International Journal of Environmental Science and Technology (2023) 20:13535–13550
https://doi.org/10.1007/s13762-023-04900-1

ORIGINAL PAPER

Modeling air quality PM2.5 forecasting using deep sparse attention-based transformer networks

Z. Zhang[1] · S. Zhang[2]

Received: 23 March 2022 / Revised: 4 January 2023 / Accepted: 16 March 2023 / Published online: 4 April 2023

Abstract

Air quality forecasting is of great importance in environmental protection, government decision-making, people's daily health, etc. Existing research methods have failed to effectively modeling long-term and complex relationships in time series PM2.5 data and exhibited low precision in long-term prediction. To address this issue, in this paper a new lightweight deep learning model using sparse attention-based Transformer networks (STN) consisting of encoder and decoder layers, in which a multi-head sparse attention mechanism is adopted to reduce the time complexity, is proposed to learn long-term dependencies and complex relationships from time series PM2.5 data for modeling air quality forecasting. Extensive experiments on two real-world datasets in China, *i.e.*, Beijing PM2.5 dataset and Taizhou PM2.5 dataset, show that our proposed method not only has relatively small time complexity, but also outperforms state-of-the-art methods, demonstrating the effectiveness of the proposed STN method on both short-term and long-term air quality prediction tasks. In particular, on singe-step PM2.5 forecasting tasks our proposed method achieves R^2 of 0.937 and reduces RMSE to 19.04 μg/m^3 and MAE to 11.13 μg/m^3 on Beijing PM2.5 dataset. Also, our proposed method obtains R^2 of 0.924 and reduces RMSE to 5.79 μg/m^3 and MAE to 3.76 μg/m^3 on Taizhou PM2.5 dataset. For long-term time step prediction, our proposed method still performs best among all used methods on multi-step PM2.5 forecasting results for the next 6, 12, 24, and 48 h on two real-world datasets.

Keywords Air quality forecasting · Deep learning · Sparse attention · Transformer · Long-term dependency

Introduction

With the acceleration and deepening of industrialization and urbanization, air pollution has been a more and more serious problem, which heavily threatens to human health with a variety of respiratory diseases such as chronic pharyngitis, chronic bronchitis, and bronchial asthma (Chang et al. 2020; Schwartz 1993; Yan et al. 2020). Besides, heavy air pollution will lead to a haze, resulting in the low atmosphere visibility, traffic accidents, flight delays, and so on. Therefore, how to realize an accurate air quality forecasting has gradually drawn extensive attentions in recent years, due to its importance in environmental protection (Liao et al. 2015), government decision-making (Zheng et al. 2015), people's daily health (Ha Chi and Kim Oanh 2021), etc.

So far, a large number of big cities have established air quality monitoring stations in urban areas to observe the city's real-time PM2.5 and other air pollutants such as PM10, CO, O_3, NO_2, SO_2, etc. (Li and Cheng 2021; Wang et al. 2022a, b). In China, the air quality status of different cities in the east, north, and northeast of China is sometimes more notable in the world, since prior studies have been reported the chemical composition and mass concentration of PM2.5 in these areas of China (Gautam et al. 2019). Long-term exposure to PM2.5 easily causes the respiratory diseases (Chai et al. 2019; Yang et al. 2020). As a result, air pollution caused by PM2.5 has been regarded as a crucial problem threatening to people's daily health. Hence, it is of great importance to perform early diagnosis of air pollution occurrence and PM2.5 concentration estimation for air quality forecasting. At present, tremendous efforts have

Editorial responsibility: Samareh Mirkia.

✉ S. Zhang
tzczsq@163.com

[1] Zhejiang Provincial Key Laboratory of Evolutionary Ecology and Conservation, Taizhou University, Taizhou 318000, Zhejiang, China

[2] Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, Zhejiang, People's Republic of China

been made to focus on air quality forecasting (Janarthanan et al. 2021; Liu et al. 2021; Mao et al. 2021; Voukantsis et al. 2011; Yi et al. 2019; Zhu et al. 2018). Existing approaches for air quality prediction can be divided into two categories: deterministic methods and statistical methods. In particular, deterministic methods usually work in a model-driven manner. That is, they utilize the aerodynamic theory to construct a numeric model to simulate the pollutant discharge and diffusion of atmospheric pollution concentration. The representative deterministic methods contain Nested Air Quality Prediction Modeling System (NAQPMS) (Wang et al. 2001), Chemical Transport Models (CTMs) (Mihailovic et al. 2009; Ponomarev et al. 2020), Weather Research and Forecasting (WRF) (Powers et al. 2017), Community Multiscale Air Quality (CMAQ) (Zhang et al. 2014), the complicated WRF-SMOKE-CMAQ model (de Almeida Albuquerque et al. 2018), and so on. However, these deterministic methods may provide inaccurate prediction results owing to the lack of real observations (Kukkonen et al. 2003). In addition, since a variety of parameters in these models are required to be decided by experience, they easily suffer from the expensive computation cost (Xu et al. 2017).

By contrast, statistical methods usually work in a data-driven manner. In other words, based on the observed data they directly employ a statistical modeling strategy to forecast air pollutant concentrations. The conventional linear statistical methods for air quality prediction include Autoregressive Moving Average (ARMA) (Graupe et al. 1975), Autoregressive Integrated Moving Average (ARIMA) (Cekim 2020; Jian et al. 2012), Autoregressive Distributed Lag (ARDL) (Abedi et al. 2020). Nevertheless, these linear statistical methods are based on the assumption that there exist linear relationships between data variables and target labels. This does not comform to the non-linearity of real-world observed data. Therefore, these linear statistical methods may not obtain promising performance on air quality forecasting tasks. To address this issue, an alternative to these liner statistial methods is to adopt nonlinear statistical machine learning methods for air quality forecasting. The representative nonlinear statistical machine learning methods are Support Vector Regression (SVR) (Chu et al. 2021; Yang et al. 2018), Artificial Neural Network (ANN) (Agarwal et al. 2020; Arhami et al. 2013), Random Forest (RF) (Gariazzo et al. 2020), eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016), and so on. Among these nonlinear statistical machine learning methods, ANNs have become one of the most popular approaches for air quality forecasting. For instance, Ding et al. (2016) employed sparse response back-propagation training feed-forward neural networks to predict air pollutant concentration. Zhao et al. (2020) integrated forward neural networks and recurrent neural networks to predict air quality hourly in Northwest of China. Liu and Zhang (2021) developed a method of AQI (air quality index) time series prediction by means of a hybrid data decomposition and echo state networks. In recent years, ensemble learning for different ANNs has been an attractive direction. In particular, an ensemble method based on 10 distinct ANNs was used to estimate air pollution health risks (Araujo et al. 2020). Wang et al. (2020) proposed a double decomposition and optimal combination ensemble learning method for interval-valued AQI forecasting. However, due to the used single-layer network structure, these tranditional nonlinear statistical learning methods belong to shallow leaning methods, resuting in their limited feature learning ability and prediction performance on air quality forecasting tasks.

To allievate the above-mentioned problem, recently emerged deep learning techqniques (Hinton and Salakhutdinov 2006; LeCun et al. 2015) may present a possible solution. With the aid of deep multi-layer network structures, deep learning techqniques are capable of learning high-level feature representations from input data and exhibit excellent performance in the fields of computer vision, natural language processing, signal processing, and so on. The well-known deep learning techniques contain Deep Belief Network (DBN) (Hinton and Salakhutdinov 2006), Convolutional Neural Network (CNN) (Krizhevsky et al. 2012), Recurrent Neural Network (RNN) (Elman 1990) and its variant of Long Short-term Memory (LSTM) (Hochreiter and Schmidhuber 1997), and so on. At present, a variety of deep learning techniques have been successfully applied for air quality forecasting (Akbal and Ünlü 2022; Dhakal et al. 2021; Wong et al. 2021; Yang et al. 2021; Zhang et al. 2020a, 2022; Zhou et al. 2022). For instance, a deep stacked autoencoder (AE) model (Li et al. 2016), as a variant of DBN, was used to learn inherent air features for air quality prediction. Image-based air quality prediction based on CNN (Chakma et al. 2017; Zhang et al. 2016) was proposed, in which CNNs were leaverged to recognize natural images into different categories on the basis of their PM2.5 concentrations. An end-to-end deep learning model comprising of CNNs and Gradient Boosting Machine (GBM) (Luo et al. 2020) was proposed for PM2.5 concentration prediction in Shanghai City, China. A Graph-based LSTM (GLSTM) model (Gao and Li 2021) was presented to predict PM2.5 concentration in Gansu Province of Northwest in China.

In recent years, various hybrid deep learning structures have drawn extensive attention for air quality forecasting. In particular, a hybrid deep learning framework combining Variational Mode Decomposition (VMD) and Bi-directional LSTM (BiLSTM) (Zhang et al. 2021) was developed to predict PM2.5 changes in cities in China. A transfer learning-based BiLSTM (Ma et al. 2019) was utilized to improve air quality prediction performance. A spatio-temporal Convolutional LSTM Extended (C-LSTME) model (Wen et al. 2019), in which CNNs and LSTMs were integrated to learn

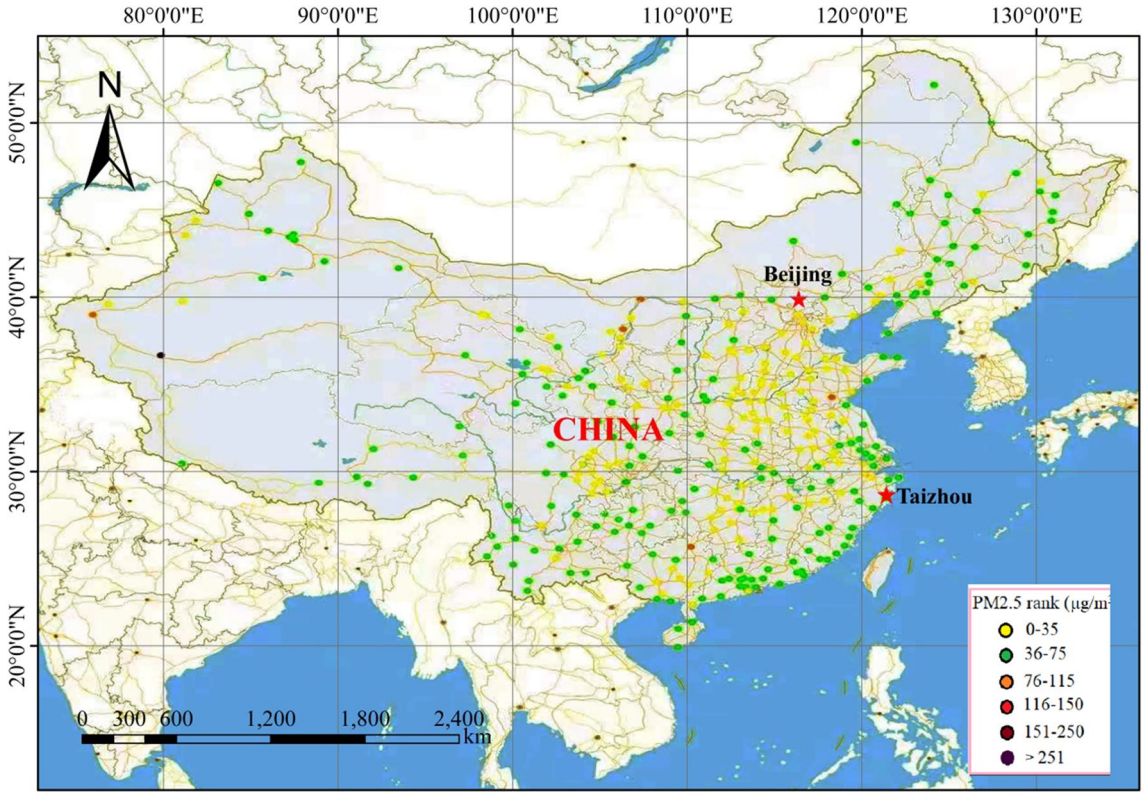


Fig. 1 Distribution of China's air quality monitoring stations (the color of each station denotes the rank of daily average PM2.5 on November 1, 2019, as depicted in the bottom right of the figure. For interpretation related to color in this figure legend, the readers see the details from the website https://www.aqistudy.cn/)

high-level spatio-temporal features, was presented to predict air quality concentration. Although these deep learning methods mentioned above have achieved good performance on air quality forecasting tasks, they may still have a drawback. That is, owing to the existed "gradient vanishing and exploding" problems in RNNs and LSTMs, as well as the limited spatial learning ability of convolutional filters in CNNs, these sequence-aligned methods are restricted in modeling long-term and complex relationships in time series PM2.5 data.

To mitigate the above-mentioned issue, in recent year the developed Transformer (Vaswani et al. 2017) method, originally proposed for machine translation tasks in natural language processing, provides possible cues for long-term air quality prediction. The original Transformer model is constructed based on self-attention mechanisms without any recurrent structures and convolutions. The motivation of the used self-attention mechanisms in the Transformer is twofold. First, compared with recurrent structures it can deal with more direct information flow across the whole sequence data, thereby allowing for more direct gradient flow. Second, it can perform faster training than recurrent structures, since most operations can be implemented in parallel. So far, self-attention-based Transformers have shown superior performance to RNNs and LSTMs in the ability of capturing long-range dependencies in the fields of machine translation (Neishi and Yoshinaga 2019; Vaswani et al. 2017), speech recognition (Chen et al. 2021; Zeyer et al. 2019), image segmentation and classification (Bazi et al. 2021; Duke et al. 2021; Lanchantin et al. 2021), electricity-consuming load analysis (Yue et al. 2020; Zhou et al. 2021), and so on. Although self-attention-based Transformers may own powerful capability of modeling long-range dependencies of sequence data, they still need large time and memory that increases quadratically with the sequence length. Besides, few studies attempt to explore Transformer-based methods for long-term air quality forecasting. To address these two issues, this paper proposes a new lightweight deep learning model for air quality forecasting based on sparse attention-based Transformer networks (STN) so as to model long-term and complex relationships from time series PM2.5 data. In our STN, a multi-head sparse attention mechanism is designed to learn long-term dependencies on the long span of time series PM2.5 data and meanwhile reduce the time complexity. Moreover, the proposed STN method can deal with the whole time series PM2.5 data for each time employ with the aid of self-attention mechanisms.

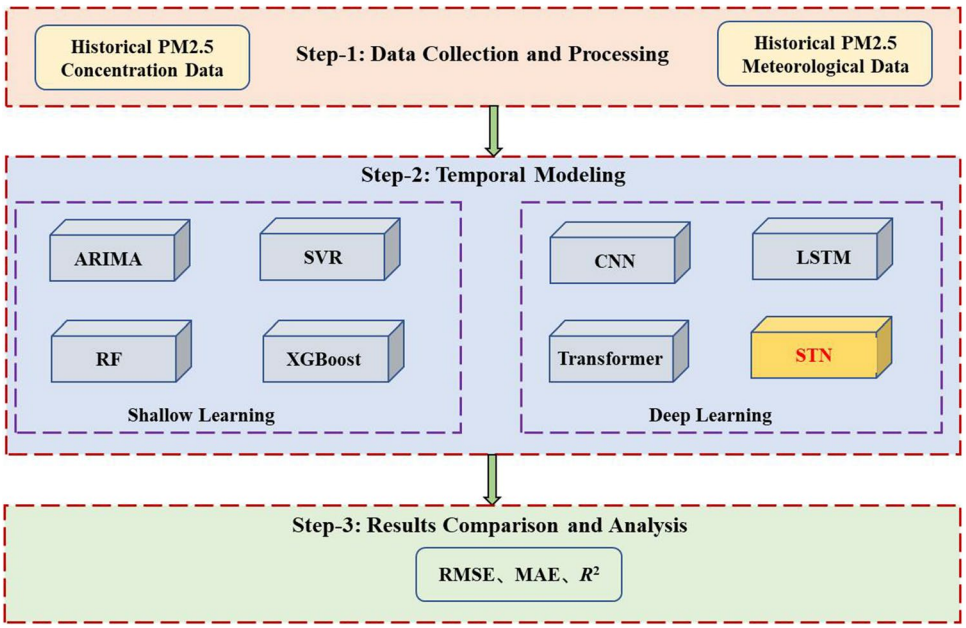


Fig. 2 Methodology structure of modeling air quality PM2.5 forecasting based on shallow learning and deep learning methods

The main contributions of this paper are summarized in three aspects: (1) a new lightweight deep learning model based on sparse attention-based Transformer networks (STN) is designed to learn long-term dependencies and complex relationships from time series PM2.5 data for deep air quality forecasting. The proposed STN method adopts a multi-head sparse attention mechanism in the encoder and decoder to learn long-term temporal dynamical information from time series PM2.5 data, and reduce time complexity simultaneously; (2) to the best of our knowledge, this is the first attempt to exploit deep sparse attention-based Transformer networks for air quality forecasting. The proposed STN method can process the entire time series PM2.5 data at the same time owing to the used self-attention mechanism. Unlike previous sequence-aligned methods, our method does not need to deal with time series PM2.5 data in an ordered

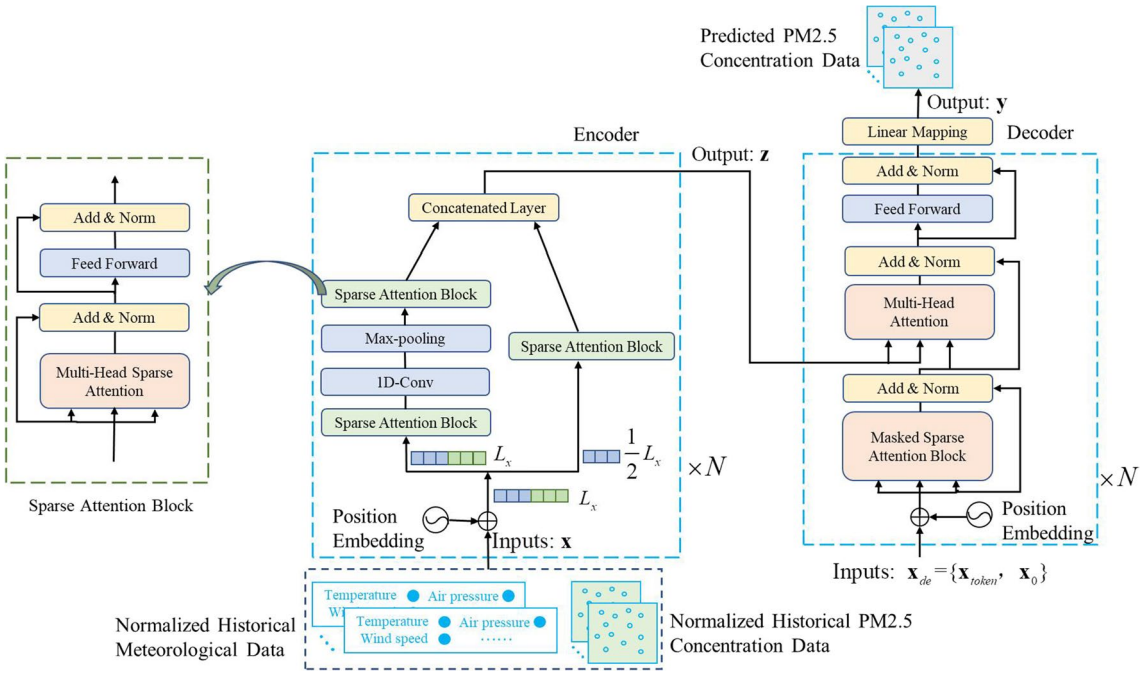


Fig. 3 Framework of proposed STN model for air quality (PM2.5 concentration) forecasting

sequence way; (3) this paper presents a comparative analysis of traditional ARIMA, SVR, RF, XGBoost, and recently developed deep learning models like CNN, LSTM, the original Transformer as well as our STN method. Extensive experiments on two real-world datasets in China, i.e., Beijing PM2.5 dataset and Taizhou PM2.5 dataset, show that our method not only has relatively small time complexity, but also outperforms state-of-the-arts, demonstrating the effectiveness of the proposed STN method on both short-term and long-term air quality prediction tasks.

Materials and methods

To evaluate the performance of the proposed method on air quality forecasting tasks, we employ two real-world air quality PM2.5 databases to conduct air quality forecasting experiments. One is the Beijing PM2.5 dataset (Liang et al. 2015) available at https://www.kaggle.com/djhavera/beijing-pm25-data-data-set. The other is Taizhou PM2.5 dataset, which was collected by our teams from Taizhou city.

Study area

In this work, we choose two typical cities, *i.e.*, Beijing and Taizhou, for studying air quality prediction, as depicted in Fig. 1. Beijing city is the Capital of China and at 116°66′ east longitude and 40°13′ north latitude. Taizhou city is located in the southeast of Zhejiang Province and at 121°42′ east longitude and 28°65′ north latitude. Figure 1 shows the distribution of China's all air quality monitoring stations and the ranking of PM2.5 values corresponding to each station on November 1, 2019. Here, the rank of PM2.5 in Fig. 1 is determined by the Ambient Air Quality Standard (GB 3095-2012) in China (Zhang et al. 2020b).

Data description

The used Beijing PM2.5 dataset (Liang et al. 2015) is hourly air quality database consisting of PM2.5 data (http://www.mee.gov.cn/) of the US Embassy in Beijing and meteorological data (http://tianqi.2345.com/) from Beijing Capital International Airport. This dataset includes eight feature items, *i.e.*, PM2.5 concentration (μg/m^3), dew point, temperature, pressure, combined wind direction, cumulated wind speed (m/s), cumulated hours of snow, cumulated hours of rain. The original dataset is recorded with an hourly interval ranging from 01/01/2010 to 12/31/2014, yielding a total of around 43,800 records. For year-independent experiments, the first four-year data are used for training, whereas the last year data (01/01/2014–12/31/2014) are selected as the testing set. For model validation, we randomly split 10% of the whole training set as the validation set. In this case, we keep that the training, and testing sets come from different years, thereby making such year-independent air quality forecasting experiments more practical. Note that such year-independent experiments are more difficult than the common year-dependent experiments in which the training and testing sets are derived from the same year.

The used hourly Taizhou PM2.5 dataset is collected from the single Hongjia monitoring station, which is located in Jiaojiang urban district from Taizhou city in Zhejiang Province. This dataset also contains eight feature items, including PM2.5 concentration (μg/m^3), dew point, temperature, pressure, combined wind direction, cumulated wind speed (m/s), cumulated hours of rain, cumulated hours of relative humidity. It consists of around 26,000 hourly records ranging from 01/01/2017 to 12/31/2019. In our experiments, the first two-year data are used as the training set, and the last year data (01/01/2019–12/31/2019) are adopted as the testing set. The randomly divided 10% of the whole training set is employed as the validation set.

Methods

Figure 2 shows the methodology structure of modeling air quality PM2.5 forecasting based on shallow learning and deep learning methods. The methodology structure starts with data collection and processing. In particular, historical PM2.5 concentration and meteorological data are collected from monitoring stations and then cleaned by means of eliminating outliers and padding missing values with a linear interpolation way. Data normalization for all air quality time series data is performed before feeding data into the used models. In the next stage of temporal modeling, various models, including shallow learning models like ARIMA, SVR, RF, XGBoost, as well as deep learning models like CNN, LSTM, Transformer, and our designed STN, are employed to model temporal dynamics from time series PM2.5 data for air quality forecasting. All used models are trained and evaluated on the collected training and testing data sets. Finally, we present the result comparison and analysis according to the used typical evaluation metrics like root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2).

Similar to the conventional Transformer (Vaswani et al. 2017), our designed sparse attention-based Transformer networks (STN) consist of encoder and decoder layers depending on self-attention mechanisms, as shown in Fig. 3. In order to learn long-term dependencies and complex relationships from time series PM2.5 data, this framework integrates two different self-attention mechanisms, including a multi-head sparse attention mechanism used in the encoder and decoder, in which a sparse attention block is designed to learn important queries for reducing time complexity, and a standard multi-head attention

mechanism (Vaswani et al. 2017) in the decoder. In the following, we will elaborate the details related to the designed STN model.

Problem description

Given input time series data $\mathbf{x} = \{x_1, x_2, \ldots, x_{L_x}\}$ ($x_i \in \mathbb{R}^{d_x}$) with a length L_x (historical meteorological data and PM2.5 concentration data) and input dimension d_x, the proposed method aims to predict the corresponding time series data $\mathbf{y} = \{y_1, y_2, \ldots, y_{L_y}\}$ ($y_i \in \mathbb{R}^{d_y}$) with a length L_y and input dimension d_y. The encoder maps input time series data $\mathbf{x} = \{x_1, x_2, \ldots, x_{L_x}\}$ into a hidden continuous representation $\mathbf{z} = \{z_1, z_2, \ldots, z_{L_z}\}$. Then, the decoder generates an output of $\mathbf{y} = \{y_1, y_2, \ldots, y_{L_y}\}$ from the given $\mathbf{z} = \{z_1, z_2, \ldots, z_{L_z}\}$. This inference is realized by using an step-by-step operation in which the decoder calculates a new hidden representation $\mathbf{z}_{k+1}$ from the previous $\mathbf{z}_k$ and other outputs in k-th step, and then forecasts the $(k+1)$-th time series data $\mathbf{y}_{k+1}$.

Position embedding

Since the original Transformer model (Vaswani et al. 2017) does not have recurrent structures and convolutions, it has no ability of leveraging the temporal information of time series data. It is thus needed to extract the relative or absolute position information of the tokens in time series data. To this end, position embedding, which is conducted with the nonlinear sine and cosine functions (Vaswani et al. 2017), is utilized to encode the temporal information of time series data. Position embedding is usually added at the bottoms of the encoder and decoder of the used Transformer model, as described in Fig. 3.

Encoder

Given input time series data $\mathbf{x}$, consisting of normalized historical meteorological data and PM2.5 concentration data, position embedding is used to encode the temporal information of $\mathbf{x}$ and generate the resulting vector with the length of L_x as inputs of the encoder. The designed encoder aims to compute the interrelationship of PM2.5-related data at each time point in the sequence data by means of using a sparse self-attention mechanism in an effort to capture the relevance and importance of PM2.5-related data at different times in the sequence data. For such self-attention encoder, the attention weights can be calculated by means of using the scaled dot-product attention of the tuple input (query, key, value).

Different from the original Transformer model (Vaswani et al. 2017) with the single branch, the designed encoder contains two-branch parallel pipelines: (1) one sparse attention block and (2) two sparse attention blocks cascaded with a 1D convolution with a kernel width 3 and a max-pooling with stride 2. Each sparse attention block consists of a multi-head sparse attention layer, a fully connected feed-forward network, followed by layer normalization. A residual connection (He et al. 2016) is used around each of two sub-layers. Here, the used 1D convolution and max-pooling operations are adopted for the self-attention distilling operation to extract the dominant attention, thereby decreasing the network size. In addition, the first branch path with one sparse attention block receives halving inputs $\frac{1}{2}L_x$, thereby reducing the number of self-attention distilling layers and improving robustness. In a concatenated layer, the learned feature maps of two-branch parallel pipelines are merged as the output $\mathbf{z}$ of the encoder.

Decoder

The decoder aims to learn the weighted attention composition of feature maps, and meanwhile, output predicted PM2.5 concentration data in a generative manner. The decoder is composed of a masked sparse attention block, a multi-head attention layer, a fully connected feed-forward network, and each of them is followed by layer normalization. Similar to the encoder, a residual connection (He et al. 2016) is also employed around each of three sub-layers. A linear mapping layer is used at the top of the decoder to output the PM2.5 prediction results $\mathbf{y}$. The masked sparse attention is obtained in the process of sparse attention computing by setting masked dot products to $-\infty$, avoiding auto-regressive. The decoder receives time series input data $\mathbf{x}_{de} = \{\mathbf{x}_{token}, \mathbf{x}_0\}$, where $\mathbf{x}_{token}$ represents the started tokens and $\mathbf{x}_0$ denotes the placeholder for target time series data.

Self-attention mechanism and sparse analysis

Given an input times series data matrix $\mathbf{X} \in \mathbb{R}^{L \times d_x}$ with a length L and input dimension d, in terms of the tuple input (query, key, value) the standard self-attention mechanism (Vaswani et al. 2017) computes the scaled dot-product as

$$\mathrm{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}}\right), \tag{1}$$

where the query matrix $\mathbf{Q} \in \mathbb{R}^{L \times d}$, key matrix $\mathbf{K} \in \mathbb{R}^{L \times d}$, value matrix $\mathbf{V} \in \mathbb{R}^{L \times d}$ are separately defined as

$$\begin{aligned} \mathbf{Q} &= \mathbf{X}\mathbf{W}_q, \\ \mathbf{K} &= \mathbf{X}\mathbf{W}_k, \\ \mathbf{V} &= \mathbf{X}\mathbf{W}_v, \end{aligned} \tag{2}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ denote the projection matrices. Equation (1) can be reformulated as its vector form. In particular, given the i-th query q_i from $\mathbf{Q}$, the attention score on the j -th key from $\mathbf{K}$ can be computed by

$$p(\mathrm{k}_j|\mathrm{q}_i) = \frac{\mathrm{e}^{\mathrm{q}_i\mathrm{k}_j^T/\sqrt{d}}}{\sum_{l=1}^{L} e^{\mathrm{q}_i\mathrm{k}_l^T/\sqrt{d}}}. \tag{3}$$

Then, the self-attention score of q_i over $\mathbf{K}$ can be defined as

$$\mathrm{Att}(\mathrm{q}_i, \mathbf{K}, \mathbf{V}) = \sum_{j=1}^{L} p(\mathrm{k}_j|\mathrm{q}_i)\mathrm{v}_j. \tag{4}$$

In this case, the time complexity of the standard self-attention mechanism (Vaswani et al. 2017) is $\mathrm{O}(L^2)$. For the query matrix, there is a potential sparsity, that is, a lot of redundant calculations are conducted to obtain attention scores for all queries. It is needed to choose important queries in which the calculated attention scores over all keys are far from the uniform distribution. To measure important queries, the Kullback–Leibler (K-L) divergence (Hershey and Olsen 2007) between the true distribution P of $p(\mathrm{k}_j|\mathrm{q}_i)$ and the uniform distribution U is used, as described below.

$$KL(P\|U) = \ln \sum_{j=1}^{L} e^{\frac{\mathrm{q}_i\mathrm{k}_j^{\mathrm{T}}}{\sqrt{d}}} - \frac{1}{L}\sum_{j=1}^{L}\frac{\mathrm{q}_i\mathrm{k}_j^{\mathrm{T}}}{\sqrt{d}} - \ln L \tag{5}$$

After dropping the constant $\ln L$, the sparse measurement of q_i can be expressed as

$$M_{sparse}(\mathrm{q}_i, \mathbf{K}) = \ln \sum_{j=1}^{L_K} e^{\frac{\mathrm{q}_i\mathrm{k}_j^{\mathrm{T}}}{\sqrt{d}}} - \frac{1}{L_K}\sum_{j=1}^{L_K}\frac{\mathrm{q}_i\mathrm{k}_j^{\mathrm{T}}}{\sqrt{d}} \tag{6}$$

According to the obtained values of M_{sparse}, larger M_{sparse} corresponds to more important queries in the self-attention mechanism. However, computing Eq. (6) is still expensive, since traversing all queries is needed to calculate every dot-product pairs. To further alleviate the computation issue, Eq. (6) can be approximated by using sampling ways:

$$\tilde{M}_{sparse}(\mathrm{q}_i, \tilde{\mathbf{K}}) = \max_j \left\{\frac{\mathrm{q}_i\mathrm{k}_j^{\mathrm{T}}}{\sqrt{d}}\right\} - \frac{1}{\tilde{L}}\sum_{j=1}^{\tilde{L}}\frac{\mathrm{q}_i\mathrm{k}_j^{\mathrm{T}}}{\sqrt{d}} \tag{7}$$

where $\tilde{\mathbf{K}}$ denotes the random sampling key matrix and $\tilde{L}$ denotes the random sampling number. After figuring out $\tilde{M}_{sparse}$ for each query, only top u dominant queries are employed to calculate self-attention, filling other pairs with zero. In this case, the time complexity is $\mathrm{O}(L \ln L)$ for a given sequence length of L.

Performance evaluation criteria

To evaluate the performance of different methods on air quality forecasting tasks, three typical evaluation metrics, such as root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2), were utilized for experiments. These three evaluation metrics are expressed below.

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{8}$$

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y_i^{mean})^2}, \tag{10}$$

Table 1 Comparisons of different methods on singe-step PM2.5 forecasting results for the next 1 h

Models	Beijing PM2.5 dataset			Execution time	Taizhou PM2.5 dataset			Execution time
	RMSE	MAE	R^2		RMSE	MAE	R^2	
SVR-POLY	35.75	28.28	0.821	0.61	9.00	7.03	0.821	0.24
SVR-RBF	40.37	34.24	0.807	0.72	8.98	7.17	0.825	0.28
SVR-LINEAR	24.17	16.83	0.899	0.53	6.83	5.08	0.873	0.21
ARIMA	22.55	14.67	0.902	8.24	6.74	4.81	0.882	4.46
RF	25.71	13.68	0.915	2.06	6.07	3.86	0.914	2.33
XGBoost	25.72	13.76	0.912	2.32	6.25	4.05	0.909	2.45
CNN	28.55	19.07	0.875	0.83	9.56	6.86	0.793	0.81
LSTM	20.63	12.91	0.926	1.33	6.32	4.55	0.893	1.65
Transformer	19.53	11.57	0.931	3.41	6.16	4.07	0.920	4.32
STN	**19.04**	**11.13**	**0.937**	2.18	**5.79**	**3.76**	**0.924**	2.78

Forward-step prediction size is 1 for the next 1 h (h1). Bold emphasis denotes the best method for smallest RSME (μg/m^3), MAE (μg/m^3), and the largest R^2. Execution time is measured in seconds

where y_i represents the observed PM2.5 value of i-th sample, $\hat{y}_i$ denotes the predicted PM2.5 value of of i-th sample, y_i^{mean} is the mean value of observed PM2.5 values, and n is the total number of samples. The smaller the RMSE and MAE are, the better the final prediction performance is. In this case, R^2 is often relatively larger.

Implementation details

All the experiments are implemented on a PC server configured with a NVIDIA Quadro P6000 graphics card which has a 24G memory. We adopt the open source machine learning framework, *i.e.*, Pytorch (https://pytorch.org) and Sklearn (https://scikit-learn.org/), to build all machine learning methods for air quality forecasting. In particular, the open-source Tensorflow library (https://github.com/tensorflow/) is used to configure deep learning and Transformer models. For these models, the Adam optimizer is employed, the initial learning rate is le^{-4}, the batch size is 32, the maximum of epochs is 200, and the mean squared error loss function is adopted. All air quality time series data are normalized to [0, 1]. The lookup size (window size), representing historical observations as input size of all used models, is set to 24 for its best performance. We compared our STN method with other typical techniques, including the traditional shallow learning models such as ARIMA and SVR, RF, XGBoost, as well as recently developed CNNs, LSTMs, original Transformer methods. They are described below in brief.

ARIMA is a typical linear statistical model for forecasting time series data. SVR is a kernel model based on nonlinear statistical machine learning theories which also can be used for time series data prediction. SVR was adopted with three different kernels (RBF, poly, and linear) with default parameter settings, *i.e.*, the penalty coefficient is 1, and the polynomial degree is 3. RF is a simple ensemble learning techniques based on decision tree predictors, and the number of trees in RF is set as 200. XGBoost is a tree-based boosting model that combines multiple tree models with low performance to build a stronger model, and the number of trees in XGBoost is also set as 200. CNNs are a typical deep learning model for 2D image data processing. Here, we use 1D-CNN for air quality prediction since time-series PM2.5 data are 1D. The used 1D-CNN contains 256 convolution kernels with a kernel width of 5 and a stride of 1, followed by a batch normalization layer, max-pooling layer, rectified linear unit layer, a dropout (0.3) layer, and a fully connected layer. LSTMs are a special kind of recurrent architecture used for modeling long-range dependencies more accurately on time series data in comparison with simple RNNs. We adopt BiLSTM for air quality forecasting, in which a forward LSTM and a backward LSTM are included. Since air quality data change significantly over time and has a strong relationship with the state before and after, BiLSTM may be appropriate for predicting PM2.5 data. In this study, we used a two-layer BiLSTM for air quality prediction, each of which has 256 hidden neurons, followed by a dropout (0.05) layer. For the original Transformer model (Vaswani et al. 2017) and the proposed STN method, we employ three encoders and two decoders for its promising performance. In the following section, we provided experimental results in two aspects: single-step forecasting for the next 1 h and multi-step forecasting for the next multiple hours.

Results and discussion

Single-step forecasting results

Table 1 shows a comparative analysis of single-step PM2.5 forecasting quantitative results (RMSE, MAE, R^2) for the next 1 h (h1) obtained by different used methods, including SVR (poly, rbf and linear kernel), ARIMA, RF, XGBoost, CNN, LSTM, Transformer, and the proposed STN method,

Table 2 Comparisons of different methods on multi-step PM2.5 forecasting results for the next 6 h on two real-world datasets

Models	Beijing PM2.5 dataset			Taizhou PM2.5 dataset		
	RMSE	MAE	R^2	RMSE	MAE	R^2
SVR-POLY	52.69	42.18	0.654	13.87	10.88	0.584
SVR-RBF	53.64	44.02	0.653	13.97	11.12	0.578
SVR-LINEAR	45.47	33.15	0.707	13.53	9.30	0.615
RF	47.65	28.38	0.737	11.79	8.81	0.699
XGBoost	48.03	28.66	0.731	13.21	9.34	0.612
CNN	42.40	28.01	0.742	13.06	9.24	0.622
LSTM	38.72	24.28	0.744	12.50	8.77	0.711
Transformer	36.95	23.12	0.752	11.54	7.74	0.716
STN	**36.41**	**22.09**	**0.782**	**11.04**	**7.19**	**0.731**

Forward-step prediction size is 6 for the next 6 h (h6). Bold emphasis denotes the best method for smallest RSME ($\mu g/m^3$), MAE ($\mu g/m^3$), and the largest R^2

Table 3 Comparisons of different methods on multi-step PM2.5 forecasting results for the next 12 h on Beijing and Taizhou PM2.5 datasets

Datasets	Models	RMSE			MAE			R^2		
		1 h-3 h	4 h-6 h	7 h-12 h	1 h-3 h	4 h-6 h	7 h-12 h	1 h-3 h	4 h-6 h	7 h-12 h
Beijing	SVR-POLY	44.06	60.09	70.72	35.13	49.21	57.58	0.747	0.562	0.348
	SVR-RBF	47.51	59.13	70.90	39.57	48.44	57.96	0.737	0.571	0.344
	SVR-LINEAR	34.04	54.56	69.77	23.90	42.41	56.47	0.812	0.591	0.351
	RF	36.50	56.15	71.83	20.84	35.95	48.10	0.820	0.614	0.411
	XGBoost	36.85	56.35	71.96	21.12	36.19	48.59	0.817	0.597	0.390
	CNN	33.18	48.89	60.31	23.15	32.92	42.96	0.807	0.623	0.485
	LSTM	31.17	46.14	59.40	19.68	30.50	41.39	0.829	0.658	0.487
	Transformer	29.70	43.98	56.02	19.26	29.67	39.27	0.831	0.662	0.492
	STN	**28.31**	**42.40**	**55.49**	**18.27**	**29.01**	**38.80**	**0.844**	**0.686**	**0.509**
Taizhou	SVR-POLY	11.58	15.83	17.72	9.04	12.71	14.26	0.710	0.458	0.323
	SVR-RBF	11.77	15.86	17.74	9.37	12.86	14.35	0.700	0.456	0.319
	SVR-LINEAR	11.45	14.69	17.23	8.25	11.35	13.53	0.713	0.461	0.326
	RF	11.32	14.57	16.56	7.94	10.43	11.99	0.743	0.534	0.404
	XGBoost	11.41	14.65	17.27	7.25	11.29	13.58	0.718	0.537	0.356
	CNN	11.21	14.43	16.33	8.09	10.38	11.86	0.729	0.543	0.419
	LSTM	11.01	13.72	15.94	7.75	9.37	11.58	0.783	0.617	0.448
	Transformer	10.06	13.28	15.83	7.51	9.35	11.29	0.785	0.616	0.464
	STN	**9.82**	**13.01**	**15.52**	**6.76**	**9.30**	**11.08**	**0.792**	**0.632**	**0.482**

Forward-step prediction size is 12 for the next 12 h (h12). Bold emphasis denotes the best method for smallest RSME (μg/m^3), MAE (μg/m^3), and the largest R^2

Table 4 Comparisons of different methods on multi-step PM2.5 forecasting results for the next 24 h on Beijing and Taizhou PM2.5 datasets

Datasets	Models	RMSE				MAE				R^2			
		1–3 h	4–6 h	7–12 h	13–24 h	1–3 h	4–6 h	7–12 h	13–24 h	1–3 h	4–6 h	7–12 h	13–24 h
Beijing	SVR-POLY	44.03	60.06	70.68	79.43	35.09	49.16	57.53	64.86	0.747	0.562	0.359	0.262
	SVR-RBF	47.49	59.13	70.87	79.99	39.54	48.42	57.91	65.47	0.737	0.571	0.360	0.269
	SVR-LINEAR	40.06	54.58	69.75	77.85	28.91	42.40	56.44	62.81	0.772	0.591	0.366	0.289
	RF	36.53	56.29	71.59	85.09	26.86	36.00	48.16	59.97	0.779	0.608	0.415	0.298
	XGBoost	36.92	56.34	71.88	85.11	26.13	36.24	48.70	60.73	0.784	0.613	0.411	0.294
	CNN	36.91	49.85	63.74	72.71	26.73	36.77	42.08	55.70	0.782	0.619	0.484	0.341
	LSTM	35.82	48.34	60.41	70.52	24.11	33.25	41.75	50.72	0.796	0.626	0.487	0.379
	Transformer	32.93	45.69	57.44	70.00	21.68	30.12	39.96	48.46	0.803	0.651	0.490	0.384
	STN	**32.35**	**44.87**	**56.47**	**69.12**	**20.95**	**29.53**	**38.48**	**47.71**	**0.814**	**0.666**	**0.495**	**0.390**
Taizhou	SVR-POLY	12.58	15.83	17.72	19.01	9.04	12.71	14.26	15.29	0.710	0.459	0.324	0.248
	SVR-RBF	12.77	15.87	17.74	19.08	9.36	12.86	14.34	15.34	0.702	0.450	0.319	0.240
	SVR-LINEAR	11.90	14.68	17.22	18.82	9.24	11.34	13.51	14.91	0.788	0.537	0.357	0.264
	RF	11.58	14.32	16.13	17.88	8.15	10.02	11.49	13.68	0.723	0.588	0.453	0.312
	XGBoost	11.64	14.48	16.23	18.03	8.20	10.23	11.88	13.72	0.717	0.556	0.423	0.304
	CNN	11.79	14.59	16.31	17.98	8.23	10.38	11.93	13.69	0.708	0.544	0.421	0.310
	LSTM	11.43	14.04	16.10	17.76	8.10	9.93	11.36	13.16	0.725	0.595	0.458	0.331
	Transformer	11.07	13.98	15.58	17.31	6.96	9.78	11.32	12.70	0.760	0.600	0.461	0.357
	STN	**9.89**	**13.41**	**14.37**	**16.98**	**6.69**	**9.53**	**11.23**	**12.22**	**0.780**	**0.615**	**0.474**	**0.363**

Forward-step prediction size is 24 for the next 24 h (h1-h24). Bold emphasis denotes the best method for smallest RSME (μg/m^3), MAE (μg/m^3), and the largest R^2

Table 5 Comparisons of different methods on multi-step PM2.5 forecasting results for the next 48 h on Beijing and Taizhou PM2.5 datasets

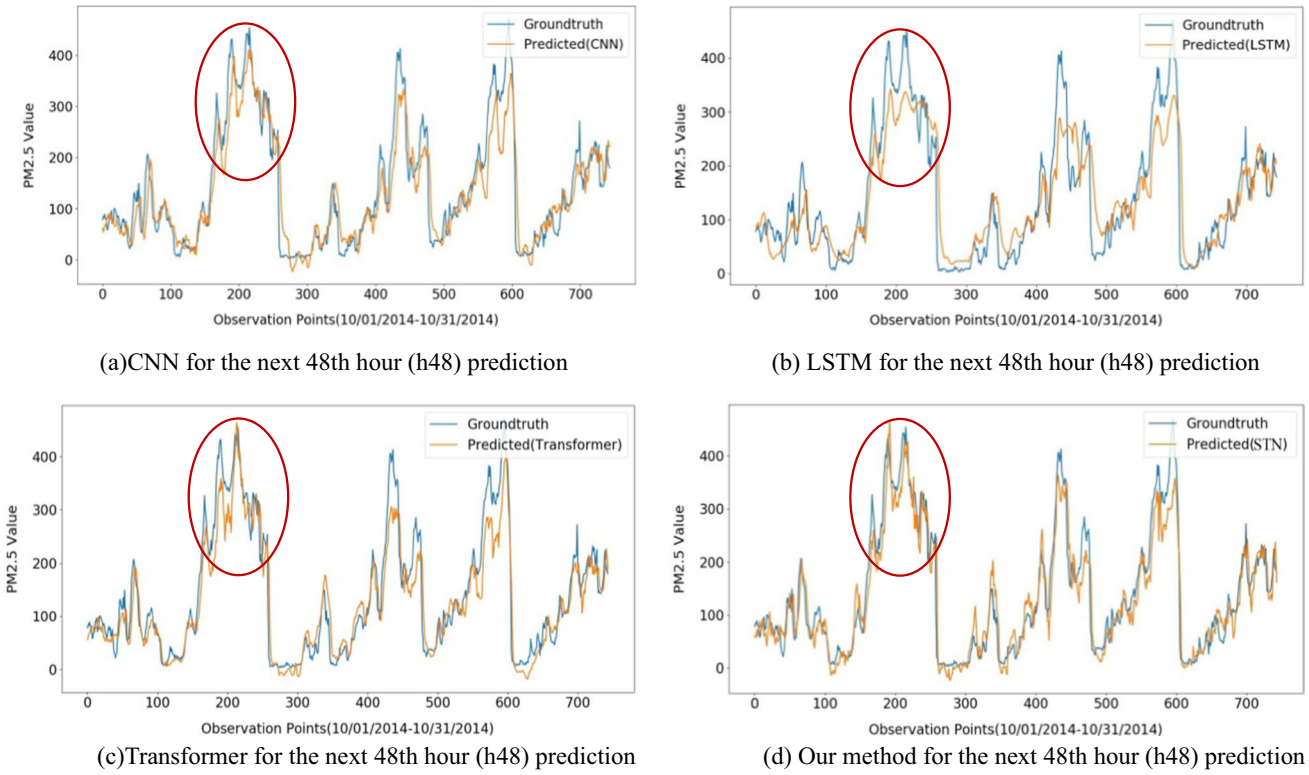
Datasets	Models	RMSE				MAE				R^2			
		1–6 h	7–12 h	13–24 h	25–48 h	1–6 h	7–12 h	13–24 h	25–48 h	1–6 h	7–12 h	13–24 h	25–48 h
Beijing	SVR-POLY	52.54	70.62	79.40	84.00	42.01	57.45	64.78	68.19	0.656	0.349	0.282	0.157
	SVR-RBF	53.57	70.81	79.93	84.09	43.92	57.85	65.37	68.31	0.655	0.345	0.281	0.152
	SVR-LINEAR	49.46	69.74	77.81	83.68	33.12	56.41	62.72	67.82	0.707	0.361	0.296	0.162
	RF	47.75	67.48	75.08	80.11	29.42	48.12	59.70	67.07	0.723	0.438	0.285	0.190
	XGBoost	48.07	67.87	75.15	80.06	29.72	48.63	59.97	67.12	0.717	0.431	0.284	0.173
	CNN	44.54	61.19	71.74	78.97	29.44	42.91	50.01	58.42	0.721	0.474	0.364	0.253
	LSTM	47.24	63.26	72.28	80.90	34.55	44.97	52.63	58.70	0.696	0.426	0.348	0.239
	Transformer	42.22	58.63	70.11	78.25	28.66	42.20	48.97	57.92	0.728	0.481	0.375	0.267
	STN	**40.53**	**57.46**	**69.94**	**77.80**	**27.52**	**40.03**	**48.23**	**57.27**	**0.736**	**0.498**	**0.397**	**0.295**
Taizhou	SVR-POLY	14.86	17.72	19.00	20.06	10.87	14.26	15.28	16.20	0.584	0.324	0.219	0.132
	SVR-RBF	14.96	17.75	19.00	20.08	11.11	14.35	15.34	16.28	0.579	0.319	0.212	0.123
	SVR-LINEAR	14.51	17.22	18.81	19.91	10.28	13.52	14.90	15.97	0.593	0.357	0.235	0.146
	RF	14.32	17.83	18.64	20.59	9.19	12.90	14.43	15.16	0.611	0.414	0.268	0.158
	XGBoost	14.43	17.91	18.92	20.77	9.32	13.01	14.57	15.78	0.607	0.394	0.251	0.151
	CNN	13.32	16.41	17.95	19.14	9.45	12.53	13.15	14.64	0.601	0.416	0.307	0.191
	LSTM	13.83	16.99	18.21	19.55	9.79	12.85	13.41	14.96	0.598	0.413	0.302	0.186
	Transformer	13.15	15.57	17.68	18.93	9.33	12.06	12.70	14.33	0.605	0.419	0.310	0.223
	STN	**12.72**	**15.06**	**17.40**	**18.13**	**9.09**	**11.29**	**12.51**	**13.90**	**0.628**	**0.424**	**0.320**	**0.249**

Forward-step prediction size is 48 for the next 48 h (h1-h48). Bold emphasis denotes the best method for smallest RSME (μg/m^3), MAE (μg/m^3), and the largest R^2

for the next 1 h on two real-world datasets, i.e., Beijing and Taizhou PM2.5 datasets. To evaluate the time computation efficiency of all models, Table 1 also presents the comparisons of the execution time for all used models, which is measured with the model's run-time implemented on the testing data.

From Table 1, we can make the following three observations, as described below.

1. Among all used methods, our STN method obtains the smallest RSME, MAE, and the highest R^2 on two real-world datasets. In particular, our method achieves the largest R^2 of 0.937 and reduces RMSE to 19.04 μg/m^3 and MAE to 11.13 μg/m^3 on Beijing PM2.5 dataset. Also, our STN method gives the largest R^2 of 0.924 and reduces RMSE to 5.79 μg/m^3 and MAE to 3.76 μg/m^3 on Taizhou PM2.5 dataset. This shows that compared with other methods such as SVR, ARIMA, RF, XGBoost, CNN, LSTM, Transformer, our STN method has more powerful ability of learn long-term dependencies and complex relationships from time series PM2.5 data for air quality forecasting. Additionally, our STN method outperforms the original Transformer method, demonstrating the advantages of our STN method on air quality forecasting tasks. The reason is that the used multi-head sparse attention mechanism in our STN has stronger ability of modeling long-term temporal dynamics from time series PM2.5 data on air quality forecasting tasks.
2. Most deep learning methods, such as LSTM, Transformer and our STN method, are superior to traditional shallow learning methods like SVR, ARIMA, RF, XGBoost on air quality prediction tasks. This indicates the advantages of deep learning methods over traditional shallow learning methods on air quality prediction tasks. Nevertheless, CNN does perform better than SVR, ARIMA, RF, and XGBoost on single-step PM2.5 prediction tasks. This shows that 2D image-based CNN is not very effective to process 1D time series PM2.5 data.
3. Among all used shallow learning methods, tree-based methods such as RF and XGBoost outperform SVR and ARIMA, demonstrating the superiority of tree-based methods to SVR and ARIMA. In addition, RF slightly performs better than XGBoost in terms of RSME, MAE, and R^2.
4. As for the computation efficiency, the ranking order of execution time for all used models is ARIMA, Transformer, STN, XGBoost, RF, LSTM, CNN, SVR-RBF, SVR-POLY, and SVR-LINEAR. Note that our STN method, as an improved version of the original Trans-



(a)CNN for the next 48th hour (h48) prediction

(b) LSTM for the next 48th hour (h48) prediction

(c)Transformer for the next 48th hour (h48) prediction

(d) Our method for the next 48th hour (h48) prediction

Fig. 4 Comparisons of multi-step ground truth and predicted PM2.5 values (μg/m^3) for the next 48 h (h48) obtained by CNN, LSTM, Transformer, and our STN method during one month (10/01/2014–10/31/2014) on Beijing PM2.5 dataset. (Each observation point in the horizontal axis represents the timescale (hour) corresponding to the obtained PM2.5 value, as depicted in the vertical axis in this figure)

former, takes less execution time compared with the original Transformer. In particular, STN separately saves 1.23 and 1.54 s on Beijing and Taizhou datasets than Transformer. This is because, in comparison with Transformer, the used multi-head sparse attention mechanism in our STN method can reduce the time complexity from $O(L^2)$ to $O(L \ln L)$, thereby yielding less execution time. This demonstrates the effectiveness of our STN method over Transformer on the time computation complexity.

Multi-step forecasting results

Table 2 presents the multi-step quantitative results of different methods on forecasting PM2.5 tasks for the next 6 h on two real-world datasets. In Table 2, the testing error of different models is the mean prediction error values in the next forward 6 h (h1–h6), thereby giving a comparative analysis of RMSE, MAE, and R^2 of SVR (poly, rbf, and linear kernel), RF, XGBoost, CNN, LSTM, Transformer, and our STN method.

As shown in Table 2, among all used models our STN method still obtains the smallest RSME, MAE, and the highest R^2 on the Beijing and Taizhou datasets, followed by Transformer, LSTM, CNN, RF, XGBoost, SVR-LINEAR, SVR-POLY, and SVR-RBF. In particular, our STN method individually yields the highest R^2 of 0.782 on Beijing PM2.5 dataset and the highest R^2 of 0.731 on Taizhou PM2.5 dataset. Additionally, our STN method reduces MAE to 22.09 μg/m^3 on Beijing PM2.5 dataset and MAE to 7.19 μg/m^3 on Taizhou PM2.5 dataset, respectively. It is worth pointing out that CNN yields better performance than traditional SVR-LINEAR and XGBoost on multi-step PM2.5 forecasting tasks for the next 6 h (h1–h6). On the contrary, CNN performs worse than SVR-LINEAR and XGBoost on single-step PM2.5 forecasting tasks for the next 1 h (h1). This indicates that CNN improves the prediction performance when the forward-step prediction size increases from the next 1–6 h.

For long-term time step prediction, Tables 3, 4 and 5 separately present performance comparisons of different methods on multi-step PM2.5 forecasting results for the next 12, 24, and 48 h on two real-world datasets. Note that for more than 6 h prediction, we split them into several intervals and trained independent models for each interval. Then, we reported the average prediction results for each interval. For instance, for the next 12 h (h1–h12) prediction, we

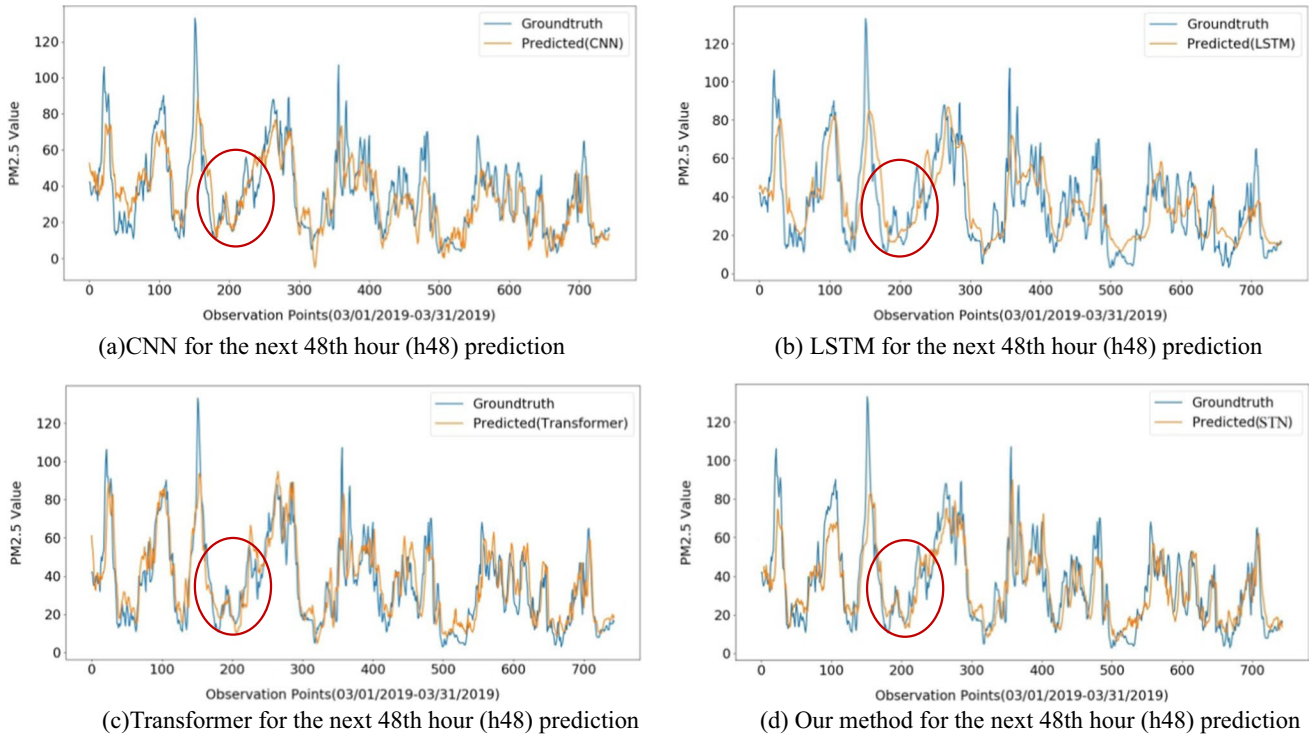


Fig. 5 Comparisons of multi-step ground truth and predicted hourly PM2.5 values (μg/m^3) for the next 48 h (h48) obtained by CNN, LSTM, Transformer, and our STN method during one month (03/01/2019–03/31/2019) on Taizhou PM2.5 dataset (each observation point in the horizontal axis represents the timescale (hour) corresponding to the obtained PM2.5 value, as depicted in the vertical axis in this figure)

divided it into three groups: 1–3, 4–6, and 7–12 h, as shown in Tables 3 and 4. For the next 24 h (h1–h24) prediction, four groups such as 1–3, 4–6, 7–12 and 13–24 h are adopted. For the next 48 h (h1–h48) prediction, four groups such as 1–6, 7–12, 13–24, 25–48 h are used.

From the results in Tables 3, 4 and 5, we can see that when the prediction time step increases, the multi-step PM2.5 forecasting performances of all used models gradually decrease. Nevertheless, it can be observed that compared with other methods, our STN method also achieves the lowest prediction error (RMSE, MAE), and the highest R^2 versus different forward prediction sizes. In addition, for the next 48 h (h1–h48), CNN performs better than LSTM, RF, XGBoost, SVR-LINEAR, demonstrating the further performance improvement in CNN on long-term air quality prediction.

To further exhibit the advantages of our STN method, we present the visualization of multi-step PM2.5 forecasting results of four deep models for the next 48 h (h1-h48) on two real-world datasets. Specially, Fig. 4 shows a comparison of multi-step ground truth and predicted PM2.5 values for the next 48 h (h48) obtained by CNN, LSTM, Transformer, and our STN method during one month (10/01/2014–10/31/2014) on Beijing PM2.5 dataset. Figure 5 presents a comparison of multi-step ground truth and predicted PM2.5 values for the next 48 h (h48) obtained by CNN, LSTM, Transformer, and our STN method during one month (03/01/2019–03/31/2019) on Taizhou PM2.5 dataset. The results in Figs. 4 and 5 indicate that our STN method performs better than other used methods when predicting PM2.5 values, especially in the time period of wave valley and peak of air quality PM2.5 testing data. Here, an illustration of the differences of different used methods is labeled with a red circle in Figs. 4 and 5.

In summary, the results in Tables 1, 2, 3, 4 and 5 and Figs. 4 and 5 on Beijing PM2.5 dataset and Taizhou PM2.5 dataset indicate that our STN method not only has relatively small time complexity, but also outperforms other used methods. This shows the advantages of our STN method on both short-term and long-term air quality prediction tasks over other used methods. More specially, on singe-step PM2.5 forecasting tasks our STN method achieves R^2 of 0.937, RMSE of 19.04, and MAE

of 11.13 on Beijing PM2.5 dataset. On Taizhou PM2.5 dataset, our STN method obtains R^2 of 0.924, RMSE of 5.79, and MAE of 3.76. For long-term PM2.5 forecasting, our STN method still gives better performance than other used methods on multi-step PM2.5 forecasting results for the next 6, 12, 24, and 48 h on two real-world datasets. In addition, it is found that the performance of all used method decreases with the increasing forward prediction size. In particular, the prediction results for the next 48 h are the worst, followed by the next 24, 12, 6, and 1 h. Besides, deep learning methods usually outperform shallow learning methods, especially for on multi-step PM2.5 forecasting tasks.

Conclusion

In this paper, we present a new lightweight method of modeling deep air quality forecasting based on sparse attention-based Transformer networks (STN) for single-step forward and multi-step forward air quality PM2.5 prediction. Our STN method, which adopts a multi-head sparse attention mechanism in the encoder and decoder to reduce the time complexity, is designed to learn long-term dependencies and complex relationships from time series PM2.5 data for air quality forecasting. Our STN method is capable of processing the entire time series PM2.5 data at the same time owing to the used self-attention mechanisms. We present a comparative analysis of traditional ARIMA, SVR, RF, XGBoost, as well as recently developed CNN, LSTM, Transformer, and our STN method. Experiment results on Beijing PM2.5 dataset and Taizhou PM2.5 dataset demonstrate that our STN method not only has relatively small time complexity, but also achieves better performance than other used methods, *i.e.*, the recently emerged deep models like the original Transformer, LSTM, CNN, and traditional ARIMA, RF, XGBoost, SVR-LINEAR, SVR-POLY, and SVR-RBF on both short-term and long-term air quality prediction tasks.

In future, it is interesting and challenging to take into account the abrupt variation in air pollution time series data for air quality forecasting. This is because such successful forecasting in advance for the sudden variation in air pollution is very beneficial to environmental protection, government decision-making, people's daily health, etc. In addition, it is also meaningful to explore more advanced deep learning models on long-term air quality prediction under different forecasting conditions. Besides, this work evaluates the performance of the proposed method based on measurement samples at two air monitoring sites in China. Therefore, it is also interesting to exploit the generalizability of the proposed STN method in larger geographical regions. Moreover, our STN method shows less time complexity than the original Transformer, but the time complexity of our STN method is still larger than traditional shallow learning methods. Therefore, how to further reduce the time complexity of our STN method is an important direction in future.

Funding This work was supported by Zhejiang Provincial National Science Foundation of China under Grant No. LY20E080013, and LZ20F020002.

Data availability statement The datasets generated during the current study are not publicly available due to the privacy but are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

Abedi A, Baygi MM, Poursafa P, Mehrara M, Amin MM, Hemami F, Zarean M (2020) Air pollution and hospitalization: an autoregressive distributed lag (ARDL) approach. Environ Sci Pollut Res 27(24):30673–30680. https://doi.org/10.1007/s11356-020-09152-x

Agarwal S, Sharma S, R S, Rahman MH, Vranckx S, Maiheu B, Blyth L, Janssen S, Gargava P, Shukla VK, Batra S, (2020) Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions. Sci Total Environ 735:139454. https://doi.org/10.1016/j.scitotenv.2020.139454

Akbal Y, Ünlü KD (2022) A deep learning approach to model daily particular matter of Ankara: key features and forecasting. Int J Environ Sci Technol 19(7):5911–5927. https://doi.org/10.1007/s13762-021-03730-3

Araujo LN, Belotti JT, Alves TA, Tadano YdS, Siqueira H (2020) Ensemble method based on Artificial Neural Networks to estimate air pollution health risks. Environ Model Softw 123:104567. https://doi.org/10.1016/j.envsoft.2019.104567

Arhami M, Kamali N, Rajabi MM (2013) Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty

analysis by Monte Carlo simulations. Environ Sci Pollut Res 20(7):4777–4789. https://doi.org/10.1007/s11356-012-1451-6
Bazi Y, Bashmal L, Rahhal MMA, Dayil RA, Ajlan NA (2021) Vision Transformers for remote sensing image classification. Remote Sens 13(3):516. https://doi.org/10.3390/rs13030516
Cekim HO (2020) Forecasting PM 10 concentrations using time series models: a case of the most polluted cities in Turkey. Environ Sci Pollut Res 27(20):25612–25624. https://doi.org/10.1007/s11356-020-08164-x
Chai G, He H, Sha Y, Zhai G, Zong S (2019) Effect of PM2.5 on daily outpatient visits for respiratory diseases in Lanzhou. China Sci Total Environ 649:1563–1572. https://doi.org/10.1016/j.scitotenv.2018.08.384
Chakma A, Vizena B, Cao T, Lin J, Zhang J (2017) Image-based air quality analysis using deep convolutional neural network. In: 2017 IEEE international conference on image processing (ICIP), Beijing, China, pp 3949–3952
Chang Q, Zhang H, Zhao Y (2020) Ambient air pollution and daily hospital admissions for respiratory system–related diseases in a heavy polluted city in Northeast China. Environ Sci Pollut Res 27:10055–10064. https://doi.org/10.1007/s11356-020-07678-8
Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, California, USA, pp 785–794
Chen X, Wu Y, Wang Z, Liu S, Li J (2021) Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In: ICASSP 2021–2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Toronto, pp 5904–5908
Chu J, Dong Y, Han X, Xie J, Xu X, Xie G (2021) Short-term prediction of urban PM 2.5 based on a hybrid modified variational mode decomposition and support vector regression model. Environ Sci Pollut Res 28(1):56–72. https://doi.org/10.1007/s11356-020-11065-8
de Almeida Albuquerque TT, de Fátima AM, Ynoue RY, Moreira DM, Andreão WL, Dos Santos FS, Nascimento EGS (2018) WRF-SMOKE-CMAQ modeling system for air quality evaluation in São Paulo megacity with a 2008 experimental campaign data. Environ Sci Pollut Res 25(36):36555–36569. https://doi.org/10.1007/s11356-018-3583-9
Dhakal S, Gautam Y, Bhattarai A (2021) Exploring a deep LSTM neural network to forecast daily PM2.5 concentration using meteorological parameters in Kathmandu Valley, Nepal. Air Qual Atmos Health 14(1):83–96. https://doi.org/10.1007/s11869-020-00915-6
Ding W, Zhang J, Leung Y (2016) Prediction of air pollutant concentration based on sparse response back-propagation training feedforward neural networks. Environ Sci Pollut Res 23(19):19481–19494. https://doi.org/10.1007/s11356-016-7149-4
Duke B, Ahmed A, Wolf C, Aarabi P, Taylor GW (2021) Sstvos: sparse spatiotemporal transformers for video object segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5912–5921
Elman JL (1990) Finding structure in time. Cogn Sci 14(2):179–211. https://doi.org/10.1016/0364-0213(90)90002-E
Gao X, Li W (2021) A graph-based LSTM model for PM2.5 forecasting. Atmos Pollut Res 12(9):101150. https://doi.org/10.1016/j.apr.2021.101150
Gariazzo C, Carlino G, Silibello C, Renzi M, Finardi S, Pepe N, Radice P, Forastiere F, Michelozzi P, Viegi G, Stafoggia M (2020) A multi-city air pollution population exposure study: combined use of chemical-transport and random-Forest models with dynamic population data. Sci Total Environ 724:138102. https://doi.org/10.1016/j.scitotenv.2020.138102
Gautam S, Patra AK, Kumar P (2019) Status and chemical characteristics of ambient PM2.5 pollutions in China: a review. Environ Dev Sustain 21(4):1649–1674. https://doi.org/10.1007/s10668-018-0123-1
Graupe D, Krause D, Moore J (1975) Identification of autoregressive moving-average parameters of time series. IEEE Trans Automat Contr 20(1):104–107. https://doi.org/10.1109/TAC.1975.1100855
Ha Chi NN, Kim Oanh NT (2021) Photochemical smog modeling of PM2.5 for assessment of associated health impacts in crowded urban area of Southeast Asia. Environ Technol Innov 21:101241. https://doi.org/10.1016/j.eti.2020.101241
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Nevada, USA, pp 770–778
Hershey JR, Olsen PA (2007) Approximating the Kullback Leibler divergence between Gaussian mixture models. In: 2007 IEEE international conference on acoustics, speech and signal processing (ICASSP'07). IEEE, Honolulu, pp IV-317–IV-320
Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507. https://doi.org/10.1126/science.112764
Hochreiter S, Schmidhuber JJNc, (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
Janarthanan R, Partheeban P, Somasundaram K, Navin Elamparithi P (2021) A deep learning approach for prediction of air quality index in a metropolitan city. Sustain Cities Soc 67:102720. https://doi.org/10.1016/j.scs.2021.102720
Jian L, Zhao Y, Zhu Y-P, Zhang M-B, Bertolatti D (2012) An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. Sci Total Environ 426:336–345. https://doi.org/10.1016/j.scitotenv.2012.03.025
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, Lake Tahoe, pp 1097–1105
Kukkonen J, Partanen L, Karppinen A, Ruuskanen J, Junninen H, Kolehmainen M, Niska H, Dorling S, Chatterton T, Foxall R, Cawley G (2003) Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmos Environ 37(32):4539–4550. https://doi.org/10.1016/S1352-2310(03)00583-1
Lanchantin J, Wang T, Ordonez V, Qi Y (2021) General multi-label image classification with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16478–16488
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
Li T, Cheng X (2021) Estimating daily full-coverage surface ozone concentration using satellite observations and a spatiotemporally embedded deep learning approach. Int J Appl Earth Obs Geoinf 101:102356. https://doi.org/10.1016/j.jag.2021.102356
Li X, Peng L, Hu Y, Shao J, Chi T (2016) Deep learning architecture for air quality predictions. Environ Sci Pollut Res 23(22):22408–22417. https://doi.org/10.1007/s11356-016-7812-9
Liang X, Zou T, Guo B, Li S, Zhang H, Zhang S, Huang H, Chen SX (2015) Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proc R Soc a: Math Phys Eng Sci 471(2182):20150257. https://doi.org/10.1098/rspa.2015.0257

Liao X, Tu H, Maddock JE, Fan S, Lan G, Wu Y, Yuan ZK, Lu Y (2015) Residents' perception of air quality, pollution sources, and air pollution control in Nanchang, China. Atmos Pollut Res 6(5):835–841. https://doi.org/10.5094/APR.2015.092

Liu H, Zhang X (2021) AQI time series prediction based on a hybrid data decomposition and echo state networks. Environ Sci and Pollut Res. https://doi.org/10.1007/s11356-021-14186-w

Liu H, Yan G, Duan Z, Chen C (2021) Intelligent modeling strategies for forecasting air quality time series: a review. Appl Soft Comput 102:106957. https://doi.org/10.1016/j.asoc.2020.106957

Luo Z, Huang F, Liu H (2020) PM2.5 concentration estimation using convolutional neural network and gradient boosting machine. J Environ Sci 98:85–93. https://doi.org/10.1016/j.jes.2020.04.042

Ma J, Cheng JCP, Lin C, Tan Y, Zhang J (2019) Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmos Environ 214:116885. https://doi.org/10.1016/j.atmosenv.2019.116885

Mao W, Wang W, Jiao L, Zhao S, Liu A (2021) Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. Sustain Cities Soc 65:102567. https://doi.org/10.1016/j.scs.2020.102567

Mihailovic DT, Alapaty K, Podrascanin Z (2009) Chemical transport models. Environ Sci Pollut Res 16(2):144–151. https://doi.org/10.1007/s11356-008-0086-0

Neishi M, Yoshinaga N (2019) On the relation between position information and sentence length in neural machine translation. In: Proceedings of the 23rd conference on computational natural language learning (CoNLL), Hong Kong, China, pp 328–338

Ponomarev N, Elansky N, Kirsanov A, Postylyakov O, Borovski A, Verevkin YM (2020) Application of atmospheric chemical transport models to validation of pollutant emissions in Moscow. Atmos Ocean Opt 33(4):362–371. https://doi.org/10.1134/S1024856020040090

Powers JG, Klemp JB, Skamarock WC, Davis CA, Dudhia J, Gill DO, Coen JL, Gochis DJ, Ahmadov R, Peckham SE (2017) The weather research and forecasting model: overview, system efforts, and future directions. Bull Am Meteorol Soc 98(8):1717–1737. https://doi.org/10.1175/BAMS-D-15-00308.1

Schwartz J (1993) Particulate air pollution and chronic respiratory disease. Environ Res 62(1):7–13. https://doi.org/10.1006/enrs.1993.1083

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, Long Beach, pp 5998–6008

Voukantsis D, Karatzas K, Kukkonen J, Räsänen T, Karppinen A, Kolehmainen M (2011) Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci Total Environ 409(7):1266–1276. https://doi.org/10.1016/j.scitotenv.2010.12.039

Wang Z, Maeda T, Hayashi M, Hsiao L-F, Liu K-Y (2001) A nested air quality prediction modeling system for urban and regional scales: application for high-ozone episode in Taiwan. Water Air Soil Pollut 130(1):391–396. https://doi.org/10.1023/A:1013833217916

Wang Z, Chen L, Zhu J, Chen H, Yuan H (2020) Double decomposition and optimal combination ensemble learning approach for interval-valued AQI forecasting using streaming data. Environ Sci Pollut Res 27(30):37802–37817. https://doi.org/10.1007/s11356-020-09891-x

Wang Y, Yuan Q, Li T, Zhu L (2022a) Global spatiotemporal estimation of daily high-resolution surface carbon monoxide concentrations using Deep Forest. J Clean Prod 350:131500. https://doi.org/10.1016/j.jclepro.2022.131500

Wang Y, Yuan Q, Zhu L, Zhang L (2022b) Spatiotemporal estimation of hourly 2-km ground-level ozone over China based on Himawari-8 using a self-adaptive geospatially local model. Geosci Front 13(1):101286. https://doi.org/10.1016/j.gsf.2021.101286

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

Wong P-Y, Lee H-Y, Chen Y-C, Zeng Y-T, Chern Y-R, Chen N-T, Candice Lung S-C, Su H-J, Wu C-D (2021) Using a land use regression model with machine learning to estimate ground level PM2.5. Environ Pollut 277:116846. https://doi.org/10.1016/j.envpol.2021.116846

Xu Y, Du P, Wang J (2017) Research and application of a hybrid model based on dynamic fuzzy synthetic evaluation for establishing air quality forecasting and early warning system: a case study in China. Environ Pollut 223:435–448. https://doi.org/10.1016/j.envpol.2017.01.043

Yan X, Zang Z, Luo N, Jiang Y, Li Z (2020) New interpretable deep learning model to monitor real-time PM2.5 concentrations from satellite data. Environ Int 144:106060. https://doi.org/10.1016/j.envint.2020.106060

Yang W, Deng M, Xu F, Wang H (2018) Prediction of hourly PM2.5 using a space-time support vector regression model. Atmos Environ 181:12–19. https://doi.org/10.1016/j.atmosenv.2018.03.015

Yang M et al (2020) Is PM1 similar to PM2.5? A new insight into the association of PM1 and PM2.5 with children's lung function. Environ Int 145:106092. https://doi.org/10.1016/j.envint.2020.106092

Yang J, Yan R, Nong M, Liao J, Li F, Sun W (2021) PM2.5 concentrations forecasting in Beijing through deep learning with different inputs, model structures and forecast time. Atmos Pollut Res 12(9):101168. https://doi.org/10.1016/j.apr.2021.101168

Yi L, Mengfan T, Kun Y, Yu Z, Xiaolu Z, Miao Z, Yan S (2019) Research on PM2.5 estimation and prediction method and changing characteristics analysis under long temporal and large spatial scale—a case study in China typical regions. Sci Total Environ 696:133983. https://doi.org/10.1016/j.scitotenv.2019.133983

Yue Z, Witzig CR, Jorde D, Jacobsen H-A (2020) BERT4NILM: a bidirectional transformer model for non-intrusive load monitoring. In: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring, New York, pp 89–93

Zeyer A, Bahar P, Irie K, Schlüter R, Ney H (2019) A comparison of transformer and LSTM encoder decoder models for ASR. In: 2019 IEEE automatic speech recognition and understanding workshop (ASRU), Singapore, pp 8–15

Zhang H, Chen G, Hu J, Chen S-H, Wiedinmyer C, Kleeman M, Ying Q (2014) Evaluation of a seven-year air quality simulation using the Weather Research and Forecasting (WRF)/Community Multiscale Air Quality (CMAQ) models in the eastern United States. Sci Total Environ 473:275–285. https://doi.org/10.1016/j.scitotenv.2013.11.121

Zhang C, Yan J, Li C, Rui X, Liu L, Bie R (2016) On estimating air pollution from photos using convolutional neural network. In: Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, pp 297–301

Zhang B, Zhang H, Zhao G, Lian J (2020a) Constructing a PM2.5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. Environ Model Softw 124:104600. https://doi.org/10.1016/j.envsoft.2019.104600

Zhang F, Shi Y, Fang D, Ma G, Nie C, Krafft T, He L, Wang Y (2020b) Monitoring history and change trends of ambient air

quality in China during the past four decades. J Environ Manage 260:110031. https://doi.org/10.1016/j.jenvman.2019.110031

Zhang Z, Zeng Y, Yan K (2021) A hybrid deep learning technology for PM2.5 air quality forecasting. Environ Sci Pollut Res 28(29):39409–39422. https://doi.org/10.1007/s11356-021-12657-8

Zhang L, Xu L, Jiang M, He P (2022) A novel hybrid ensemble model for hourly PM2.5 concentration forecasting. Int J EnvironSci Technol. https://doi.org/10.1007/s13762-022-03940-3

Zhao Z, Qin J, He Z, Li H, Yang Y, Zhang R (2020) Combining forward with recurrent neural networks for hourly air quality prediction in Northwest of China. Environ Sci Pollut Res 27(23):28931–28948. https://doi.org/10.1007/s11356-020-08948-1

Zheng Y, Yi X, Li M, Li R, Shan Z, Chang E, Li T (2015) Forecasting fine-grained air quality based on big data. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, Sydney, pp 2267–2276

Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, Zhang W (2021): Informer: beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of AAAI, pp 11106–11115

Zhou H, Zhang F, Du Z, Liu R (2022) A theory-guided graph networks based PM2.5 forecasting method. Environ Pollut 293:118569. https://doi.org/10.1016/j.envpol.2021.118569

Zhu S, Yang L, Wang W, Liu X, Lu M, Shen X (2018) Optimal-combined model for air quality index forecasting: 5 cities in North China. Environ Pollut 243:842–850. https://doi.org/10.1016/j.envpol.2018.09.025