, the raw PM2.5 sequence data undergo a frequency spectrum analysis using Fourier transforms to identify periodicity, reducing the impact of human factors on the model's parameter settings. Next, the STL is applied to extract the periodicity and trend information from the PM2.5 sequence data, mitigating the influence of noise on the prediction results. Finally, the raw PM2.5 sequence data, along with the extracted seasonal, trend, and residual components, collectively constitute a new feature space to enhance the representation of the sequence data features. During the model training and prediction process, the internal update and reset gates of the GRU are utilized to regulate the information flow, effectively capturing the dynamic characteristics of sequential data. The features extracted by the GRU are subsequently processed by the multihead self-attention mechanism, allowing multiple key information points within the GRU output to be attended to simultaneously. Different attention weights are assigned to these points, enabling the capture of long-range dependencies across time steps in the sequence data and enhancing the extraction of global correlation information. Finally, the processed features are passed through a fully connected layer, where the final output is generated.

Methodologies relevant to the proposed model

Gated recurrent unit

The gated recurrent unit (GRU) is a type of recurrent neural network that was developed as an improvement upon the LSTM network in 2014[67]. The GRU comprises two gating units, the reset gate and the update gate. Compared to the LSTM network, it offers higher computational efficiency. The gating process of the GRU in Fig. 6(d) is as follows:

At time t, the input to the GRU is $\S_t$, and the hidden state from the previous time step is $\langle_{t-1}$. The states of the reset gate and the update gate are as follows:

$$\nabla_t = \sigma\left(\mathbf{W}_r \cdot [\langle_{t-1}, \S_t] + \mathbf{b}_r\right) \tag{4}$$

$$\ddagger_t = \sigma\left(\mathbf{W}_z \cdot [\langle_{t-1}, \S_t] + \mathbf{b}_z\right) \tag{5}$$

where ∇_t represents the output of the reset gate, $\mathbf{W}_r$ is the weight matrix corresponding to the reset gate, $[\langle_{t-1}, \S_t]$ is the input vector formed by concatenating $\langle_{t-1}$ and $\S_t$, $\mathbf{b}_r$ is the bias matrix of the reset gate, $\ddagger_t$ represents the output of the update gate, $\mathbf{W}_z$ is the weight matrix corresponding to the update gate, $\mathbf{b}_z$ is the bias matrix of the update gate, and σ is the activation function.

The candidate hidden state $\langle_t{}'$ is calculated as follows:

$$\langle_t{}' = \tanh\left(\mathbf{W}_h \cdot [\nabla_t \otimes \langle_{t-1}, \S_t] + \mathbf{b}_h\right) \tag{6}$$

where $\mathbf{W}_h$ represents the weight matrix corresponding to the candidate hidden state, $\otimes$ denotes the Hadamard product operation, $\mathbf{b}_h$ represents the bias matrix corresponding to the hidden state, and $\langle_t{}'$ denotes the hyperbolic tangent activation function.

The current time step hidden layer output $\langle_t$ $(\dagger_t)$ is given by:

$$\langle_t = (1 - \ddagger_t) \otimes \langle_{t-1} + \ddagger_t \otimes \langle'_t \tag{7}$$

As indicated by Eq. (5), the update gate's output $\ddagger_t$ is controlled to retain the information from the previous time step until the current time step. When $\ddagger_t$ approaches 1, the output mainly consists of the candidate hidden state $\langle_t{}'$ at the current time step. Conversely, when $\ddagger_t$ approaches 0, the output mainly comprises the hidden state $\langle_{t-1}$ from the previous time step. Thus, the update gate and reset gate determine which data from the previous neuron input should be forgotten and which should be updated to the current state at this time step.

Multihead self-attention

The multihead self-attention mechanism is a variant of the self-attention mechanism developed to enhance the expressive and generalization capabilities of the model[43]. It utilizes multiple independent self-attention heads in parallel to capture information from different subspaces of the input sequence. Each head calculates attention weights separately, and the results are then concatenated or weight summed to obtain a richer feature representation[43,68].

The attention function of the MSAGRU model in Fig. 6(e) is represented as follows:

$$Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_k}}\right)\mathbf{V} \tag{8}$$

In the multihead self-attention mechanism, **Q**, **K**, and **V** are linearly transformed independently *h* times using different parameter matrices for each head. Then, the transformed **Q**, **K**, and **V** are input into *h* parallel heads to perform the attention function operation described above. Each parallel head can capture the unique feature information of the data in different representation subspaces within the sequence. Finally, the results from the *h* parallel heads are concatenated and linearly transformed to obtain the final output.

Model evaluation metrics

To comprehensively and objectively evaluate the performance of the model predictions, this study selected the performance metrics, root mean square error (RMSE), mean absolute error (MAE), mean bias error (MBE), symmetric mean absolute percentage error (SMAPE), Pearson correlation coefficient (PCC), and direction accuracy of the forecasting results (DA). The expressions for each metric are shown in Eq. (9) to (14) .

$$RMSE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left(y_t - y'_t\right)^2} \tag{9}$$

$$MAE = \frac{1}{m}\sum_{t=1}^{m}\left|y_t - y'_t\right| \tag{10}$$

$$SMAPE = \frac{100\%}{m}\sum_{t=1}^{m}\frac{\left|y_t - y'_t\right|}{\left(\left|y_t\right| + \left|y'_t\right|\right)/\,2} \tag{11}$$

$$PCC = \frac{\sum_{t=1}^{m}\left(y_t - \overline{y}\right)\left(y'_t - \overline{y'_t}\right)}{\sqrt{\sum_{t=1}^{m}\left(y_t - \overline{y}\right)^2 \sum_{t=1}^{m}\left(y'_t - \overline{y'_t}\right)^2}} \tag{12}$$

$$DA = \frac{1}{m-1}\sum_{t=1}^{m-1} 1\left(\mathrm{sign}\left(y'_{t+1} - y'_t\right) = \mathrm{sign}\left(y_{t+1} - y_t\right)\right) \tag{13}$$

$$MBE = \frac{1}{m}\sum_{t=1}^{m}\left(y_t - y'_t\right) \tag{14}$$

where y_t and y'_t represent the actual observed data and the predicted result at time *t*, respectively; *m* represents the length of the test sets; $\overline{y}$ and $\overline{y'}$ represent the mean values of the actual observed data and the predicted results, respectively; y_{t+1} and y'_{t+1} represent the one-dimensional observed data and predicted result at time $t+1$, respectively; and "1" is the indicator function.

The RMSE, MAE, and SMAPE are used to measure the difference between the model-predicted results and the actual observed data, where smaller values indicate a higher prediction accuracy and better performance. The PCC is used to assess the linear correlation between the predicted results and the actual observed data, with values ranging from -1 to 1. The closer the absolute value of the PCC is to 1, the stronger the linear correlation between the predicted results and the actual observed data. DA is used to evaluate the accuracy of the model in capturing the direction of data trends, which is measured by judging the consistency of the trends between the consecutive time steps, with values ranging from 0 to 1. The MBE measures the systematic bias of the model, with positive values indicating a tendency to overestimate the actual observed data and the negative values indicating underestimations. However, due to the potential cancellation of positive and negative biases, the MBE alone does not fully represent the extent of systematic bias. Therefore, in this study, the MBE is considered only as a supplementary metric for assessing model prediction bias. The numerical value of MBE and its impact on the experimental results are not discussed in this study.

Code availability

The code is available at https://github.com/zeroyi123/PM2.5-Concentrations-Prediction/tree/master.

Experimental results and analysis

In this section, the effectiveness of the proposed FSR-MSAGRU model in predicting PM2.5 is validated through ablation experiments and comparative experiments using PM2.5 from different regions.

Experimental parameter settings

The numerical experiments in this study were conducted on a computer with a 2.50 GHz Intel(R) Core(TM) i7-11700 CPU and 16.0 GB of RAM running the Windows 11 64-bit operating system. The development environment utilized was PyCharm 2022.3.2 (Community Edition), with the Python environment version 3.7.16.

In the experiments, all the neural network models were built using the Keras framework, with the main model comprising 128 neurons. The mean squared error (MSE) was employed as the loss function, and the Adam optimizer was selected. During model training, the maximum number of epochs was set to 20,000 to ensure sufficient learning of the model. To balance the memory usage and model update frequency, the batch size was set to 64. Additionally, the early stopping technique was applied to monitor the training progress and prevent overfitting. Specifically, training was stopped if the validation loss did not improve within 500 consecutive epochs. For enhanced transparency and traceability of the training process, CSV-Logger callback was used to record detailed performance metrics for each training epoch. Moreover, the model checkpoint mechanism was employed to monitor the validation loss on the validation set. Whenever an improvement in model performance was detected, the current best-performing model was saved, facilitating subsequent model evaluation, deployment, and invocation. All the other model parameters were kept at their default values. It is important to note that the final experimental results for all models are evaluated based on the average of five predictions from the best-performing model.

Experiment I: ablation experiment of the proposed model

In this section, the contributions of each component of the FSR-MSAGRU model (i.e., the multihead attention mechanism, STL data preprocessing, and feature space reconstruction) to improving the PM2.5 prediction performance are systematically evaluated through ablation experiments. Through individual comparisons of the prediction results from the baseline GRU model, the MSAGRU model, the STL data preprocessing-based GRU model (STL-GRU), the STL data preprocessing-based MSAGRU model (STL-MSAGRU), the feature space reconstruction-based GRU model (FSR-GRU), and the feature space reconstruction-based MSAGRU model (FSR-MSAGRU), the influence of the addition of each component on the prediction results is analyzed. The performance metrics of the baseline and improved models in predicting results are presented in Table 4.

The comparative curve of the prediction results from the ablation experiments is shown in Fig. 7.

By combining the comparison curves of the PM2.5 prediction results on different datasets and the performance metrics of the GRU, STL-GRU, MSAGRU, STL-MSAGRU, FSR-GRU and FSR-MSAGRU models in Table 4 and Fig. 7, the following conclusions can be drawn:

(1) The inclusion of the multihead self-attention mechanism can slightly enhance the prediction performance of the baseline model.

Table 4 shows that the improved MSAGRU model outperforms the baseline GRU model. Across the 6 different experimental datasets, the RMSE decreases, and the MAE decreases in all the datasets except for Shijiazhuang and Chengdu, where it increases slightly. The MSAGRU exhibits a better overall prediction accuracy than the GRU. The SMAPE shows similar values between the GRU and the MSAGRU, indicating a similar performance in terms of relative error. In terms of model fitting, except for Beijing, the PCC values of the MSAGRU model are better than those of the GRU model. However, the DA values decrease in the four datasets, indicating that the MSAGRU performs slightly worse in capturing the trend direction of the data, which may be attributed to the attention mechanism focusing on the global correlation properties, which sometimes weakens the dependency between the individual sequence steps. As shown in Table 4, among the 30 primary evaluation metrics across 6 different experimental datasets, the MSAGRU model outperformed the GRU model, achieving the best performance in 22 metrics. However, the inclusion of the multihead self-attention mechanism does not significantly improve prediction performance compared to the data decomposition and feature space reconstruction modules.

The impact of the multihead self-attention mechanism on improving model predictive performance varies with datasets complexity. For example, when using STL-MSAGRU model to predict PM2.5 concentrations, the data decomposition method effectively reduces sequence complexity, limiting the contribution of long-term dependencies captured by the multihead self-attention mechanism to performance improvement. As shown in Table 4, the performance gains from the multihead self-attention mechanism are more pronounced in the more complex Lanzhou dataset compared to lower-complexity datasets. Moreover, after feature space reconstruction, the PM2.5 sequence data undergo dimensional expansion, resulting in a more complex feature representation compared to the purely decomposed sequence. In this case, the long-term dependencies extracted by the multihead self-attention mechanism contribute more effectively to enhancing predictive performance. As shown in Table 4, the FSR-MSAGRU model generally outperforms the FSR-GRU model.

(2) The STL preprocessing of the sequence data can significantly enhance the prediction performance of the model.

From the metrics listed in Table 4, it can be observed that the improvement in prediction results is more pronounced in datasets such as Lanzhou, Xi'an, Shijiazhuang, and Chengdu, where the performance metrics RMSE, MAE, and SMAPE are reduced by approximately half. The improvement in model fitting is even more significant, with the PCC values all exceeding 0.94, specifically 0.9475, 0.9614, 0.9493, and 0.9596; moreover, the DA values also increased by more than 0.25. In the Beijing and Lhasa experimental datasets, after preprocessing the sequence data using the STL, the RMSE, MAE, and SMAPE of the MSAGRU data decreased by 28.85%, 29.40%, and 17.75%, respectively, and by 11.79%, 18.86%, and 22.90%, respectively. The fitting metrics PCC and DA are improved by 45.23% and 18.18% and 4.25% and 58.00%, respectively. In conclusion, on all 6 of the different experimental datasets, the performance metrics of the MSAGRU model prediction results are significantly improved after preprocessing the raw PM2.5 sequence data using the STL algorithm.

(3) The predictive performance of the model can be significantly enhanced by feature space reconstruction.

Figure 7 shows that on all 6 different experimental datasets, the prediction curves of the proposed FSR-MSAGRU model are closer to the actual data curve than are those of the GRU, MSAGRU, and STL-MSAGRU

Dataset	Model	RMSE /(μg/m^3)	MAE /(μg/m^3)	SMAPE /(%)	PCC	DA	MBE
Lanzhou	GRU	11.6285	6.6551	18.80%	0.7148	0.4094	-0.1408
	STL-GRU	5.3648	3.0980	9.24%	0.9460	0.7464	0.1726
	MSAGRU	11.2284	6.5533	18.94%	0.7309	0.4275	0.0151
	STL-MSAGRU	5.2494	3.2411	9.97%	0.9475	0.6957	-0.1645
	FSR-GRU	0.6086	0.3268	1.13%	0.9993	0.9855	-0.0499
	FSR-MSAGRU	0.5246	0.1875	0.63%	0.9995	0.9928	-0.0936
Xi’an	GRU	17.9032	12.3766	31.24%	0.8202	0.5181	-0.8399
	STL-GRU	8.8417	5.7235	15.03%	0.9597	0.7935	-0.6880
	MSAGRU	17.6251	12.3692	31.47%	0.8282	0.4891	0.9421
	STL-MSAGRU	8.6651	5.8252	15.41%	0.9614	0.7862	-0.0458
	FSR-GRU	0.5656	0.4236	1.71%	0.9998	0.9783	-0.0035
	FSR-MSAGRU	0.3301	0.2259	0.75%	0.9999	0.9964	0.0344
Beijing	GRU	21.1930	14.4576	51.84%	0.5503	0.5882	2.6320
	STL-GRU	14.7382	10.1124	41.86%	0.8031	0.6360	1.2777
	MSAGRU	20.9408	14.4503	51.84%	0.5489	0.5662	1.1008
	STL-MSAGRU	14.8991	10.2026	42.63%	0.7972	0.6691	-0.7834
	FSR-GRU	4.2872	3.0766	17.06%	0.9850	0.9191	-0.4729
	FSR-MSAGRU	4.0701	2.9749	16.33%	0.9863	0.9081	-0.0455
Shijiazhuang	GRU	17.1549	12.8019	32.49%	0.7490	0.5507	2.0505
	STL-GRU	7.6280	5.7090	16.21%	0.9561	0.7971	-0.1310
	MSAGRU	17.0979	12.9560	33.24%	0.7493	0.5543	1.8308
	STL-MSAGRU	8.0901	6.1116	17.22%	0.9493	0.8080	-0.5207
	FSR-GRU	1.0322	0.8166	2.53%	0.9993	0.9928	-0.2723
	FSR-MSAGRU	0.1681	0.1257	0.45%	0.9999	0.9999	0.0111
Chengdu	GRU	11.6151	8.4413	26.39%	0.7878	0.5181	1.3740
	STL-GRU	5.1486	3.9717	14.09%	0.9647	0.7754	0.8519
	MSAGRU	11.4977	8.6241	27.27%	0.7944	0.5109	-0.0929
	STL-MSAGRU	5.3238	4.0010	13.99%	0.9596	0.7935	0.0596
	FSR-GRU	0.4492	0.3514	1.31%	0.9997	0.9783	0.0970
	FSR-MSAGRU	0.1854	0.1195	0.53%	1.0000	0.9855	0.0272
Lhasa	GRU	2.1926	1.6367	22.06%	0.8314	0.3824	0.3797
	STL-GRU	1.4493	1.0628	13.52%	0.9256	0.6029	0.0198
	MSAGRU	2.1333	1.5813	21.12%	0.8367	0.3676	0.3406
	STL-MSAGRU	1.8817	1.2830	16.28%	0.8722	0.5809	-0.1272
	FSR-GRU	0.8170	0.6331	10.19%	0.9776	0.8125	0.1131
	FSR-MSAGRU	0.5615	0.4099	6.61%	0.9897	0.8493	-0.1322

Table 4. Performance metrics of ablation experiment prediction results.

models, demonstrating the excellent prediction capability of the FSR-MSAGRU model. Compared to the prediction results of the MSAGRU and STL-MSAGRU models, the FSR-MSAGRU model exhibits outstanding performance in datasets such as Lanzhou, Xi’an, Shijiazhuang, and Chengdu, with a reduction of more than 90% in the performance metrics RMSE, MAE, and SMAPE. In the Beijing and Lhasa experimental datasets, the RMSE, MAE, and SMAPE of the FSR-MSAGRU model are reduced by at least 59% compared to those of the STL-MSAGRU model and by at least 68% compared to those of the MSAGRU model. Additionally, the FSR-MSAGRU model consistently demonstrated high PCC values exceeding 0.98 across all the datasets, indicating a strong positive correlation between the prediction results and the actual data. Despite slightly lower DA metrics in the Beijing and Lhasa datasets, the FSR-MSAGRU model maintains robust performances, validating its effectiveness in PM2.5 prediction. The experimental results above validate the superiority of the proposed FSR-MSAGRU model in predicting PM2.5, fully illustrating that feature space reconstruction significantly enhances the model’s predictive performance.

Experiment II: comparative experiment of the proposed model with other models

The purpose of the experiments in this section is to verify the effectiveness of the proposed FSR-MSAGRU model in predicting PM2.5 through comparative experiments with mainstream single models and hybrid models. The selected mainstream single models include the convolutional neural network (CNN), Elman regression neural network, LSTM, and BiLSTM[35]. The hybrid models include the convolutional neural network and gated recurrent unit hybrid model (CNN-GRU)[56], the GRU based on complete ensemble empirical mode decomposition with

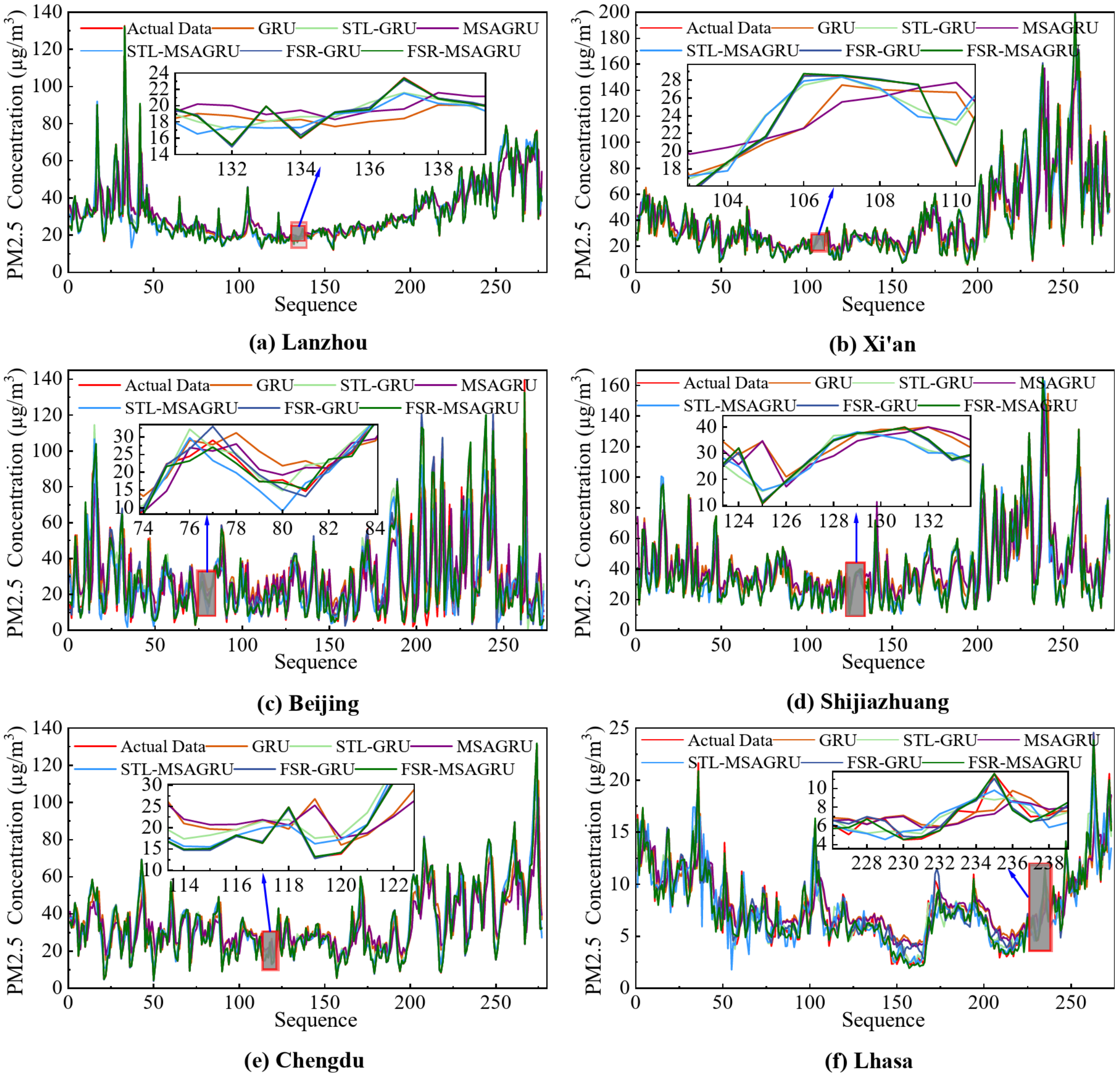


Fig. 7. Comparison curves of the ablation experiment prediction results.

adaptive noise (CEEMDAN-GRU), the LSTM hybrid model based on the STL (STL-LSTM), the LSTM hybrid model based on second-order trend decomposition with the STL (STL2-LSTM), the integrated 3D-CNN and GRU network model (3D-CNN-GRU)[36], and the ensemble empirical mode decomposition, attention mechanism and long short-term memory network model (EEMD-ALSTM)[69]. The performance metrics of the PM2.5 prediction results on the Lanzhou dataset are presented in Table 5. The comparative curves of the PM2.5 prediction results and model fitting scatter plots for single models and hybrid models with the FSR-MSAGRU model on the Lanzhou dataset are shown in Fig. 8.

The variations in the prediction performance metrics of the different models on various datasets are depicted in Fig. 9 through radar charts, histograms, scatter plots, etc.

Detailed comparisons and analyses are conducted between the PM2.5 prediction results of the aforementioned single models, hybrid models, and the FSR-MSAGRU model on 6 different datasets. The following conclusions can be drawn:

(1) The prediction results of the FSR-MSAGRU model outperform those of the mainstream single models.

By comparing the scatter plots of the prediction values versus the actual values between the single models and the FSR-MSAGRU model shown in Fig. 8, it is evident that on the Lanzhou dataset, the scatter plots between the prediction values and actual values of the FSR-MSAGRU model are more tightly clustered around the ideal line. Furthermore, from the curves of the prediction results and their locally magnified views, it is visually apparent that the consistency between the prediction curve of the FSR-MSAGRU model and the actual value curve is significantly greater than that of the single models. As demonstrated in Table 5, a comprehensive comparison of the performance metrics on the Lanzhou dataset between the single model and the FSR-MSAGRU model reveals that the FSR-MSAGRU model exhibits superior prediction accuracy and model fitting performance. The prediction results on the remaining 5 experimental datasets, as depicted in Fig. 9, similarly demonstrate that the

Dataset	Model	RMSE /(μg/m^3)	MAE /(μg/m^3)	SMAPE /(%)	PCC	DA	MBE
Lanzhou	CNN	11.9206	7.4553	21.65%	0.6921	0.4058	1.1967
	Elman	11.4293	6.5316	18.66%	0.7205	0.4746	-0.2394
	LSTM	11.5811	6.9034	19.73%	0.7193	0.4167	1.2016
	BiLSTM[35]	11.3894	6.4826	18.63%	0.7231	0.4275	-0.4265
	GRU	11.6285	6.6551	18.80%	0.7148	0.4094	-0.1408
	MSAGRU	11.2284	6.5533	18.94%	0.7309	0.4275	0.0151
	CNN-GRU[56]	11.7766	7.3760	21.55%	0.6970	0.4565	0.6116
	CEEMDAN-GRU	7.1517	4.2856	13.67%	0.9003	0.7174	-0.2237
	STL-MSAGRU	5.2494	3.2411	9.97%	0.9475	0.6957	-0.1645
	STL-LSTM	6.0625	3.3259	9.64%	0.9302	0.7428	-0.0977
	STL2-LSTM	3.8915	2.4367	8.16%	0.9722	0.7790	-0.0390
	3D CNN-GRU[36]	10.5887	6.8007	19.80%	0.7749	0.5797	1.6954
	EEMD-ALSTM[69]	5.6154	4.0844	14.74%	0.9432	0.7246	-1.3517
	FSR-MSAGRU	**0.5246**	**0.1875**	**0.63%**	**0.9995**	**0.9928**	-0.0936

Table 5. Performance metrics of different models (Lanzhou).

FSR-MSAGRU model achieves the optimal predictive performance. These experimental results demonstrate that the proposed FSR-MSAGRU model outperforms mainstream single models in terms of prediction performance.

(2) The prediction performance of the FSR-MSAGRU model surpasses that of the experimental reference hybrid models.

From Fig. 8, it is evident that the scatter plots of the hybrid models demonstrate a more centralized distribution around the actual values on the Lanzhou dataset, with the prediction trajectories of most hybrid models closely aligning with the actual data. Nonetheless, in contrast to the prediction performance of the FSR-MSAGRU model, the prediction performance of hybrid models still exhibits certain shortcomings. A comprehensive analysis of the performance metrics of the prediction results between the hybrid models and the FSR-MSAGRU model, as shown in Fig. 9, indicates that the hybrid models CEEMDAN-GRU, STL-LSTM, STL2-LSTM and EEMD-ALSTM demonstrate certain prediction advantages on the 6 different experimental datasets. However, the overall prediction effectiveness of these hybrid models still falls short of that of the FSR-MSAGRU model. These experimental results demonstrate that the prediction performance of the proposed FSR-MSAGRU model for PM2.5 surpasses that of all the other experimental reference hybrid models.

Discussion of the experiments

Discussion on model performance evaluation

Diebold–Mariano test

The Diebold–Mariano test is a common nonparametric statistical test used to compare the prediction accuracy of two or more models. Its main purpose is to examine whether there is a significant difference between the prediction errors of two models, thus determining the superiority or inferiority of the models. In this study, the Diebold–Mariano test is conducted using the RMSE as the loss function. The prediction errors of the FSR-MSAGRU model versus the actual values are used as the benchmark data to evaluate the effectiveness of the FSR-MSAGRU model. The DM statistic is negative if the reference model prediction performance is superior to that of the FSR-MSAGRU model, and positive if the FSR-MSAGRU model prediction performance is superior to that of the reference model. A larger absolute value of the DM statistic indicates a more significant difference in the prediction performance between the two models. The results of the Diebold–Mariano test for the 13 reference models on the 6 different experimental datasets are presented in Table 6.

Table 6 shows that at a significance level of 0.05, the Diebold–Mariano test rejects the null Hypothesis H0: there is no significant difference between the two models being tested. Therefore, there is a significant difference in the prediction performance between the FSR-MSAGRU model proposed in this study and all the reference models, with positive values of the DM statistic indicating that the prediction performance of the FSR-MSAGRU model is superior to that of all the reference models.

A comprehensive evaluation of the prediction results of the various models

To comprehensively evaluate the prediction results of each model, the statistical metrics (standard deviation, root mean square deviation (RMSD), and the correlation coefficient) of the different model prediction results are displayed through Taylor diagrams in this section. By comparing Taylor diagrams, the matching degree between the prediction results of the different models and the actual observations can be intuitively compared, enabling the assessment of differences in prediction performances among the various models. Taylor diagrams of the prediction results for the 14 models across 6 different datasets are shown in Fig. 10. In the Taylor diagrams depicted in Fig. 10, the red dashed circular arc represents the standard deviation of the actual observations. The closer the model prediction results are to the red dashed circular arc, the closer the variation range of the model prediction results is to the actual observations, indicating an improved consistency of the model. Moreover, the closer the model prediction results are to the "Actual" point marked in the diagram, the closer the

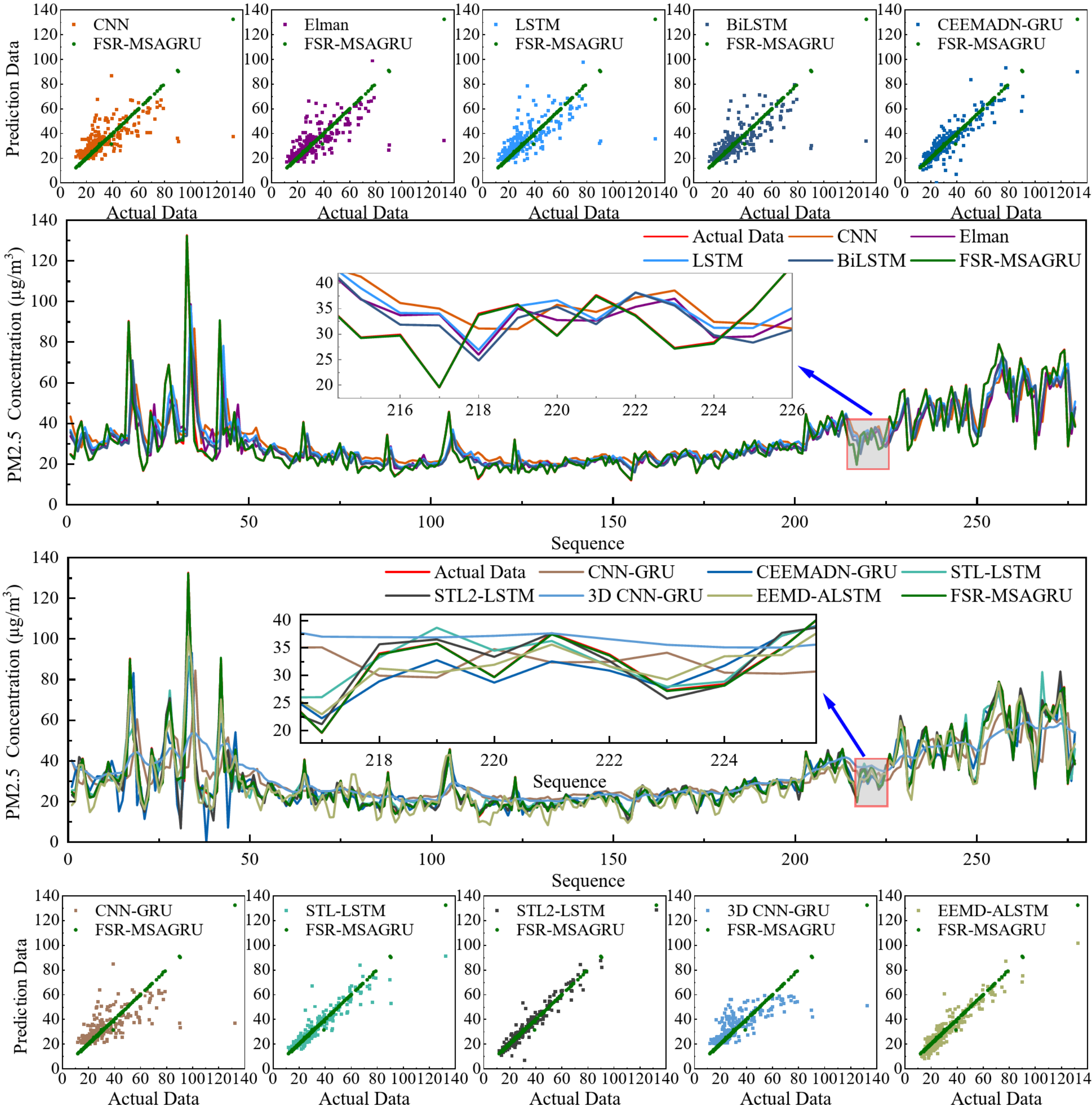


Fig. 8. Comparison curves and scatter plots of the single models, hybrid models, and FSR-MSAGRU model prediction results (Lanzhou).

model prediction results are to the actual observations, indicating a lower RMSD and higher correlation, thus demonstrating the superior prediction performance of the model.

Figure 10 shows that for the Lanzhou dataset, the correlation between the prediction results of the single models and the actual observations is generally weak. On the Xi'an dataset, the weakest correlation with the actual observations is demonstrated by the 3D CNN-GRU model. On the Beijing dataset, very similar correlations with the actual observations are observed for the CEEMDAN-GRU and STL2-LSTM models, the EEMD-ALSTM model achieved the second-best predictive performance among the 14 models. On the Shijiazhuang dataset, the STL2-LSTM model prediction results exhibit the highest standard deviation, indicating a greater variation range in its prediction results. On the Chengdu dataset, the predictive performance of the STL2-LSTM and EEMD-ALSTM models is essentially comparable. On the Lhasa dataset, the prediction performance of the EEMD-ALSTM model is only slightly inferior to that of the FSR-MSAGRU model. Overall, the FSR-MSAGRU model's performance across the 6 different experimental datasets is closest to the "Actual" point marked in the diagram. Compared to the other models, the FSR-MSAGRU model demonstrates better consistency and correlation with the actual observations, indicating that the prediction performance of the FSR-MSAGRU model is superior to that of all the other reference models across the different experimental datasets.

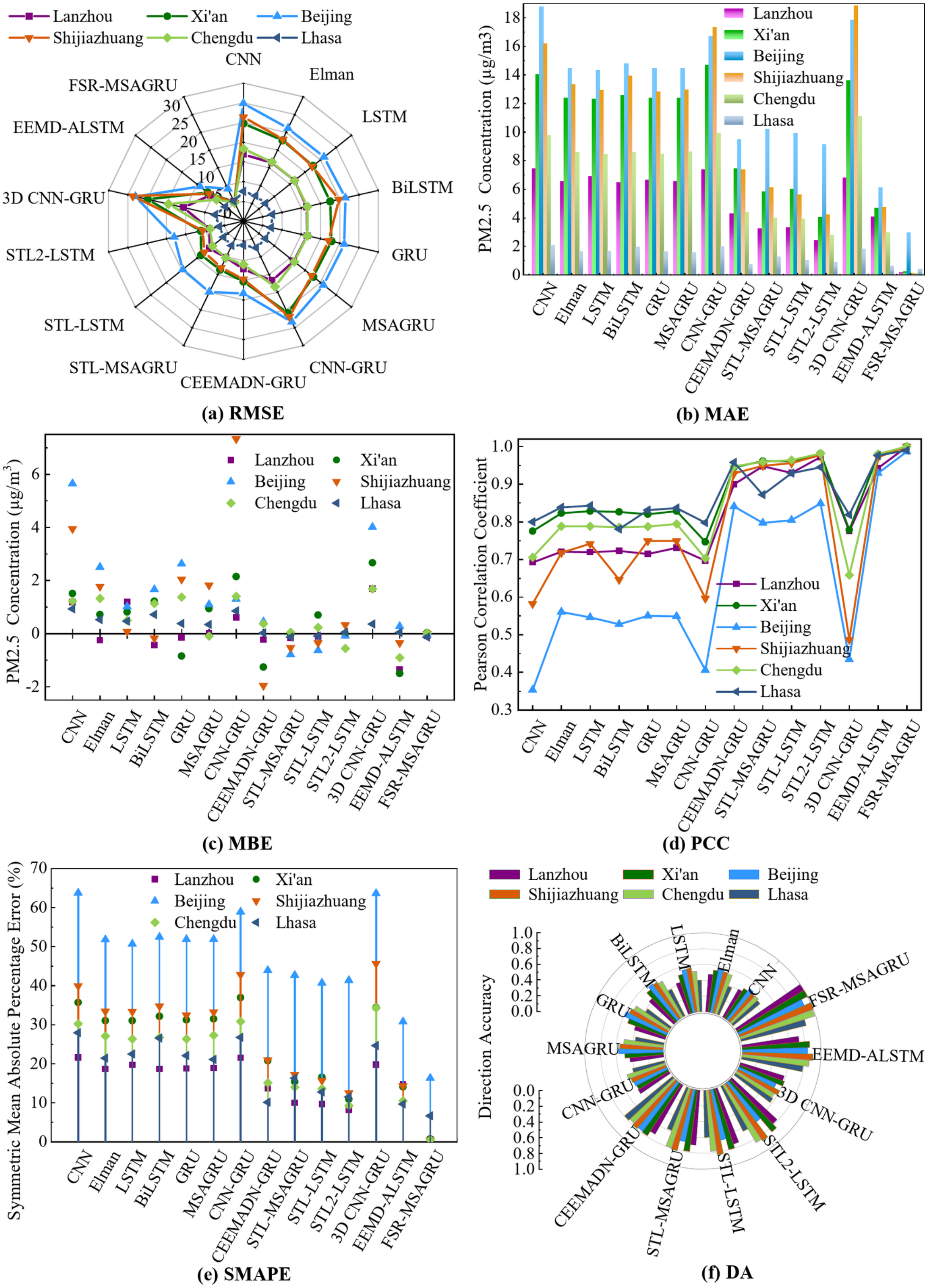


Fig. 9. Performance metrics of the different models on various datasets.

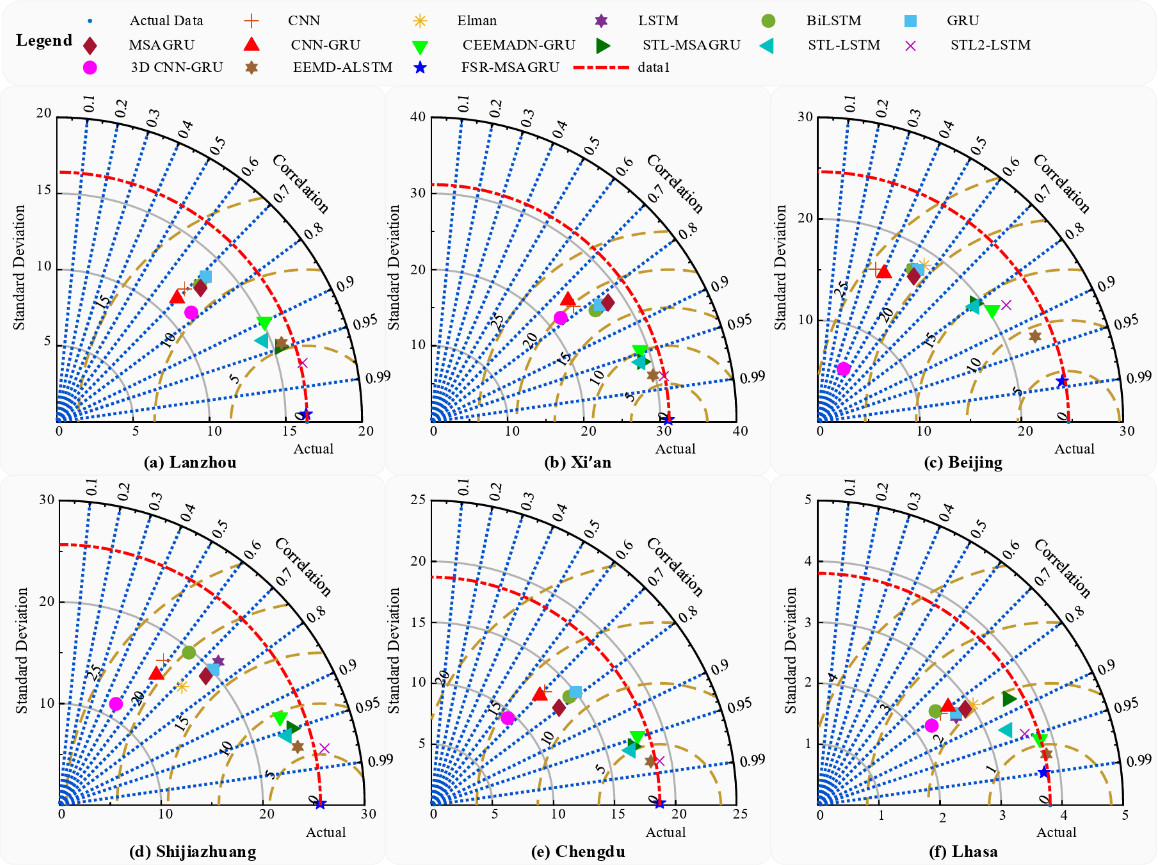
Discussion on single-model prediction performance and improvements

(1) None of the single models can maintain the best prediction performance.

From the performance metrics of the 5 single models depicted in Fig. 9, it is evident that none of the single models achieved an optimal performance across all the datasets. Specifically, on the Lanzhou dataset, the BiLSTM model attains the highest prediction accuracy. The LSTM and Elman models demonstrate relatively strong predictive performance on the Xi'an and Chengdu datasets. On the Beijing dataset, the Elman model outperforms the other models in terms of the RMSE and PCC. For the Shijiazhuang dataset, the GRU model

Model	Lanzhou	Xi'an	Beijing	Shijiazhuang	Chengdu	Lhasa
CNN	3.7492*	112.7258*	8.2718*	139.4179*	124.5073*	134.2444*
Elman	3.2273**	113.3857*	85.5030*	117.2315*	134.7368*	114.5253*
LSTM	3.4162*	113.5411*	84.9819*	129.8716*	130.1819*	109.3463*
BiLSTM[35]	3.1751**	119.2811*	81.3544*	92.6590*	134.8324*	125.0313*
GRU	3.2959*	104.1842*	83.5771*	142.4837*	130.0248*	106.4364*
MSAGRU	3.2685**	99.8772*	86.9596*	145.8403*	138.6332*	104.0910*
CNN-GRU [56]	3.6621*	102.6903*	7.1443*	9.8557*	120.9061*	129.3515*
CEEMDAN-GRU	73.6663*	118.2614*	89.3533*	148.6870*	110.4164*	76.3547*
STL-MSAGRU	74.0572*	92.9375*	77.3538*	126.1690*	6.7944*	96.1930*
STL-LSTM	51.8039*	94.8233*	4.5786*	118.6124*	8.2040*	94.8237*
STL2-LSTM	61.3682*	99.6880*	4.8423*	135.4397*	9.2011*	91.0683*
3D CNN-GRU[36]	4.1259*	92.2599*	120.6974*	139.8668*	128.5119*	99.4922*
EEMD-ALSTM[69]	79.9051*	107.8205*	77.1872*	138.4936*	10.4169*	74.4446*
FSR-AGRU	–	–	–	–	–	–

Table 6. Diebold–Mariano test results of the reference model and the proposed model on the different datasets. " *" indicates 0.01 significance level, " **" indicates 0.05 significance level.



(a) Lanzhou (b) Xi'an (c) Beijing (d) Shijiazhuang (e) Chengdu (f) Lhasa

Fig. 10. Taylor diagram of the prediction results of the different models on the various datasets.

outperforms the other models in terms of the RMSE, MAE, SMAPE, and PCC. The Elman model performs best on the Lhasa dataset in terms of prediction accuracy, but the LSTM model achieves the best correlation with the actual data. The above discussion illustrates that single models exhibit varied performances in PM2.5 prediction across the different regions, making it difficult to discern the strength of model generalization. Therefore, exploring methods to improve the prediction performance of single models is necessary.

(2) The improvement of single-model predictions necessitates appropriate methods.

As demonstrated in Sect. 4.3, the inclusion of the multihead self-attention mechanism can enhance the prediction performance of the GRU. By comparing the MSAGRU model with five single models—CNN, Elman,

Dataset	Decomposition method	RMSE /(μg/m^3)	MAE /(μg/m^3)	SMAPE /(%)	PCC	DA	MBE
Lanzhou	CEEMDAN-GRU	7.1517	4.2856	13.67%	0.9003	0.7174	-0.2237
	STL-GRU	5.3648	3.0980	9.24%	0.9460	0.7464	0.1726
Xi’an	CEEMDAN-GRU	10.2638	7.4251	20.74%	0.9452	0.7826	-1.2538
	STL-GRU	8.8417	5.7235	15.03%	0.9597	0.7935	-0.6880
Beijing	CEEMDAN-GRU	13.3516	9.4665	43.94%	0.8408	0.7721	0.4473
	STL-GRU	14.7382	10.1124	41.86%	0.8031	0.6360	1.2777
Shijiazhuang	CEEMDAN-GRU	9.7250	7.3547	20.97%	0.9287	0.7899	-1.9463
	STL-GRU	7.6280	5.7090	16.21%	0.9561	0.7971	-0.1310
Chengdu	CEEMDAN-GRU	5.9986	4.4002	15.13%	0.9475	0.7645	0.3757
	STL-GRU	5.1486	3.9717	14.09%	0.9647	0.7754	0.8519
Lhasa	CEEMDAN-GRU	1.0983	0.7557	10.11%	0.9583	0.7868	0.0307
	STL-GRU	1.4493	1.0628	13.52%	0.9256	0.6029	0.0198

Table 7. Comparison of the prediction results of the different decomposition methods.

LSTM, BiLSTM, and GRU—on 6 different experimental datasets, it was found that the MSAGRU model achieved the highest performance, obtaining 10 out of 30 optimal primary evaluation metrics in Fig. 9. The LSTM model, with 9 optimal metrics, performs the next best, while the GRU model achieves only 3 optimal metrics. This comparison indicates that the prediction performance of the MSAGRU model, enhanced by the inclusion of the multihead attention mechanism, is slightly better than that of the other mainstream single models, further reinforcing the experimental conclusion in Sect. 4.3. In contrast, the CNN-GRU model, an improvement of the CNN and GRU hybrid, not only fails to improve the prediction performance of the baseline model but also performs worse than the single models. Therefore, in terms of improving the single models, it is necessary to select appropriate methods for model improvement, or to choose suitable hybrids of the single models to enhance the prediction performance; otherwise, these methods may have adverse effects.

Discussion on the impact of data preprocessing on model performance

The experimental results in Sect. 4.3 demonstrate that preprocessing the input data using the STL can lead to a significant improvement in the prediction performance of the model. To validate the influence of data preprocessing on the prediction results, the CEEMDAN data decomposition method, the EEMD data decomposition method, and the method of second-order trend decomposition with the STL, were incorporated into this study for comparison. By comparing the prediction results of models incorporating data decomposition (CEEMDAN-GRU, EEMD-ALSTM, STL-MSAGRU, STL-LSTM, and STL2-LSTM) with those of single models (CNN, Elman, LSTM, BiLSTM, GRU, and CNN-GRU), as depicted in Fig. 9, it can be observed that the prediction results of the four models incorporating data decomposition preprocessing are superior to those of any single model. Therefore, data decomposition preprocessing can effectively enhance the prediction performance of the main model.

To verify the performance of the STL method, this study preprocessed the raw sequence data using both the CEEMDAN and the STL methods on 6 different experimental datasets. The GRU model was then used to predict the preprocessed data. The performance metrics of the GRU model using the two different data decomposition methods are presented in Table 7.

Table 7 shows that on the datasets of Lanzhou, Xi’an, Shijiazhuang, and Chengdu, the GRU model using the STL data decomposition method (STL-GRU) outperforms the GRU model using the CEEMDAN data decomposition method (CEEMADN-GRU). However, on the Lhasa dataset, the CEEMADN-GRU model performed better. In the Beijing dataset, except for the SMAPE indicator, the CEEMADN-GRU model outperforms the STL-GRU model in the remaining four metrics. Analyzing the principles and processes of the two decomposition methods, it can be observed that the reason for the inferior performance of the STL-GRU model compared to the CEEMADN-GRU model on the datasets of Beijing and Lhasa lies in the fact that the STL data decomposition method for these datasets is set to 8 days. During data processing, the STL method may overlook or fail to adequately capture periodic fluctuations with a period length shorter than 8 days, thereby influencing the subsequent predictive performance of the GRU model.

Conclusion

Accurate predictions of PM2.5 concentrations are vital for safeguarding public health and addressing environmental issues. However, machine learning models often overlook the periodic and global features of the sequence data when predicting PM2.5. In response to these challenges, a novel hybrid data-driven modeling approach to PM2.5 concentrations prediction, termed the FSR-MSAGRU model, which excels in improving feature representations through feature space reconstructions and capturing global features, is proposed in this study. The FSR-MSAGRU model demonstrated excellent prediction performance across 6 distinct experimental datasets. The conclusions drawn from this study are summarized as follows:

(1) The incorporation of the multihead self-attention mechanism can enhance the prediction performance of the GRU. The MSAGRU model, which incorporates perception global feature information, was proposed in this study by innovatively extending the basic gated recurrent unit through the introduction of a multihead

self-attention mechanism. Through parallel processing via multiple self-attention heads, the model achieves a simultaneous focus on various crucial pieces of information within the GRU. Additionally, it learns distinct attention weights among these information points, facilitating the capture of long-range dependencies across the sequence data within the GRU.

(2) Data preprocessing, combined with the STL decomposition method, can enhance the prediction performance of the model. According to the data description analysis in, Sect. 2.1 PM2.5 sequence data exhibit pronounced periodic features, highlighting the crucial importance of accurately extracting the periodic features of the sequences during data preprocessing. Therefore, in this study, the STL data decomposition method is adopted to refine the raw sequence data into three subsequences, enabling precise capture of the periodic variations and trend directions of the PM2.5 sequence data, and effectively reducing the interference of the sequence data noise on the prediction results. Moreover, in determining the STL parameter values, Fourier transform analysis is applied to the PM2.5 sequence data to perform frequency domain analysis, providing accurate periodic parameter values for the STL based on a scientific analysis and reducing the biases introduced by manual parameter settings.

(3) The model prediction performance is significantly enhanced by feature space reconstruction. Although the method of separately predicting subsequences and then aggregating them to determine the final prediction results can yield a certain level of prediction accuracy, its limitation lies in the fact that the main model treats each subsequence as an independent entity, overlooking the correlations between the subsequences as well as between the subsequences and the raw data. Therefore, a feature space reconstruction method for data preprocessing is proposed in this study to enhance feature representation. The reconstructed feature space is expanded from one-dimensional feature vectors to four-dimensional feature vectors, enabling the model to capture more complex and hidden feature patterns and relationships.

(4) The FSR-MSAGRU model proposed in this study demonstrates promising predictive performance and generalizability. Six experimental datasets from different regions with varying levels of complexity were selected to verify the effectiveness of the proposed FSR-MSAGRU model. The SMAPE values of the results predicted by the FSR-MASGRU model are 0.63%, 0.75%, 16.33%, 0.45%, 0.53%, and 6.61%, representing reductions of 96.66%, 97.60%, 68.50%, 98.62%, 97.99%, and 70.05%, respectively, compared to those of the baseline model. The fitting performance metric PCC of the FSR-MASGRU model exceeds 0.98 for all the datasets. These results underscore the exceptional prediction performance of the proposed FSR-MSAGRU model across diverse datasets and its robust generalization capability.

The FSR-MSAGRU model proposed in this study provides an accurate and stable method for short-term PM2.5 concentrations prediction in different regions. In future research, comprehensive consideration will be given to multiple environmental factors, with an in-depth analysis of the interrelationships among the multifeature data. Effective multistep prediction methods will be explored to enhance the prediction capabilities over broader time ranges. Additionally, Fig. 9(c) shows that most neural network models exhibit a positive bias in PM2.5 prediction. This provides a direction for future research on error corrections in these models and warrants further investigation. Furthermore, fully utilizing the multi-head self-attention mechanism remains a limitation of this study and an important direction for future research.

Data availability

Data is provided within the manuscript or supplementary information files. The code is available at https://github.com/zeroyi123/PM2.5-Concentrations-Prediction/tree/master.

Received: 23 June 2024; Accepted: 2 May 2025
Published online: 16 May 2025

References

1. Wu, Z. et al. Prediction of air pollutant concentrations based on the long short-term memory neural network. *J. Hazard. Mater.* **465**, 133099. https://doi.org/10.1016/j.jhazmat.2023.133099 (2024).
2. Bryan, L. & Landrigan, P. PM(2.5) pollution in Texas: a Geospatial analysis of health impact functions. *Front. Public. Health.* **11**, 1286755. https://doi.org/10.3389/fpubh.2023.1286755 (2023).
3. Kim, K. H., Kabir, E. & Kabir, S. A review on the human health impact of airborne particulate matter. *Environ. Int.* **74**, 136–143. https://doi.org/10.1016/j.envint.2014.10.005 (2015).
4. Hill, W. et al. Lung adenocarcinoma promotion by air pollutants. *Nature* **616**, 159–167. https://doi.org/10.1038/s41586-023-05874-3 (2023).
5. Manisalidis, I., Stavropoulou, E., Stavropoulos, A. & Bezirtzoglou, E. Environmental and health impacts of air pollution: A review. *Front. Public. Health.* **8**, 14. https://doi.org/10.3389/fpubh.2020.00014 (2020).
6. Zhang, L. et al. Short-term and long-term effects of PM(2.5) on acute nasopharyngitis in 10 communities of Guangdong, China. *Sci. Total Environ.* **688**, 136–142. https://doi.org/10.1016/j.scitotenv.2019.05.470 (2019).
7. Pandya, A., Nanavaty, R., Pipariya, K., Shah, M. A. & Comparative Systematic study of machine learning (ML) approaches for particulate matter (PM) prediction. *Arch. Comput. Method E.* **31**, 595–614. https://doi.org/10.1007/s11831-023-09994-x (2023).
8. Zhu, J., Wu, P., Chen, H., Zhou, L. & Tao, Z. A. Hybrid forecasting approach to air quality time series based on endpoint condition and combined forecasting model. *Int. J. Environ. Res. Public. Health.* **15** https://doi.org/10.3390/ijerph15091941 (2018).
9. Huang, G., Li, X., Zhang, B. & Ren, J. PM2.5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. *Sci. Total Environ.* **768**, 144516. https://doi.org/10.1016/j.scitotenv.2020.144516 (2021).
10. Byun, D. & Schere, K. L. Review of the governing equations, computational algorithms, and other components of the Models-3 community multiscale air quality (CMAQ) modeling system. *Appl. Mech. Rev.* **59**, 51–77. https://doi.org/10.1115/1.2128636 (2006).
11. Yatsyshyn, T., Shkitsa, L., Popov, O. & Liakh, M. Development of mathematical models of gas leakage and its propagation in atmospheric air at an emergency gas well gushing. *Eastern-European J. Enterp. Technol.* **5**, 49–59. https://doi.org/10.15587/1729-4061.2019.179097 (2019).

12. Zhou, G. Q. et al. Numerical air quality forecasting over Eastern China: an operational application of WRF-Chem. *Atmos. Environ.* **153**, 94–108. https://doi.org/10.1016/j.atmosenv.2017.01.020 (2017).
13. Tesche, T. et al. CMAQ/CAMx annual 2002 performance evaluation over the Eastern US. *Atmos. Environ.* **40**, 4906–4919. https://doi.org/10.1016/j.atmosenv.2005.08.046 (2006).
14. Bey, I. et al. Global modeling of tropospheric chemistry with assimilated meteorology: model description and evaluation. *J. Geophys. Research: Atmos.* **106**, 23073–23095. https://doi.org/10.1029/2001jd000807 (2001).
15. Saide, P. E. et al. Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model. *Atmos. Environ.* **45**, 2769–2780. https://doi.org/10.1016/j.atmosenv.2011.02.001 (2011).
16. Gao, M. et al. Modeling study of the 2010 regional haze event in the North China plain. *Atmos. Chem. Phys.* **16**, 1673–1691. https://doi.org/10.5194/acp-16-1673-2016 (2016).
17. Wu, H. et al. Improving PM(2.5) forecasts in China using an initial error transport model. *Environ. Sci. Technol.* **54**, 10493–10501. https://doi.org/10.1021/acs.est.0c01680 (2020). Improving.
18. Zheng, B. et al. Heterogeneous chemistry: a mechanism missing in current models to explain secondary inorganic aerosol formation during the January 2013 haze episode in North China. *Atmos. Chem. Phys.* **15**, 2031–2049. https://doi.org/10.5194/acp-15-2031-2015 (2015).
19. Zhang, C., Wang, S., Wu, Y., Zhu, X. & Shen, W. A long-term prediction method for PM2.5 concentration based on Spatiotemporal graph attention recurrent neural network and grey Wolf optimization algorithm. *J. Environ. Chem. Eng.* **12**. https://doi.org/10.1016/j.jece.2023.111716 (2024).
20. Panneerselvam, B. et al. A novel approach for the prediction and analysis of daily concentrations of particulate matter using machine learning. *Sci. Total Environ.* **897**. https://doi.org/10.1016/j.scitotenv.2023.166178 (2023).
21. Pyae, T. S. & Kallawicha, K. First Temporal distribution model of ambient air pollutants (PM2.5, PM10, and O3) in Yangon City, Myanmar during 2019–2021. *Environ. Pollut.* **347**. https://doi.org/10.1016/j.envpol.2024.123718 (2024).
22. De Ridder, K., Kumar, U., Lauwaet, D., Blyth, L. & Lefebvre, W. Kalman filter-based air quality forecast adjustment. *Atmos. Environ.* **50**, 381–384. https://doi.org/10.1016/j.atmosenv.2012.01.032 (2012).
23. Gogikar, P., Tripathy, M. R., Rajagopal, M., Paul, K. K. & Tyagi, B. PM2.5 Estimation using multiple linear regression approach over industrial and non-industrial stations of India. *J. Amb Intel Hum. Comp.* **12**, 2975–2991. https://doi.org/10.1007/s12652-020-02457-2 (2020).
24. Kim, J., Wang, X., Kang, C., Yu, J. & Li, P. Forecasting air pollutant concentration using a novel Spatiotemporal deep learning model based on clustering, feature selection and empirical wavelet transform. *Sci. Total Environ.* **801**, 149654. https://doi.org/10.1016/j.scitotenv.2021.149654 (2021).
25. Lu, T. et al. National empirical models of air pollution using microscale measures of the urban environment. *Environ. Sci. Technol.* **55**, 15519–15530. https://doi.org/10.1021/acs.est.1c04047 (2021).
26. Luo, H. Y., Wang, D. Y., Yue, C. Q., Liu, Y. L. & Guo, H. X. Research and application of a novel hybrid decomposition-ensemble learning paradigm with error correction for daily PM forecasting. *Atmos. Res.* **201**, 34–45. https://doi.org/10.1016/j.atmosres.2017.10.009 (2018).
27. Liu, Z., Huang, X. & Wang, X. PM(2.5) prediction based on modified Whale optimization algorithm and support vector regression. *Sci. Rep.* **14**, 23296. https://doi.org/10.1038/s41598-024-74122-z (2024).
28. Guo, Q., He, Z. & Wang, Z. Simulating daily PM2. 5 concentrations using wavelet analysis and artificial neural network with remote sensing and surface observation data. *Chemosphere* **340**, 139886. https://doi.org/10.1016/j.chemosphere.2023.139886 (2023).
29. Ibrir, A., Kerchich, Y., Hadidi, N., Merabet, H. & Hentabli, M. Prediction of the concentrations of PM1, PM2.5, PM4, and PM10 by using the hybrid dragonfly-SVM algorithm. *Air Qual. Atmos. Health.* **14**, 313–323. https://doi.org/10.1007/s11869-020-00936-1 (2020).
30. Ayturan, Y. A. et al. Short-term prediction of PM2.5 pollution with deep learning methods. *Global Nest J.* **22**, 126–131. https://doi.org/10.30955/gnj.003208 (2020).
31. Chae, M., Han, S. & Lee, H. Outdoor particulate matter correlation analysis and prediction based deep learning in the Korea. *Electronics-Switz* **9** https://doi.org/10.3390/electronics9071146 (2020).
32. Wang, L. et al. A hybrid Spatiotemporal model combining graph attention network and gated recurrent unit for regional composite air pollution prediction and collaborative control. *Sustain. Cities Soc.* **116**. https://doi.org/10.1016/j.scs.2024.105925 (2024).
33. Paulpandi, C., Chinnasamy, M. & Nagalingam Rajendiran, S. Multi-Site air pollutant prediction using long short term memory. *Comput. Syst. Sci. Eng.* **43**, 1341–1355. https://doi.org/10.32604/csse.2022.023882 (2022).
34. Seng, D., Zhang, Q., Zhang, X., Chen, G. & Chen, X. Spatiotemporal prediction of air quality based on LSTM neural network. *Alex Eng. J.* **60**, 2021–2032. https://doi.org/10.1016/j.aej.2020.12.009 (2021).
35. Zhang, M., Wu, D. & Xue, R. Hourly prediction of PM2.5 concentration in Beijing based on Bi-LSTM neural network. *Multimed Tools Appl.* **80**, 24455–24468. https://doi.org/10.1007/s11042-021-10852-w (2021).
36. Faraji, M., Nadi, S., Ghaffarpasand, O., Homayoni, S. & Downey, K. An integrated 3D CNN-GRU deep learning method for short-term prediction of PM2.5 concentration in urban environment. *Sci. Total Environ.* **834**, 155324. https://doi.org/10.1016/j.scitotenv.2022.155324 (2022).
37. Gilik, A., Ogrenci, A. S. & Ozmen, A. Air quality prediction using CNN + LSTM-based hybrid deep learning architecture. *Environ. Sci. Pollut Res. Int.* **29**, 11920–11938. https://doi.org/10.1007/s11356-021-16227-w (2022).
38. Zhu, M. & Xie, J. Investigation of nearby monitoring station for hourly PM2.5 forecasting using parallel multi-input 1D-CNN-biLSTM. *Expert Syst. Appl.* **211**. https://doi.org/10.1016/j.eswa.2022.118707 (2023).
39. Galassi, A., Lippi, M. & Torroni, P. Attention in natural Language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4291–4308. https://doi.org/10.1109/TNNLS.2020.3019893 (2021).
40. Li, Y., Lu, G., Li, J., Zhang, Z. & Zhang, D. Facial expression recognition in the wild using Multi-Level features and attention mechanisms. *Ieee T Affect. Comput.* **14**, 451–462. https://doi.org/10.1109/taffc.2020.3031602 (2023).
41. Wu, F. et al. A novel hybrid model for hourly PM2.5 prediction considering air pollution factors, meteorological parameters and GNSS-ZTD. *Environ. Modell Softw.* **167**. https://doi.org/10.1016/j.envsoft.2023.105780 (2023).
42. Zheng, Q. et al. Application of complete ensemble empirical mode decomposition based multi-stream informer (CEEMD-MsI) in PM2.5 concentration long-term prediction. *Expert Syst. Appl.* **245**. https://doi.org/10.1016/j.eswa.2023.123008 (2024).
43. Tao, W. et al. EEG-Based emotion recognition via Channel-Wise attention and self attention. *Ieee T Affect. Comput.* **14**, 382–393. https://doi.org/10.1109/taffc.2020.3025777 (2023).
44. Dubey, G., Singh, H. P., Maurya, R. K., Sheoran, K. & Dhand, G. A hybrid forecasting system using convolutional-based extreme learning with extended elephant herd optimization for time-series prediction. *Soft Comput.* https://doi.org/10.1007/s00500-023-09499-6 (2024).
45. Mirzaei, S., Liao, T. L., Hsu, C. Y. & Modeling PM2.5 urbane pollution using hybrid models incorporating decomposition and multiple factors. *Urban Clim.* **60** https://doi.org/10.1016/j.uclim.2025.102338 (2025).
46. Dong, L., Hua, P., Gui, D. & Zhang, J. Extraction of multi-scale features enhances the deep learning-based daily PM(2.5) forecasting in cities. *Chemosphere* **308**, 136252. https://doi.org/10.1016/j.chemosphere.2022.136252 (2022).
47. Jin, X. et al. Integrated predictor based on decomposition mechanism for PM2.5 Long-Term prediction. *Appl. Sci.* **9**. https://doi.org/10.3390/app9214533 (2019).

48. Wang, P., He, X., Feng, H. & Zhang, G. A multivariate Short-Term trend Information-Based time series forecasting algorithm for PM2.5 daily concentration prediction. *Sustainability-Basel* **15**. https://doi.org/10.3390/su152316264 (2023).
49. Ding, S., Zhang, H., Tao, Z. & Li, R. Integrating data decomposition and machine learning methods: an empirical proposition and analysis for renewable energy generation forecasting. *Expert Syst. Appl.* **204**. https://doi.org/10.1016/j.eswa.2022.117635 (2022).
50. Yang, S. et al. A novel hybrid model based on STL decomposition and one-dimensional convolutional neural networks with positional encoding for significant wave height forecast. *Renew. Energ.* **173**, 531–543. https://doi.org/10.1016/j.renene.2021.04.010 (2021).
51. Huang, G., Zhao, X. & Lu, Q. A new cross-domain prediction model of air pollutant concentration based on secure federated learning and optimized LSTM neural network. *Environ. Sci. Pollut Res. Int.* **30**, 5103–5125. https://doi.org/10.1007/s11356-022-22454-6 (2023).
52. Zhang, L. et al. Air quality predictions with a semi-supervised bidirectional LSTM neural network. *Atmos. Pollut Res.* **12**, 328–339. https://doi.org/10.1016/j.apr.2020.09.003 (2021).
53. Wen, C. et al. 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 (2019).
54. Pak, U. et al. A deep learning approach via multifractal detrended fluctuation analysis for PM2.5 prediction. *J. Atmos. Sol-Terr Phy.* **268**. https://doi.org/10.1016/j.jastp.2025.106444 (2025).
55. Amanollahi, J. & Ausati, S. PM2.5 concentration forecasting using ANFIS, EEMD-GRNN, MLP, and MLR models: a case study of Tehran, Iran. *Air Qual. Atmos. Health.* **13**, 161–171. https://doi.org/10.1007/s11869-019-00779-5 (2019).
56. Yeo, I., Choi, Y., Lops, Y., Sayeed, A. & Efficient PM2.5 forecasting using geographical correlation based on integrated deep learning algorithms. *Neural Comput. Appl.* **33**, 15073–15089. https://doi.org/10.1007/s00521-021-06082-8 (2021).
57. Xu, S., Li, W., Zhu, Y. & Xu, A. A novel hybrid model for six main pollutant concentrations forecasting based on improved LSTM neural networks. *Sci. Rep.* **12**, 14434. https://doi.org/10.1038/s41598-022-17754-3 (2022).
58. Tao, H. et al. PM2.5 concentration forecasting: development of integrated multivariate variational mode decomposition with kernel ridge regression and weighted mean of vectors optimization. *Atmos. Pollut Res.* **15**. https://doi.org/10.1016/j.apr.2024.102125 (2024).
59. Zhang, B., Zhang, H., Zhao, G. & Lian, J. Constructing a PM2.5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. *Environ. Modell Softw.* **124**. https://doi.org/10.1016/j.envsoft.2019.104600 (2020).
60. Li, S. et al. Urban PM2.5 concentration prediction via Attention-Based CNN–LSTM. *Appl. Sci.* **10**. https://doi.org/10.3390/app10061953 (2020).
61. Zhu, J., Deng, F., Zhao, J. & Zheng, H. Attention-based parallel networks (APNet) for PM(2.5) Spatiotemporal prediction. *Sci. Total Environ.* **769**, 145082. https://doi.org/10.1016/j.scitotenv.2021.145082 (2021).
62. Chu, Y. et al. Three-hourly PM2.5 and O3 concentrations prediction based on time series decomposition and LSTM model with attention mechanism. *Atmos. Pollut Res.* **14**. https://doi.org/10.1016/j.apr.2023.101879 (2023).
63. Li, D., Liu, J. & Zhao, Y. Forecasting of PM2.5 Concentration in Beijing Using Hybrid Deep Learning Framework Based on Attention Mechanism. *Applied Sciences* 12. https://doi.org/10.3390/app122111155 (2022).
64. Teng, M. et al. Long-term PM2.5 concentration prediction based on improved empirical mode decomposition and deep neural network combined with noise reduction auto-encoder- A case study in Beijing. *J. Clean. Prod.* **428**. https://doi.org/10.1016/j.jclepro.2023.139449 (2023).
65. Peng, Z. et al. Application of machine learning in atmospheric pollution research: A state-of-art review. *Sci. Total Environ.* **910**. https://doi.org/10.1016/j.scitotenv.2023.168588 (2024).
66. Ma, Z., Zheng, W., Chen, X. & Yin, L. Joint embedding VQA model based on dynamic word vector. *Peerj Comput. Sci.* **7**, e353. https://doi.org/10.7717/peerj-cs.353 (2021).
67. Bonassi, F., Farina, M. & Scattolini, R. On the stability properties of gated recurrent units neural networks. *Syst. Control Lett.* **157**. https://doi.org/10.1016/j.sysconle.2021.105049 (2021).
68. Yu, X., Zhang, D., Zhu, T. & Jiang, X. Novel hybrid multi-head self-attention and multifractal algorithm for non-stationary time series prediction. *Inf. Sci.* **613**, 541–555. https://doi.org/10.1016/j.ins.2022.08.126 (2022).
69. Liu, Z., Ji, D. & Wang, L. PM(2.5) concentration prediction based on EEMD-ALSTM. *Sci. Rep.* **14**, 12636. https://doi.org/10.1038/s41598-024-63620-9 (2024).

Acknowledgements

This research was funded by the NSFC (National Natural Science Foundation of China) project (grant number: 42371377).

Author contributions

Author Contributions Statement: X.Y., methodology, formal analysis, data curation, and writing– original draft; Y.B., conceptualization, supervision, project administration, and writing– review & editing; Q.Y., investigation; L.D., validation; W.S., software; W.L., visualization; H.R., resources; Q.S., validation. All authors have reviewed and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00911-9.

Correspondence and requests for materials should be addressed to Y.B.