Environmental Pollution 336 (2023) 122402

Contents lists available at ScienceDirect

Environmental Pollution

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

Spatiotemporal informer: A new approach based on spatiotemporal embedding and attention for air quality forecasting☆

Yang Feng[a], Ju-Song Kim[a,c], Jin-Won Yu[a,c], Kuk-Chol Ri[b,d], Song-Jun Yun[c], Il-Nam Han[c], Zhanfeng Qi[a], Xiaoli Wang[a,*]

[a] School of Environmental Science and Safety Engineering, Tianjin University of Technology, Tianjin, 300384, China
[b] Department of Foreign Languages and Literature, Kim Il Sung University, Pyongyang, 950001, Democratic People's Republic of Korea
[c] Department of Mathematics, University of Science, Pyongyang, 999091, Democratic People's Republic of Korea
[d] School of Foreign Languages, Tianjin University, Tianjin, 300350, China

ARTICLE INFO

Keywords:
Air quality
Deep learning
Informer
Spatiotemporal attention
Spatiotemporal embedding

ABSTRACT

Accurate prediction of air pollution is essential for public health protection. Air quality, however, is difficult to predict due to the complex dynamics, and its accurate forecast still remains a challenge. This study suggests a spatiotemporal Informer model, which uses a new spatiotemporal embedding and spatiotemporal attention, to improve AQI forecast accuracy. In the first phase of the proposed forecast mechanism, the input data is transformed by the spatiotemporal embedding. Next, the spatiotemporal attention is applied to extract spatiotemporal features from the embedded data. The final forecast is obtained based on the attention tensors. In the proposed forecast model, the input is a 3-dimensional data that consists of air quality data (AQI, $PM_{2.5}$, O_3, SO_2, NO_2, CO) and geographic information, and the output is a multi-positional, multi-temporal data that shows the AQI forecast result of all the monitoring stations in the study area. The proposed forecast model was evaluated by air quality data of 34 monitoring stations in Beijing, China. Experiments showed that the proposed forecast model could provide highly accurate AQI forecast: the average of MAPE values for from 1 h to 20 h ahead forecast was 11.61%, and it was much smaller than other models. Moreover, the proposed model provided a highly accurate and stable forecast even at the extreme points. These results demonstrated that the proposed spatiotemporal embedding and attention techniques could sufficiently capture the spatiotemporal correlation characteristics of air quality data, and that the proposed spatiotemporal Informer could be successfully applied for air quality forecasting.

1. Introduction

Over the past decades, the rapid socio-economic development and urbanization has led to the deterioration of air quality, and air pollution continues to be a serious environmental problem (Cai et al., 2018). According to the world air quality report 2022 (IQAir, 2023), poor air quality accounts for 93 billion days lived with illness each year worldwide, and the annual average $PM_{2.5}$ in less than 10% of countries and regions in the world have reached the WHO guidelines. Exposure to air pollution causes several health conditions which include asthma, cancer, lung illnesses, heart disease, and premature mortality. WHO's report presents that every year, approximately 7 million people including 600 thousand children die from air pollution before their life expectancy (WHO, 2021a). Furthermore, recent research found that the Covid-19 infection is also significantly related to air pollution (Bilal Bashir et al., 2020; Yao et al., 2021). In addition to human health, air pollution has a serious impact on the sustainable development of the world's socioeconomics and ecosystem (Jiang et al., 2018). The annual economic cost of air pollution is approximately 8 trillion dollars, more than 6.1 percent of the global annual GDP (IQAir, 2023). Air pollution is no longer not only an accidental natural disaster, but an essential global issue that is closely related to all the people in the society.

Given the importance to the risk of air pollution, many countries and regions have been implementing and strengthening measures to prevent it (Ma et al., 2020). Air quality monitoring systems are established literally everywhere to monitor air pollutant emission and obtain

☆ This paper has been recommended for acceptance by Pavlos Kassomenos.

* Corresponding author.
E-mail address: wangxiaoli@tjut.edu.cn (X. Wang).

https://doi.org/10.1016/j.envpol.2023.122402
Received 1 July 2023; Received in revised form 1 August 2023; Accepted 16 August 2023
Available online 17 August 2023

real-time air quality data (Jain et al., 2021). However, air quality monitoring systems only supply the current and past air quality data, and with this data, it is difficult to take efficient measures to prevent future air pollution (Feng et al., 2019). Since a timely prediction of air pollution can help to protect public health and minimize the risk of air pollution, it is significant to predict air quality accurately in advance (Pelaez et al., 2020; Qi et al., 2019).

Therefore, studies to develop air quality forecast models have been widely conducted (Garaga et al., 2018; Liao et al., 2021). Air quality forecast models are mainly divided into 2 categories: process-driven and data-driven methods (Xu and Yoneda, 2021).

The main idea of the process-driven methods is to predict air quality based on the simulation of the air pollution process by the physicochemical laws (Karambelas et al., 2018). Stern et al. (2008) applied chemical transport models to predict PM_{10} in northern Germany. Saide et al. (2011) used the weather research and forecasting model coupled with chemistry analysis to predict CO concentration in Santiago, Chile.

Although these studies showed the high predictability of the process-driven models, it requires detailed information on the air pollution sources and the propagation and transition process of air pollutants, and it also needs an expertise in physicochemical process of air pollution. Such a complicated process limited the application of process-driven models (Kumar et al., 2016).

Compared to the process-driven models, the data-driven models are built only by using the observed environmental data without any detailed physicochemical process. Therefore, the development of a data-driven model is much simpler and, moreover, the models that are built for one site can be easily applied to other sites (Bai et al., 2019). Due to the advantages of the data-driven models, they have been widely applied to air quality forecast. The data-driven methods can be divided into 3 categories: classical statistical, traditional machine learning, and deep learning methods (Zhang et al., 2022a). The most commonly used classical statistics-based models include autoregressive integrated moving average (Nieto et al., 2018; Zhang et al., 2018), generalized auto-regressive conditional heteroskedasticity (Kumar and De Ridder, 2010), multiple linear regression (Ma et al., 2014; Stadlober et al., 2008), and the traditional machine learning models including support vector machine (Leong et al., 2020; Zhou et al., 2019), random forest (Masmoudi et al., 2020; Zhan et al., 2018), artificial neural networks (Antanasijevic et al., 2013; Kamal et al., 2006). These models are easy-to-use and fast owing to their simple structures, but they are not appropriate for air quality forecast because the input data is disorganized and it causes that they could not simulate time delay and time order (Liu et al., 2020; Yan et al., 2021).

Along with the development of big data technology, deep learning technologies have been making a rapid progress. Compared to the traditional machine learning models, deep learning models can learn and analyse the hidden information of a big data more accurately owing to the deep hidden layers (Du et al., 2021; Kong et al., 2020). Due to the advantages, deep learning has been widely applied to the image processing, linguistic analysis, voice recognition (Kimura and Saeki, 2020; Song et al., 2020; Xu et al., 2018), and it gradually appeared in many studies on air quality forecast, proving its high capacity and efficiency (Zhang et al., 2022a).

The typical air quality forecast models based on deep learning include convolutional neural networks (Kow et al., 2020; Rijal et al., 2018; Sayeed et al., 2020), recurrent neural networks (Athira et al., 2018; Li et al., 2017; Wu et al., 2022a; Xu et al., 2021), Transformers (Wang et al., 2022b; Xu et al., 2023; Yu et al., 2023). Among these deep learning-based forecast models, Transformer-based models can quantitively evaluate the relevance of the features by attention mechanism, and conduct encoding-decoding on the features of different weights (Vaswani et al., 2017). Transformer, with these structural advantages, could be more effective in forecasting than recurrent neural networks because it could overcome the gradient vanishing, and it is also more efficient in computation. In recent years, Transformer-based models have been widely applied to air quality forecasting. Chen et al. (2022) used CNN-Transformer to predict O_3 concentration in Beijing, China. Xu et al. (2023) used wavelet-Transformer to predict $PM_{2.5}$ and PM_{10} in Guilin, China. Yu et al. (2023) used spatiotemporal Transformer to predict $PM_{2.5}$ in the wildfire-prone areas in Los Angeles, U.S.

In recent years, Informer, a variant of Transformer, has been drawing the attention of many researchers (Zhou et al., 2021). Informer was developed by combining Transformer with the temporal embedding and sparse attention mechanism, and therefore, it can more accurately express the dynamic and time dependent properties of a time series and is more proper for time series modelling such as air quality forecasting. Al-qaness et al. (2023) applied ResInformer to predict $PM_{2.5}$ concentration of 3 cities in China. Zhou and Yang (2022) combined Informer with a vision processing technology to predict $PM_{2.5}$ in a hospital. These studies demonstrated that Informer-based models could be effectively applied to air quality forecasting. However, these studies, which are mainly focused on temporal correlation, did not fully consider the spatiotemporal properties. Air quality data not only has a temporal correlation characteristics but also has a spatial correlation characteristics, and therefore it is necessary to analyse the spatiotemporal characteristics to obtain a more accurate forecast (Sun et al., 2021; Yu et al., 2023; Zhang et al., 2022b). Although the idea of the spatiotemporal embedding was suggested in (Yu et al., 2023), its attention mechanism, which is an essential part of Informer, did not consider the spatial properties. Therefore, we can expect to get a more thorough and accurate air quality forecast model by analysing the spatial properties in the attention process.

In this paper, we propose a new spatiotemporal Informer-based forecast model that can sufficiently analyse the spatiotemporal properties of air quality data.

The main novelties of this paper are as follows.

1) A novel spatiotemporal embedding method is proposed to express the spatiotemporal properties between air quality parameters. Differently from (Yu et al., 2023), we separately conducted the embedding process for the static and dynamic factors, and combined these results. Additionally, this study used the DTW distance instead of the correlation coefficient. DTW distance was proved to be more accurate than the correlation coefficient to express the relationship between time series (Kim et al., 2021; Soh et al., 2018).
2) A novel spatiotemporal self-attention mechanism is suggested. Although the spatiotemporal embedding method can express the spatiotemporal properties of air quality data to a certain extent, it could be more efficient if it is combined with a proper attention mechanism. Therefore, this study proposes a new spatiotemporal self-attention mechanism in which the self-attention calculations were implemented both in the spatial and temporal dimensions.
3) The geographical information is utilized in the spatial embedding calculation. Most of the studies on Transformer-based forecast model did not take the geographical information into account. However, research found that the geographical information could be effectively used to improve the accuracy of the forecast models (Ma et al., 2019; Wang et al., 2022a). In this regard, this research uses the geographical information (distance between the monitoring stations, longitude, latitude, and the distance from the monitoring stations to the main highways) in the spatial embedding calculation, so as to fully utilize the spatial properties of the monitoring stations.
4) A new statistical imputation method is proposed. Our method could consider the relationship between the air quality parameters in the spatial, temporal and parameter dimensions.

In this research, we choose Beijing, the capital of China, as the study area and assess the multi-positional, muti-temporal forecast performance of the proposed model. Several other forecast models are also used to compare to our model.

The rest part of this paper is organized as follows. Section 2 gives the

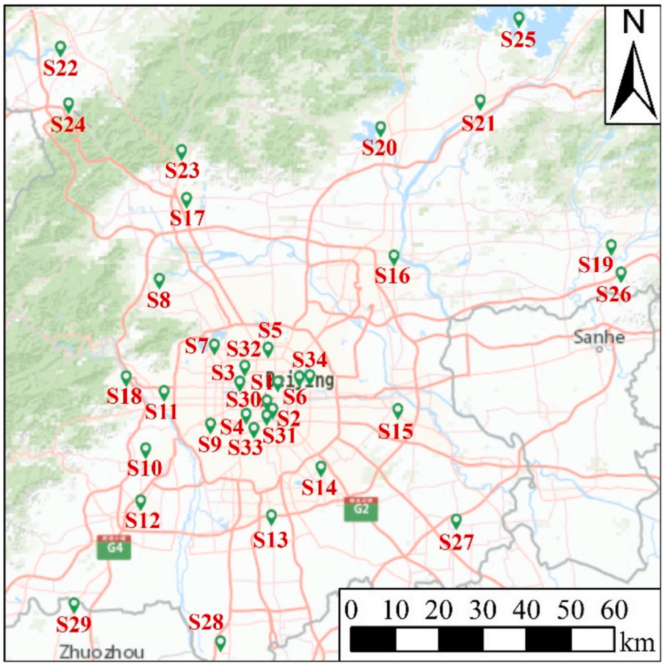


Fig. 1. Distribution of the 34 air quality monitoring stations.

detailed information on the study area, the air quality data and its statistical properties. Section 3 introduces the architecture and forecast process of the proposed model. Section 4 discusses the forecast results and Section 5 gives final conclusion on this study.

2. Study area and data description

2.1. Study area

Beijing is the capital of the People's Republic of China, the centre of the politics, culture, international exchanges, scientific and technological innovation. It has a total area of 16,410.54 km^2 and a resident population of 21, 189, 095 according to the 7th census in November 2020. Beijing is a world-famous capital and a modern international city, which has been rated as a world-class city by Globalization and World Cities. While Beijing's role in the socio-economic development of the country is increasing day by day, high population density and industrial concentration caused serious air pollution in Beijing. According to 2022 Beijing Ecological and Environmental Bulletin (BMBEE China, 2023), the average annual $PM_{2.5}$ concentration in Beijing was 30 μg/m^3 in 2022, barely reaching 'Moderate' category (MEE China, 2012a). It is 6 times the air quality guideline level (5 μg/m^3) indicated by WHO (WHO, 2021b), and it shows that Beijing's air pollution is still a serious issue. In view of the negative impact of air pollution on human health and socioeconomic development of the city, finding effective ways to manage air pollution in Beijing has become the focus of many studies at present.

2.2. Data collection

We collected hourly monitored air quality data from 0 a.m., January 1, 2018 to 11 p.m., December 31, 2019 published by the Ministry of Ecology and Environment of the People's Republic of China (http://mee.gov.cn/). Fig. 1 shows the distribution of 34 air quality monitoring stations (S1~S34) in Beijing. Our data consists of AQI and 5 other air pollutants ($PM_{2.5}$, O_3, SO_2, NO_2, CO). We divide the data into 3 parts: training set (from 0 a.m., January 1, 2018 to 11 p.m., October 31, 2019), validation set (from 0 a.m., November 1, 2019 to 11 p.m., November 30, 2019) and test set (from 0 a.m., December 1, 2019 to 11 p.m., December 31, 2019).

2.3. Data pre-processing

Data pre-processing, which handles the outliers and missing values, is the first step to build the forecast model. Some studies applied the machine learning and generative adversarial network for data imputation, but these imputation methods usually take long time (Che et al., 2018; Wu et al., 2022b). In consideration of the long training time of a deep learning model, it is not appropriate to spend a long time for data pre-processing. In this case, statistical imputation method is a proper option. The statistical imputation methods could be found in many previous studies, but most of them researches did not fully analyse the spatiotemporal and parameter aspects of a time series (Garcia-Laencina et al., 2010; Junger and de Leon, 2015).

This study proposes a new data imputation method that handle the missing values in the spatial, temporal and dimensions. Firstly, our method uses the boxplot method to detect the outliers and treat them as missing values. And then, all the missing values are handled as follows.

1) For each air quality parameter, we find 'acceptably missing cases' in which the missing values are from less than 20% of the monitoring stations. For each 'acceptably missing case', we build a linear regression model in which the dependent variables are the monitoring stations of the missing values and the independent variables are the rest monitoring stations, and it is applied to impute the missing values.
2) For each air quality parameter, we find the missing values except for 3 or more consecutive ones, and impute them by a linear imputation based on the corresponding air quality time series.
3) For each monitoring station, we find the cases in which the monitored value of only one of 6 air quality parameters is missing. For each case, a linear regression model is applied to impute the missing value.
4) After the above three imputation processes, all the cases containing the rest missing values are deleted.

Since the proposed spatiotemporal imputation method builds a linear regression model by analysing the correlation in the spatial, temporal and parameter dimensions, it can thoroughly consider the relationship between the factors.

3. Methodology

3.1. Background

3.1.1. Correlation analysis

3.1.1.1. Spatial correlation analysis. This study uses Spearman's correlation because the AQI time series has a non-normal distribution. As can be seen from Fig. S1, the correlation coefficients between the monitoring stations near each other are generally greater than those between those relatively far away. For the monitoring stations that are less than 20 km away, the correlation coefficient is relatively high (>0.9), and for those that are more than 100 km away, it is relatively low (<0.7).

3.1.1.2. Temporal correlation analysis. The temporal correlation coefficients of 34 AQI time series are shown in Fig. S2. Fig. S2 shows an apparent downward trend as the time delay increases, and it shows that the current AQI is mainly affected by the temporally close AQI values. We can see a severe change when the time delay is 20 h, and the correlation coefficients are nearly 0 when the time delay is around 80 h.

3.1.1.3. Correlation between AQI and other air quality parameters. Fig. S3 shows the distribution of correlation coefficients between the AQI and other air quality parameters of 34 monitoring stations. Fig. S3 shows that O_3 has a negative correlation, and other 4 parameters have positive

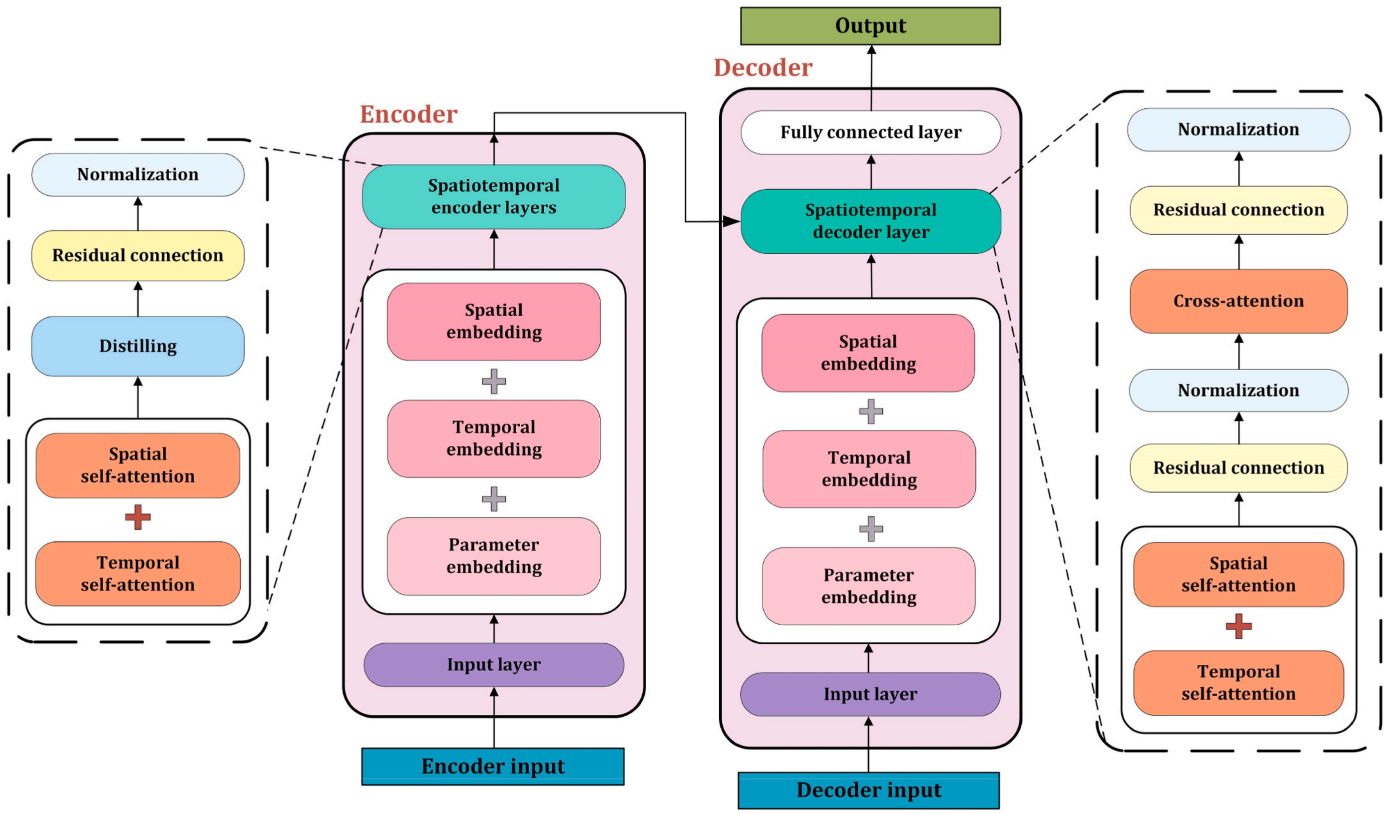


Fig. 2. Architecture of the proposed ST-Informer.

correlation with the AQI. Particularly, the correlation coefficients between $PM_{2.5}$ and AQI are greater than 0.8, and it demonstrates that $PM_{2.5}$ is the most significant factor which causes air pollution.

As can be seen from the above three correlation analyses, air quality parameters are deeply correlated in the spatial, temporal and parameter dimensions. It also proves that our idea to build a forecast model based on the spatial, temporal and parameter correlation analysis is quite credible.

3.1.2. Spatiotemporal air quality forecast

Spatiotemporal air quality forecast is to obtain a multi-positional, multi-temporal air quality forecast using the past data and the other environmental information. Let N_S is the total number of the monitoring stations, N_P is the number of air quality parameters, L_{in} is the length of a time series to be used as the input, and L_{out} is the forecast period. Then, our forecast problem can be described as follows.

$$X_t = \{x_{t-L_{in}+1}, \cdots, x_{t-1}, x_t\} \in \mathbb{R}^{L_{in}\times N_S\times N_P} \rightarrow Y_t = \{y_{t+1}, y_{t+2}, \cdots, y_{t+L_{out}}\} \in \mathbb{R}^{L_{out}\times N_s}, \forall t \tag{1}$$

where t is a time point, $x_u \in \mathbb{R}^{N_S\times N_P}$ is the monitored data matrix at time point u ($u = t - L_{in} + 1, \cdots, t$), $y_v \in \mathbb{R}^{N_s}$ is the AQI forecast result at time point v ($v = t + 1, \cdots, t + L_{out}$).

3.1.3. Informer for air quality forecast

The foundation of our forecast model is Informer (Zhou et al., 2021). The detailed description on Informer can be found in Appendix A. Informer, which is a variant of Transformer, is regarded as a powerful tool to simulate a time series because it can efficiently express the long temporal correlation of a time series by the temporal embedding, sparse attention mechanism and distilling technique.

However, the embedding process of Informer did not consider the correlation characteristics in the spatial and parameter dimensions simultaneously, therefore, it could fail to express the multi-dimensional spatiotemporal dependence. Moreover, since it's encoder-decoder uses the global attention, Informer mainly focuses on the temporal analysis, and therefore, it is not proper to forecast a spatiotemporal parameter. Air quality depends on pollution sources, wind speed and direction, and geographical factors, so it generally has a dynamic spatiotemporal distribution. Therefore, for a more accurate air quality forecast, we need to combine Informer with other techniques to analyse the spatiotemporal features.

3.2. Overview of the proposed spatiotemporal informer

This study proposes a spatiotemporal Informer (ST-informer) to overcome the shortcoming of Informer. The forecast process of ST-Informer (Fig. 2) is summarized as follows.

1) The input tensor is processed in spatial, temporal and parameter dimension by the spatiotemporal embedding. The details will be presented in 3.3.
2) The spatiotemporal self-attention is applied to extract the spatiotemporal features, and then, it is used to obtain the final AQI forecast. The details will be presented in 3.4.

3.3. Spatial, temporal and parameter embedding

The first step of the encoder and decoder is the embedding process to transform the input data into a more convenient form for the next step. This study suggests a new spatiotemporal embedding that combines the spatial, temporal and parameter embedding results.

3.3.1. Spatial embedding

Spatial embedding consists of a dynamic factor embedding and a static factor embedding. Here, the dynamic factors mean the factors that change with time and spatial position, such as air quality parameters, while the static factors are the unchanging ones with time, such as longitude and latitude. The spatial embedding process is as follows.

3.3.1.1. Dynamic factor embedding. We use DTW distance (see Appendix A) for the dynamic factor embedding. DTW distance was proved to be more effective to express the correlation between time series than the correlation coefficient (Kim et al., 2021; Soh et al., 2018). Below we present the dynamic factor embedding process.

Step 1 A latent feature matrix $h_D \in \mathbb{R}^{L_{in} \times N_s}$ is obtained by transforming the input tensor $x_D \in \mathbb{R}^{L_{in} \times N_s \times N_{P,D}}$ ($N_{P,D}$: the number of the dynamic factors).

$$h_D = x_D * w_{D,1} + b_{D,1} \tag{2}$$

where $w_{D,1} \in \mathbb{R}^{N_{P,D}}$ and $b_{D,1} \in \mathbb{R}^{L_{in} \times N_s}$ is the weight and bias, respectively.

Step 2 DTW distances based on $h_D \in \mathbb{R}^{L_{in} \times N_s}$ among all the monitoring stations are calculated and standardized, and then, the dynamic correlation between the monitoring stations is calculated. $R_D(i, j)$, the dynamic correlation between the monitoring station I and j, *is as follows.*

$$R_D(i,j) = 1 - \mathscr{d}_{DTW}(h_D(i), h_D(j)) \tag{3}$$

where $h_D(i), h_D(j) \in \mathbb{R}^{L_{in}}$ is ith and jth vector of h_D, $\mathscr{d}_{DTW}(h_D(i), h_D(j))$ is the standardized DTW distance between $h_D(i)$ and $h_D(j)$.

Step 3 For each time point t, dynamic embedding matrix $\mathscr{E}^S_{D,t} \in \mathbb{R}^{N_s \times d_{model}}$ (d_{model}: the output size) is obtained as follows.

$$\mathscr{E}^S_{D,t} = R_D * x_{D,t} * w_{D,2} + b_{D,2} \tag{4}$$

where $w_{D,2} \in \mathbb{R}^{N_{P,D} \times d_{model}}$ and $b_{D,2} \in \mathbb{R}^{N_s \times d_{model}}$ is the weight and bias, respectively.

As a result, we obtain a dynamic embedding tensor $\mathscr{E}^S_D \in \mathbb{R}^{L_{in} \times N_s \times d_{model}} = \{ \mathscr{E}^S_{D,t} | t = 1, 2, \cdots, L_{in} \}$.

3.3.1.2. Static factor embedding.

Step 1 A latent feature matrix $h_S \in \mathbb{R}^{N_s \times d_{model}}$ is obtained by transforming the input matrix $x_S \in \mathbb{R}^{N_s \times N_{P,S}}$ ($N_{P,S}$: number of the static factors).

$$h_S = x_S * w_S + b_S \tag{5}$$

where $w_S \in \mathbb{R}^{N_{P,S} \times d_{model}}$ and $b_S \in \mathbb{R}^{N_s \times d_{model}}$ is the weight and bias, respectively.

Step 2 The geographic distances among the monitoring stations are calculated and standardized, and then the static correlation $R_S \in \mathbb{R}^{N_S \times N_S}$ is calculated. $R_S(i, j)$, the static correlation between the monitoring station I and j is as follows.

$$R_S(i,j) = 1 - \mathscr{d}_{Geo}(i,j) \tag{6}$$

where $\mathscr{d}_{Geo}(i,j)$ is the standardized geographic distance between the monitoring station i and j.

Step 3 Static embedding matrix $\mathscr{E}^S_S \in \mathbb{R}^{N_s \times d_{model}}$ is obtained by multiplying the static correlation matrix R_S and latent feature matrix h_S.

3) A spatial embedding tensor, $\mathscr{E}^S$ is obtained by the summation of the dynamic embedding tensor $\mathscr{E}^S_D$ and static embedding tensor $\mathscr{E}^S_S$.

3.3.2. Temporal embedding

This study applied the temporal embedding proposed by (Zhou et al., 2021). It consists of the positional embedding and the feature embedding. The temporal embedding tensor is calculated as follows.

1) Positional embedding tensor is an embedding matrix based on the temporal position of a time series.

$$\begin{cases} \mathscr{E}^T_{Pos}(l, 2j) = \sin\left(p / 10000^{2j/d_{model}}\right) \\ \mathscr{E}^T_{Pos}(l, 2j+1) = \cos\left(p / 10000^{2j/d_{model}}\right) \end{cases}, j \in \{1, 2, \cdots, [d_{model} / 2]\} \tag{7}$$

where $p \in \{1, 2, \cdots, L_{in}\}$ is the temporal position of the time series.

2) Feature embedding tensor is an embedding matrix based on the input timestamp. As the same as in (Zhou et al., 2021), it is obtained by combining 4 parts (hour, date, month, day of the week).
3) The final temporal embedding tensor $\mathscr{E}^T$ is calculated by combining the positional and the feature embedding tensor.

3.3.3. Parameter embedding

For the $L_{in} \times N_S \times N_P$ dimension input data, a 2-dimensional convolution operation with d_{model} kernels is performed to get a $L_{in} \times N_S \times d_{model}$ dimension parameter embedding tensor $\mathscr{E}^P$. Parameter embedding can fully analyse the interdependence and the dynamic changes of the parameters because it handles the input features without separating temporally.

The final embedding tensor, $\mathscr{E}$ is obtained by the summation of the spatial embedding tensor $\mathscr{E}^S$, the temporal embedding tensor $\mathscr{E}^T$ and the parameter embedding tensor $\mathscr{E}^P$.

The embedding process is conducted in both of encoder and decoder. For the inputs of encoder and decoder, we denote the corresponding embedding tensors by $\mathscr{E}_{enc}$, $\mathscr{E}_{dec}$, respectively.

3.4. Spatiotemporal self-attention

In the encoder-decoder, the next step after the embedding process is the self-attention. Differently from Informer (Zhou et al., 2021), in the proposed spatiotemporal self-attention, the attention calculations are performed in both of the spatial and temporal dimensions and they are combined to get the spatiotemporal self-attention tensor. In the calculation of the spatial self-attention, the spatial correlation the attention tensors are calculated in the special dimension at each time point. On the other hand, in the calculation of the temporal self-attention, the attention tensors are calculated in the temporal dimension at each monitoring station. Therefore, the spatiotemporal self-attention, which combines the spatial and temporal self-attention, can fully consider the spatiotemporal correlation characteristics of the embedding tensor.

Below are the detailed calculation processes of the spatial and temporal self-attention.

3.4.1. Spatial self-attention

The input is the embedding tensor $\mathscr{E}_{enc} \in \mathbb{R}^{L_{in} \times N_s \times d_{model}}$, and it is processed by the traditional multi-head self-attention.

Step 1 For each time point t and jth head ($j \in \{1, 2, \cdots, n_{head}\}$, n_{head} : number of the attention heads), the query, key and value matrices ($Q^j_{S,t}$; $K^j_{S,t}$; $V^j_{S,t} \in \mathbb{R}^{N_s \times d}(d = d_{model} / n_{head})$) are calculated by a linear mapping on $\mathscr{E}_{enc,t} \in \mathbb{R}^{N_s \times d_{model}}$.
Step 2 Attention for the time point t and jth head is calculated as follows.

$$\mathscr{A}^j_{S,t}\left(Q^j_{S,t}, K^j_{S,t}, V^j_{S,t}\right) = \text{Softmax}\left(\frac{Q^j_{S,t}\left(K^j_{S,t}\right)^T}{\sqrt{d}}\right) V^j_{S,t} \tag{8}$$

Step 3 Obtain the spatial self-attention tensor $\mathscr{A}_S(Q_S, K_S, V_S) \in \mathbb{R}^{L_{in} \times N_s \times d_{model}} = \{\mathscr{A}_{S,t} \in \mathbb{R}^{N_s \times d_{model}} | t = 1, 2, \cdots, L_{in}\}$. Here, $\mathscr{A}_{S,t} \in \mathbb{R}^{N_s \times d_{model}}$ is the concatenation of $\mathscr{A}^j_{S,t}, j \in \{1, 2, \cdots, n_{head}\}$.

3.4.2. Temporal self-attention

The temporal self-attention tensor is calculated by the probability sparse self-attention method.

Step 1 For each monitoring station s and jth head ($j \in \{1, 2, \cdots, n_{head}\}$, n_{head} : number of the attention heads), the query, key and value matrices ($Q^j_{T,s}, K^j_{T,s}, V^j_{T,s} \in \mathbb{R}^{L_{in} \times d}$) are calculated by a linear mapping on $\mathscr{E}_{enc,s} \in \mathbb{R}^{L_{in} \times d_{model}}$.

Step 2 For each $k \in \{1, 2, \cdots, L_{in}\}$, the sparse metric of kth query is calculated as follows.

$$M\left(Q^j_{T,s}(k), K^j_{T,s}\right) = \ln \sum_{l=1}^{L_{in}} \exp\left(\frac{Q^j_{T,s}(k)\left(K^j_{T,s}\right)^T(l)}{\sqrt{d}}\right) - \frac{1}{L_{in}} \sum_{l=1}^{L_{in}} \frac{Q^j_{T,s}(k)\left(K^j_{T,s}\right)^T(l)}{\sqrt{d}} \tag{9}$$

Step 3 Among $\{M(Q^j_{T,s}(k), K^j_{T,s}) | k = 1, 2, \cdots, L_{in}\}$, the top 5 ln L_{in} elements are chosen. The sparse query matrix $\overline{Q}^j_{T,s}$ is comprised of the chosen $Q^j_{T,s}(k)$.

Step 4 Calculate the temporal self-attention for monitoring station s and jth head.

$$\mathscr{A}^j_{T,s}\left(Q^j_{T,s}, K^j_{T,s}, V^j_{T,s}\right) = \text{Softmax}\left(\frac{\overline{Q}^j_{T,s}\left(K^j_{T,s}\right)^T}{\sqrt{d}}\right) V^j_{T,s} \tag{10}$$

Step 5 Obtain the temporal self-attention tensor $\mathscr{A}_T \in \mathbb{R}^{L_{in} \times N_s \times d_{model}} = \{\mathscr{A}_{T,s} \in \mathbb{R}^{L_{in} \times d_{model}} | s = 1, 2, \cdots, N_S\}$. Here, $\mathscr{A}_{T,s} \in \mathbb{R}^{L_{in} \times d_{model}}$ is the concatenation of $\mathscr{A}^j_{T,s}(Q^j_{T,s}, K^j_{T,s}, V^j_{T,s}), j \in \{1, 2, \cdots, n_{head}\}$.

Finally, the spatiotemporal self-attention tensor $\mathscr{A}$ is obtained by the summation of the spatial self-attention tensor $\mathscr{A}_S$ and temporal self-attention tensor $\mathscr{A}_T$.

3.5. Other calculation processes in encoder-decoder

3.5.1. Encoder layers

Except for the spatiotemporal embedding (see 3.3) and spatiotemporal self-attention (see 3.4), the rest part of encoder is as the same as in (Zhou et al., 2021), that is, the dropout, residual connection, normalization, distilling, etc. Here, the distilling operation is to reduce the redundant information and improve the computation speed.

3.5.2. Decoder layer

Our decoder is a single decoder layer that consists of 2 attention layers. In the first attention layer, the spatiotemporal self-attention calculation (see 3.4) is conducted. The rest part of decoder is as the same as in (Zhou et al., 2021), that is, in the second attention layer, cross-attention calculation is conducted based on the self-attention tensors from the encoder and first attention layer of the decoder. Final output is obtained through the dropout, residual connection, and normalization operation.

3.6. Evaluation metrics

3.6.1. Metrics for forecast error

In order to assess the forecast error, this study employs the following 3 statistical metrics: mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE).

$$\text{MAE} = \frac{\sum_{t=1}^{L} |Y_t - \hat{Y}_t|}{L}, \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{L}\sum_{t=1}^{L} (Y_t - \hat{Y}_t)^2}, \tag{12}$$

$$\text{MAPE} = \frac{1}{L}\sum_{t=1}^{L} \frac{|Y_t - \hat{Y}_t|}{Y_t} \tag{13}$$

where Y_t is actual value, $\hat{Y}_t$ is predicted value and L is the number of samples.

3.6.2. Metrics for AQI category prediction accuracy

When forecasting AQI, it is significant to get an accurate AQI category. In this study, we referred to the AQI classification (Table S1) from "Technical Regulation on Ambient Air Quality Index (HJ633-2012)" (MEE China, 2012b) published by Ministry of Ecology and Environment of the People's Republic of China. For each actual and predicted AQI value, we obtain actual AQI category ($r(Y_t)$) and predicted AQI category ($r(\hat{Y}_t)$). And then, we use the following 3 metrics to evaluate the overall AQI category prediction accuracy.

$$\text{A}_{\text{Cat}} = \frac{1}{L}\sum_{t=1}^{L} \alpha_t, \alpha_t = \begin{cases} 1, if\ r(Y_t) = r(\hat{Y}_t), \\ 0, \text{otherwise}, \end{cases} \tag{14}$$

$$\text{U}_{\text{Cat}} = \frac{1}{L}\sum_{t=1}^{L} \beta_t, \beta_t = \begin{cases} 1, if\ r(Y_t) < r(\hat{Y}_t), \\ 0, \text{otherwise}, \end{cases} \tag{15}$$

$$\text{O}_{\text{Cat}} = \frac{1}{L}\sum_{t=1}^{L} \gamma_t, \gamma_t = \begin{cases} 1, if\ r(Y_t) > r(\hat{Y}_t), \\ 0, \text{otherwise}, \end{cases} \tag{16}$$

where A_{Cat} is the AQI category prediction accuracy, U_{Cat} is the rate of the underestimated prediction, O_{Cat} is the rate of the overestimated prediction.

3.6.3. Metrics for comparison between models

When assessing the performance of a model, it is common to compare it with other models. This study introduces the following metric to quantitatively evaluate the advantage of the model of interest over the other models.

$$\text{IP}\left(\frac{\text{Model}}{\text{Model}_0}\right) = \frac{\text{SI(Model)} - \text{SI(Model}_0)}{\text{SI(Model}_0)} \times 100\% \tag{17}$$

where IP(•) means improvement percentage, Model is the model of interest, Model_0 is a competitor model, and SI(•) is a statistical indicator chosen among MAE, RMSE, MAPE. The SI(•) used in this study is MAPE.

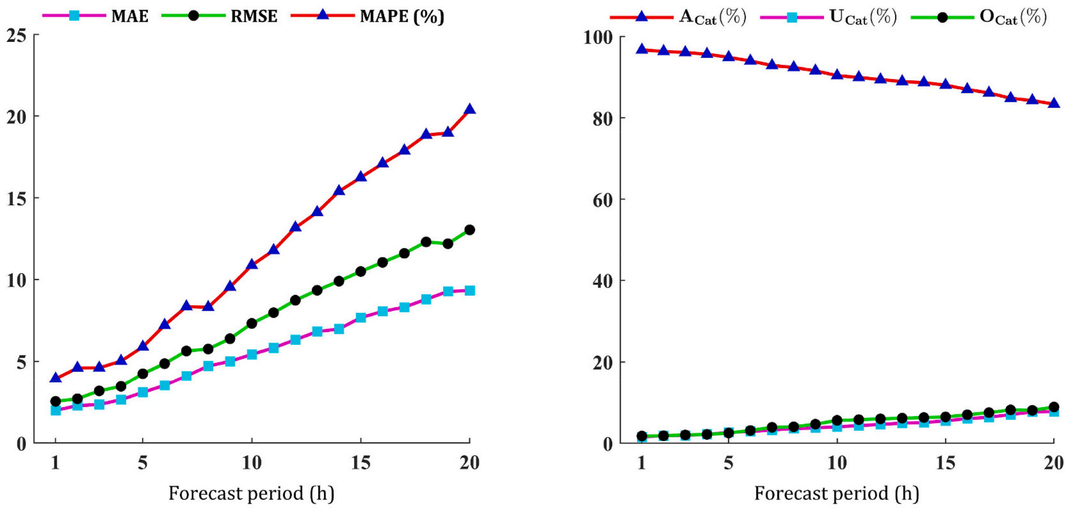
4. Result and discussion

4.1. Hyperparameter optimization

In this study, we develop a forecast model that predicts the AQI values in the future 20 h ($L_{out} = 20$). The hyperparameters of our model are optimized based on the training and validation sets. By grid search, we select the optimized input length L_{in} in {20, 40, 80, 120}, the number of encoder layers in {2, 3, 4}, the number of the heads of multi-head attention n_{head} in {8, 16}, output size of multi-head attention d_{model} in {128, 256, 512}. Other options of our model (optimizer, learning rate, batch size) are as the same as in (Zhou et al., 2021).

4.2. Forecast performance of the proposed model

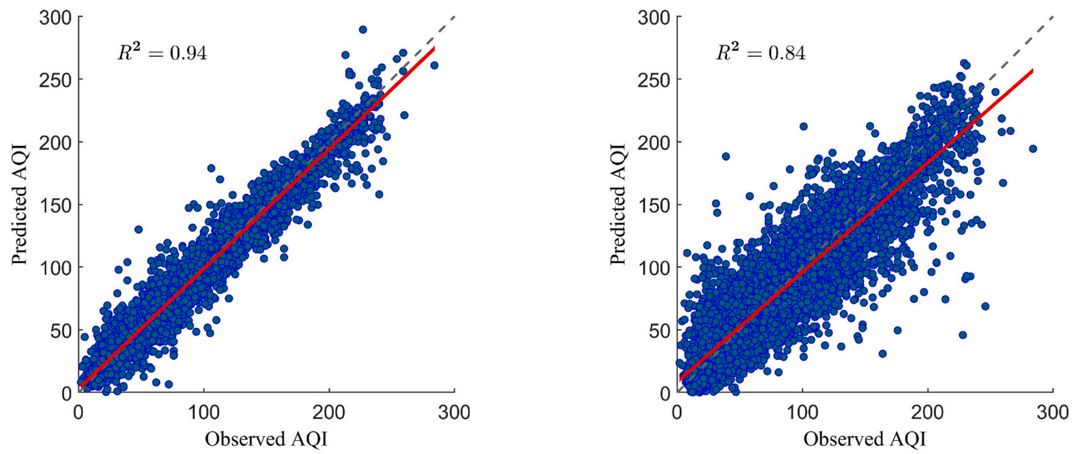
After the hyperparameters were determined, we again trained the forecast model by the sumset of the training and validation set. And we evaluated the performance of the forecast model by the testing set.



(a) Forecast error

(b) AQI category prediction accuracy

Fig. 3. Forecast performance of the proposed model.



(a) 10h ahead forecast

(b) 20h ahead forecast

Fig. 4. Correlation between the actual and predicted AQI (p-value<0.01).

4.2.1. AQI forecast result

Fig. 3 (a) shows the forecast error of our model (MAE, RMSE, MAPE). The average and maximum values of MAE, RMSE, MAPE are 5.62, 7.63, 11.61%; 9.33, 13.04, 20.37%, respectively. From the result, it can be seen that the forecast error of our model is relatively low. It slowly increases with the increase of the forecast period.

4.2.2. AQI category prediction result

AQI category prediction accuracy of our model is calculated by Eq. (14)–(16). As can be seen from Fig. 3 (b), all the A_{Cat} values are higher than 83.38%, U_{Cat} and O_{Cat} values are less than 8.87%. And this shows that our model has a great capacity in AQI category prediction.

4.2.3. Correlation between the actual and predicted AQI values

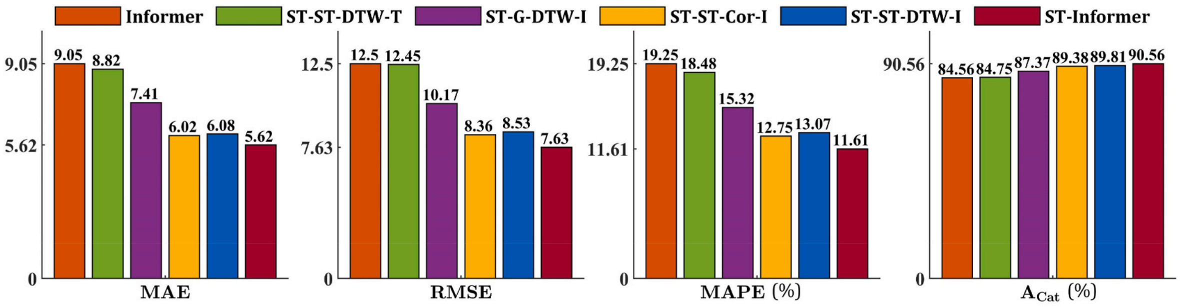
For the convenience of our experiment, here, we take 2 forecast periods (10 h, 20 h) to conduct the correlation analysis. Fig. 4 presents the scatter plots showing the correlation between the actual and predicted values. It can be seen that the coefficient of determination (R^2) values are considerably high ($R^2 \geq 0.84$). And it shows that the proposed forecast model could sufficiently express the dynamic changes of the AQI time series.

4.3. Comparison with other models

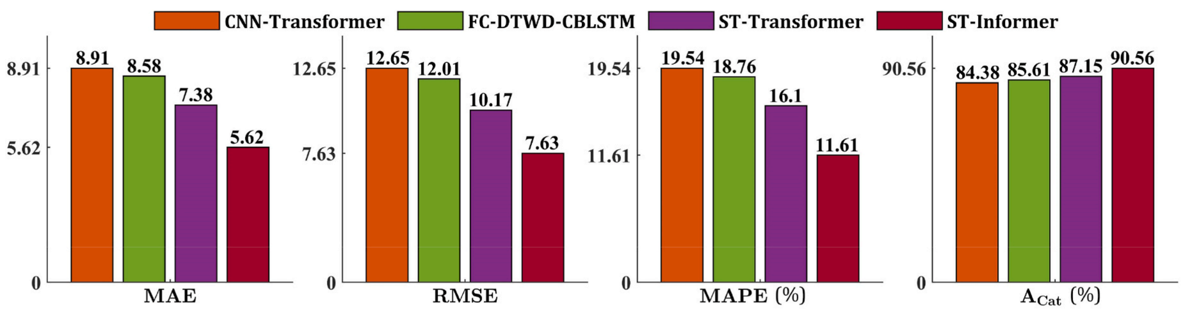
4.3.1. Comparison with similar models

The structure of the models is shown in Table S2. The hyperparameters of these models are obtained in a similar way to ST-Informer (see 4.1). And then, our model ST-Informer is compared with these models. From the comparison result (Fig. 5 (a)), we can draw the following conclusions.

1) The forecast accuracy of the Informer-based spatiotemporal models (ST-G-DTW-I, ST-ST-Cor-I, ST-ST-DTW-I, ST-Informer) are higher than the Transformer-based spatiotemporal model (ST-ST-DTW-T). IP values (Eq. (17)) of ST-G-DTW-I, ST-ST-Cor-I, ST-ST-DTW-I and ST-Informer over ST-ST-DTW-T are 17.10%,31.01%, 29.27% and 37.18%, respectively. This shows that Informer has a great capacity to extract the spatiotemporal features.
2) ST-G-DTW-I that applied the spatiotemporal embedding showed better forecast performance than Informer: IP of ST-G-DTW-I over Informer is 20.42%. And this proves that the proposed spatiotemporal embedding could sufficiently express the spatiotemporal features of the input data.



(a) Comparison with similar models



(b) Comparison with the models from the previous studies

Fig. 5. Comparison of the forecast models.

3) The models that applied the spatiotemporal self-attention (ST-ST-Cor-I, ST-ST-DTW-I, ST-Informer) showed better forecast performance than ST-G-DTW-I. IP values of ST-ST-Cor-I, ST-ST-DTW-I and ST-Informer over ST-G-DTW-I are 16.78%, 14.69% and 24.22%, respectively. This shows that the proposed self-attention could significantly enhance the feature extraction capacity of Informer.
4) ST-Informer, which applied the DTW distance-based spatial embedding, showed better forecast performance than ST-ST-Cor-I that applied correlation coefficient-based spatial embedding: IP of ST-Informer over ST-ST-Cor-I is 8.94%. And this proves that DTW distance is a better option than correlation coefficient to express the correlation between time series.
5) ST-Informer, which applied the geographical distance-based static embedding, showed better forecast performance than ST-ST-DTW-I: IP of ST-Informer over ST-ST-DTW-I is 11.17%. It shows that the proposed geographical distance-based static embedding is an efficient method to improve the AQI forecast accuracy.

4.3.2. Comparison with the models from the previous studies

Here, we compare ST-Informer with the models from the previous studies: CNN-Transformer (Chen et al., 2022), FC-DTWD-CBLSTM (Kim et al., 2021), ST-Transformer (Yu et al., 2023). From the comparison result (Fig. 5 (b)), it can be seen that ST-Informer shows better forecast performance than other models.

4.3.3. Comparison with other imputation methods

In this part, we evaluate the proposed spatiotemporal imputation method by comparing with several common imputation methods (Table S3). Data pre-processing is conducted by the imputation methods, and then, the pre-processed data is used to train and test the

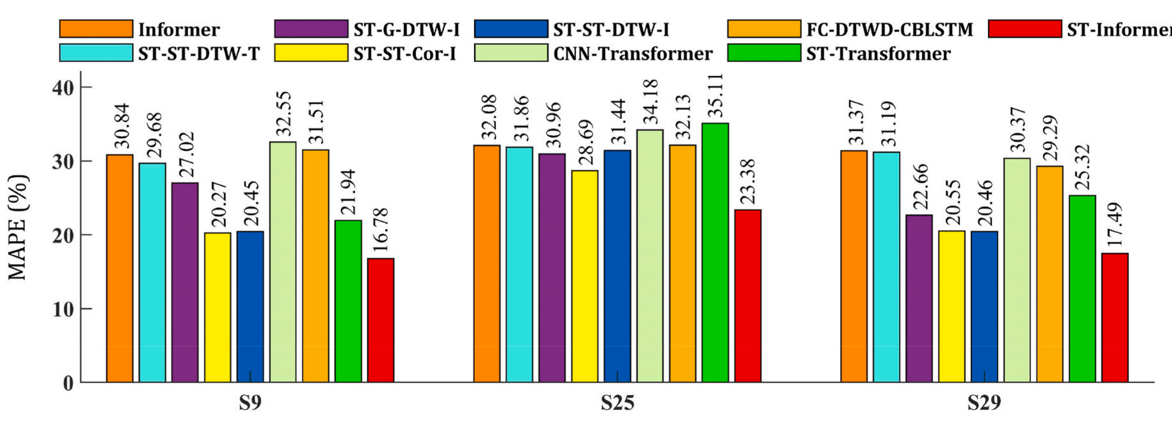


Fig. 6. MAPE of the forecast results at a single monitoring station.

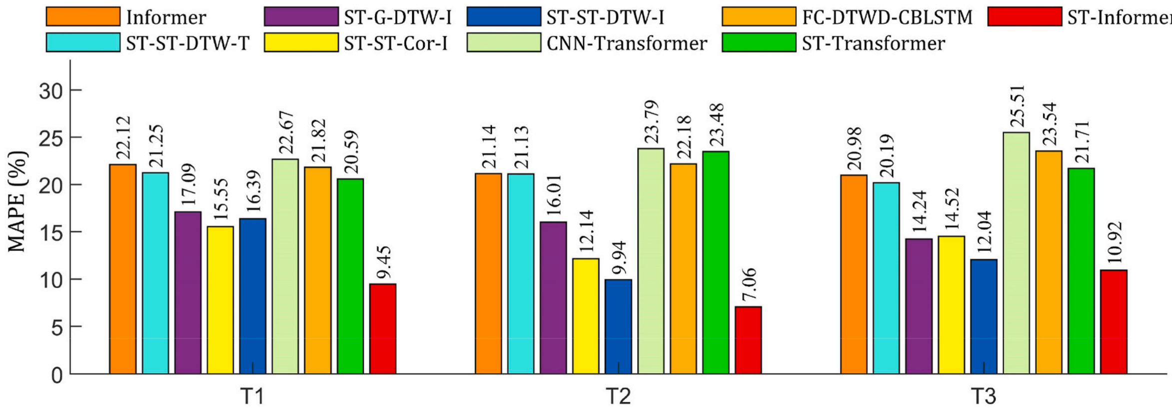


Fig. 7. MAPE of the forecast results at a single time point (%).

proposed forecast model. From the comparison result (Fig. S4), it is obvious that the forecast performance of the model that was built based on the proposed imputation is better than others. It proves that the proposed imputation method can handle the missing values in a more reasonable way, so as to help the forecast model to fully analyse the spatiotemporal characteristics of an air quality data.

As the above experiments show, the proposed forecast model ST-Informer has a high capacity for AQI forecasting.

4.4. Discussion

In 4.3, we assessed the overall performance of the forecast models by calculating the average forecast accuracy for the total test set for all the monitoring stations in the future 20 h. However, the main purpose of AQI forecasting is to get an accurate forecast at a specific monitoring station or at a specific time point. In view of the difficulty of forecasting a AQI time series with highly dynamic changes, it is an essential evaluation for an AQI forecast model.

Therefore, in this part, we conduct an experiment to assess the forecast performance at a specific monitoring station and at a specific time point. For the convenience of our experiment, here, we only provide the result for 20 h ahead prediction.

4.4.1. Forecast performance at a single station

In the study area, 3 typical monitoring stations (S9, S25, S29) are selected. S9 is located in the city centre and air quality there is the worst in the city centre, S25 is located in the northern part of the city and air quality there is the best in the city, S29 is located in the southern part of the city and air quality there is the worst in the city.

Here, we conduct the following 3 experiments.

1) For the 9 forecast models (see 4.3), the forecast performance at S9, S25, and S29 was analysed. Fig. 6 shows the forecast accuracy of these models, and Table S4 shows IP values of ST-Informer over the other 8 models. As can be seen from Fig. 6, ST-informer provided a more accurate forecast than others. From Table S4, we can see that all the IP values are greater than 14.52%.
2) For the 9 forecast models, AQI category prediction at S9, S25, and S29 was analysed. Prediction result of the 9 models is shown in Table S5. Results show that A_{Cat}, U_{Cat}, O_{Cat} values of ST-Informer are better than the others at all the 3 stations.
3) The agreement between the predicted and actual AQI values is analysed. Here, among the 8 competitor models, we selected 2 models (Informer, ST-G-DTW-I) that have a similar structure to ST-Informer. The result is shown in Fig. S5 and Fig. S6. As Fig. S5 shows, the R^2 value of ST-Informer is the greatest, and the points are mostly concentrated around the regression line. Moreover, as Fig. S6 shows, the predicted values of ST-Informer are closer to the actual values than other models. It is more apparent for the points near the extrema (299th point of S9, 250th point of S25, 306th point of S29).

The above 3 experiments proves that the proposed forecast model ST-Informer can provide a highly accurate forecast at a single monitoring station.

4.4.2. Forecast performance at a single time point

In order to analyse the forecast performance at a single time point, we select 3 typical time points: 7 a.m., December 29, 2019 (the highest AQI in the morning), 10 p.m., December 28, 2019 (the highest AQI in the evening), 6 p.m., December 28, 2019 (the highest AQI in the afternoon). At these three time points, AQI of the most of the monitoring stations belong to 'Unhealthy' category. Denote these 3 time points by T1, T2, and T3.

1) At T1, T2, and T3, forecast performance of the 9 forecast models (see 4.3) was analysed. Fig. 7 shows the forecast error of the models, and Table S6 presents the IP values of ST-Informer over the others. As Fig. 7 shows, ST-Informer provided better forecast than other models. Moreover, as can be seen from Table S6, most of the IP values are greater than 20%.
2) At T1, T2, and T3, AQI category prediction of the 9 forecast models was analysed. The result is shown in Table S7. As Table S7 shows, all the A_{Cat} values of ST-Informer are greater than other 8 models. ST-Informer also provided good U_{Cat} and O_{Cat} values for all the three time points. Although some models showed better U_{Cat} values (ST-ST-Cor-I, CNN-Transformer, FC-DTWD-CBLSTM, and ST-Transformer at T2) and O_{Cat} values (ST-ST-Cor-I and ST-ST-DTW-I at T3) than ST-Informer, the overall prediction accuracy of these models is much worse than ST-Informer.
3) At T1, T2, and T3, the agreement between the predicted and actual AQI values was analysed. Here, among the 8 competitor models, we selected 2 models (Informer, ST-G-DTW-I) that have a similar structure to ST-Informer. The result is shown in Fig. S7 and Fig. S8. As can be seen from Fig. S7, the R^2 value of ST-Informer is the greatest, and the points are mostly concentrated around the regression line. Moreover, as Fig. S8 shows, the predicted values of ST-Informer are closer to the actual values than other models.

As this experiment shows, ST-Informer can provide highly accurate forecast at a single time point.

4.4.3. Stability of forecast

In this part, we assess the forecast stability of ST-Informer by using standard deviation of error (SDE). Three monitoring stations (S9, S25, S29) are selected for this experiment. Fig. S9 shows the forecast error of Informer, ST-G-DTW-I and ST-Informer, and the corresponding SDE

Table 1
SDE values of the models.

Model	S9	S25	S29
Informer	21.51	17.42	25.02
ST-G-DTW-I	15.65	13.76	19.28
ST-Informer	10.94	10.41	12.20

values are presented in Table 1. In Fig. S9, the two black lines are the lower and upper bound of the forecast error by ST-Informer. As can be seen from Fig. S9, some of the forecast errors of the other models are distributed outside of the space between the two black lines. In addition, as Table 1 shows, SDE for our model is smaller than others. Overall, this result demonstrates our model can provide more stable and accurate forecast than other models.

5. Conclusion

With the purpose of analysing the spatiotemporal characteristics of air quality data more accurately, this study proposed a new air quality forecast model based on a spatiotemporal embedding and self-attention. In the proposed spatiotemporal embedding and self-attention processes, the correlation characteristics were fully considered in the spatial, temporal and parameter dimensions, and therefore, the proposed forecast model could sufficiently analyse the highly dynamic characteristics of air quality data. The proposed forecast model was evaluated by the air quality data of Beijing, China. In the from 1 h to 20 h ahead air quality prediction, the proposed model provided a more accurate result than other models. Moreover, experiments showed that the proposed model could provide a highly stable and accurate forecast at a single monitoring station and at a single time point. These results proved that our model could be used as a powerful tool for air quality forecasting.

However, this study was conducted for only one city area, and so the robustness of the proposed forecast model was not sufficiently evaluated. Additionally, the proposed model was only limited to a short-term air quality forecasting.

Our future studies will aim to conduct a long-term or countrywide air quality forecasting, and the forecasting for other environmental indicators, such as water quality, rainfall, and solar radiation.

Credit author statement

Yang Feng: Conceptualization, Methodology, Writing–Original draft preparation, Ju-Song Kim: Conceptualization, Methodology, Writing–Review and editing, Jin-Won Yu: Conceptualization, Methodology, Writing–Original draft preparation, Kuk-Chol Ri: Writing–Review and editing, Song-Jun Yun: Data collection and curation, Software, Investigation, Il-Nam Han: Methodology, Data Process, Visualization, Zhanfeng Qi: Writing–Review and editing, Supervision, Xiaoli Wang: Writing–Review and editing, Supervision.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgement

We would like to acknowledge the significant support by The National Natural Science Foundation of China (Grant No. 42006168), The Key Research and Development Program of Tianjin (Grant No. 20YFZCSN01040) and Xinjiang Air Pollution Prevention and Control Program – Emergency Control Capability Development Project for Heavily Polluted Weather in Xinjiang (grant No. GK 2022-137).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2023.122402.

References

Al-qaness, M.A.A., Dahou, A., Ewees, A.A., Abualigah, L., Huai, J.Z., Abd Elaziz, M., Helmi, A.M., 2023. ResInformer: residual transformer-based artificial time-series forecasting model for PM2.5 concentration in three major Chinese cities. Mathematics 11, 476. https://doi.org/10.3390/math11020476.

Antanasijevic, D.Z., Pocajt, V.V., Povrenovic, D.S., Ristic, M.D., Peric-Grujic, A.A., 2013. PM10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci. Total Environ. 443, 511–519. https://doi.org/10.1016/j.scitotenv.2012.10.110.

Athira, V., Geetha, P., Vinayakumar, P., Soman, K.P., 2018. DeepAirNet: applying recurrent networks for air quality prediction. Procedia Comput. Sci. 132, 1394–1403. https://doi.org/10.1016/j.procs.2018.05.068.

Bai, Y., Li, Y., Zeng, B., Li, C., Zhang, J., 2019. Hourly PM2.5 concentration forecast using stacked autoencoder model with emphasis on seasonality. J. Clean. Prod. 224, 739–750. https://doi.org/10.1016/j.jclepro.2019.03.253.

Bilal Bashir, M.F., Benghoul, M., Numan, U., Shakoor, A., Komal, B., Bashir, M.A., Bashir, M., Tan, D.J., 2020. Environmental pollution and COVID-19 outbreak: insights from Germany. Air. Qual. Atmos. Hlth. 13, 1385–1394. https://doi.org/10.1007/s11869-020-00893-9.

BMBEE China (Beijing Municipal Bureau of Ecology and Environment, China), 2023. Beijing Ecological and Environmental Bulletin 2022. https://sthjj.beijing.gov.cn/bjhrb/index/xxgk69/sthjlyzwg/1718880/1718881/1718882/326119689/index.html.

Cai, K., Li, S.S., Zheng, F.B., Yu, C., Zhang, X.Y., Liu, Y., Li, Y.J., 2018. Spatio-temporal variations in NO2 and PM2.5 over the central plains economic region of China during 2005-2015 based on satellite observations. Aerosol Air Qual. Res. 18, 1221–1235. https://doi.org/10.4209/aaqr.2017.10.0394.

Che, Z.P., Purushotham, S., Cho, K., Sontag, D., Liu, Y., 2018. Recurrent neural networks for multivariate time series with missing values. Sci. Rep. 8, 6085. https://doi.org/10.1038/s41598-018-24271-9.

Chen, Y.B., Chen, X.M., Xu, A.L., Sun, Q., Peng, X.Y., 2022. A hybrid CNN-Transformer model for ozone concentration prediction. Air. Qual. Atmos. Hlth. 15, 1533–1546. https://doi.org/10.1007/s11869-022-01197-w.

Du, S.D., Li, T.R., Yang, Y., Horng, S.J., 2021. Deep air quality forecasting using hybrid deep learning framework. Ieee T. Knowl. Data En. 33, 2412–2424. https://doi.org/10.1109/tkde.2019.2954510.

Feng, R., Zheng, H.J., Gao, H., Zhang, A.R., Huang, C., Zhang, J.X., Luo, K., Fan, J.R., 2019. Recurrent neural network and random forest for analysis and accurate forecast of atmospheric pollutants: a case study in Hangzhou, China. J. Clean. Prod. 231, 1005–1015. https://doi.org/10.1016/j.jclepro.2019.05.319.

Garaga, R., Sahu, S.K., Kota, S.H., 2018. A review of air quality modeling studies in India: local and regional scale. Curr. Pollut. Rep. 4, 59–73. https://doi.org/10.1007/s40726-018-0081-0.

Garcia-Laencina, P.J., Sancho-Gomez, J.L., Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: a review. Neural Comput. Appl. 19, 263–282. https://doi.org/10.1007/s00521-009-0295-6.

IQAir, 2023. 2022 World Air Quality Report: Region & City PM2.5 Ranking. https://www.iqair.com/world-air-quality-report.

Jain, S., Presto, A.A., Zimmerman, N., 2021. Spatial modeling of daily PM2.5, NO2, and CO concentrations measured by a low-cost sensor network: comparison of linear, machine learning, and hybrid land use models. Environ. Sci. Technol. 55, 8631–8641. https://doi.org/10.1021/acs.est.1c02653.

Jiang, P., Yang, J., Huang, C.H., Liu, H.K., 2018. The contribution of socioeconomic factors to PM2.5 pollution in urban China. Environ. Pollut. 233, 977–985. https://doi.org/10.1016/j.envpol.2017.09.090.

Junger, W.L., de Leon, A.P., 2015. Imputation of missing data in time series for air pollutants. Atmos. Environ. 102, 96–104. https://doi.org/10.1016/j.atmosenv.2014.11.049.

Kamal, M.M., Jailani, R., Shauri, R.L.A., 2006. Prediction of ambient air quality based on neural network technique. Conference on Research & Development 115–119. https://doi.org/10.1109/SCORED.2006.4339321.

Karambelas, A., Holloway, T., Kiesewetter, G., Heyes, C., 2018. Constraining the uncertainty in emissions over India with a regional air quality model evaluation. Atmos. Environ. 174, 194–203. https://doi.org/10.1016/j.atmosenv.2017.11.052.

Kim, J.S., Wang, X.L., Kang, C., Yu, J.W., Li, P.H., 2021. 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.

Kimura, N., Saeki, T., 2020. End-to-end deep learning speech recognition model for silent speech challenge. Interspeech 2020, 1024–1026.

Kong, L.Q., Liu, Z.F., Wu, J.G., 2020. A systematic review of big data-based urban sustainability research: state-of-the-science and future directions. J. Clean. Prod. 273, 123142 https://doi.org/10.1016/j.jclepro.2020.123142.

Kow, P.Y., Wang, Y.S., Zhou, Y.L., Kao, I.F., Issermann, M., Chang, L.C., Chang, F.J., 2020. Seamless integration of convolutional and back-propagation neural networks for regional multi-step-ahead PM2.5 forecasting. J. Clean. Prod. 261, 121285 https://doi.org/10.1016/j.jclepro.2020.121285.

Kumar, A., Patil, R.S., Dikshit, A.K., Islam, S., Kumar, R., 2016. Evaluation of control strategies for industrial air pollution sources using American meteorological society/environmental protection agency regulatory model with simulated meteorology by weather research and forecasting model. J. Clean. Prod. 116, 110–117. https://doi.org/10.1016/j.jclepro.2015.12.079.

Kumar, U., De Ridder, K., 2010. GARCH modelling in association with FFT-ARIMA to forecast ozone episodes. Atmos. Environ. 44, 4252–4265. https://doi.org/10.1016/j.atmosenv.2010.06.055.

Leong, W.C., Kelani, R.O., Ahmad, Z., 2020. Prediction of air pollution index (API) using support vector machine (SVM). J. Environ. Chem. Eng. 8, 103208 https://doi.org/10.1016/j.jece.2019.103208.

Li, X., Peng, L., Yao, X.J., Cui, S.L., Hu, Y., You, C.Z., Chi, T.H., 2017. Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environ. Pollut. 231, 997–1004. https://doi.org/10.1016/j.envpol.2017.08.114.

Liao, K., Huang, X.H., Dang, H.F., Ren, Y., Zuo, S.D., Duan, C.S., 2021. Statistical approaches for forecasting primary air pollutants: a review. Atmosphere-basel 12, 686. https://doi.org/10.3390/atmos12060686.

Liu, H., Yin, S., Chen, C., Duan, Z., 2020. Data multi-scale decomposition strategies for air pollution forecasting: a comprehensive review. J. Clean. Prod. 277, 124023 https://doi.org/10.1016/j.jclepro.2020.124023.

Ma, J., Ding, Y.X., Cheng, J.C.P., Jiang, F.F., Tan, Y., Gan, V.J.L., Wan, Z.W., 2020. Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. J. Clean. Prod. 244, 118955 https://doi.org/10.1016/j.jclepro.2019.118955.

Ma, J., Ding, Y.X., Cheng, J.C.P., Jiang, F.F., Wan, Z.W., 2019. A temporal-spatial interpolation and extrapolation method based on geographic long short-term memory neural network for PM2.5. J. Clean. Prod. 237, 117729 https://doi.org/10.1016/j.jclepro.2019.117729.

Ma, Z.W., Hu, X.F., Huang, L., Bi, J., Liu, Y., 2014. Estimating ground-level PM2.5 in China using satellite remote sensing. Environ. Sci. Technol. 48, 7436–7444. https://doi.org/10.1021/es5009399.

Masmoudi, S., Elghazel, H., Taieb, D., Yazar, O., Kallel, A., 2020. A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. Sci. Total Environ. 715, 136991 https://doi.org/10.1016/j.scitotenv.2020.136991.

MEE China (Ministry of Ecology and Environment, China), 2012a. Ambient Air Quality Standards (GB 3095-2012). https://sthjj.beijing.gov.cn/bjhrb/index/xxgk69/sthjlyzwg/1718880/1718881/1718882/326119689/index.html.

MEE China (Ministry of Ecology and Environment, China), 2012b. Technical Regulation on Ambient Air Quality Index. HJ-633-2012). https://sthjj.beijing.gov.cn/bjhrb/index/xxgk69/sthjlyzwg/1718880/1718881/1718882/326119689/index.html.

Nieto, P.J.G., Lasheras, F.S., Garcia-Gonzalo, E., Juez, F.J.D., 2018. PM10 concentration forecasting in the metropolitan area of Oviedo (northern Spain) using models based on SVM, MLP, VARMA and ARIMA: a case study. Sci. Total Environ. 621, 753–761. https://doi.org/10.1016/j.scitotenv.2017.11.291.

Pelaez, L.M.G., Santos, J.M., Albuquerque, T.T.D., Reis, N.C., Andreao, W.L., Andrade, M.D., 2020. Air quality status and trends over large cities in south America. Environ. Sci. Pol. 114, 422–435. https://doi.org/10.1016/j.envsci.2020.09.009.

Qi, Y.L., Li, Q., Karimian, H., Liu, D., 2019. A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. Sci. Total Environ. 664, 1–10. https://doi.org/10.1016/j.scitotenv.2019.01.333.

Rijal, N., Gutta, R.T., Cao, T., Lin, J., Bo, Q., Zhang, J., 2018. Ensemble of deep neural networks for estimating particulate matter from images. 3rd International Conference on Image, Vision and Computing (ICIVC) 733–738. https://doi.org/10.1109/ICIVC.2018.8492790.

Saide, P.E., Carmichael, G.R., Spak, S.N., Gallardo, L., Osses, A.E., Mena-Carrasco, M.A., Pagowski, M., 2011. 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.

Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A., Jung, J., 2020. Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance. Neural Network. 121, 396–408. https://doi.org/10.1016/j.neunet.2019.09.033.

Soh, P.W., Chang, J.W., Huang, J.W., 2018. Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access 6, 38186–38199. https://doi.org/10.1109/access.2018.2849820.

Song, Y., Huang, Z.L., Shen, C.Y., Shi, H., Lange, D.A., 2020. Deep learning-based automated image segmentation for concrete petrographic analysis. Cement Concr. Res. 135, 106118 https://doi.org/10.1016/j.cemconres.2020.106118.

Stadlober, E., Hormann, S., Pfeiler, B., 2008. Quality and performance of a PM10 daily forecasting model. Atmos. Environ. 42, 1098–1109. https://doi.org/10.1016/j.atmosenv.2007.10.073.

Stern, R., Builtjes, P., Schaap, M., Timmermans, R., Vautard, R., Hodzic, A., Memmesheimer, M., Feldmann, H., Renner, E., Wolke, R., Kerschbaumer, A., 2008. A model inter-comparison study focussing on episodes with elevated PM10 concentrations. Atmos. Environ. 42, 4567–4588. https://doi.org/10.1016/j.atmosenv.2008.01.068.

Sun, X.T., Xu, W., Jiang, H.X., Wang, Q.L., 2021. A deep multitask learning approach for air quality prediction. Ann. Oper. Res. 303, 51–79. https://doi.org/10.1007/s10479-020-03734-1.

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Proceedings of the 31st Conference on Neural Information Processing Systems 30. https://doi.org/10.48550/arXiv.1706.03762.

Wang, D.S., Wang, H.W., Lu, K.F., Peng, Z.R., Zhao, J.H., 2022a. Regional prediction of ozone and fine particulate matter using diffusion convolutional recurrent neural network. Int. J. Environ. Res. Publ. Health 19, 3988. https://doi.org/10.3390/ijerph19073988.

Wang, Z.Y., Yang, Y.D., Yue, S.L., 2022b. Air quality classification and measurement based on double output vision transformer. Ieee Internet Things 9, 20975–20984. https://doi.org/10.1109/jiot.2022.3176126.

WHO, 2021a. New WHO Global Air Quality Guidelines Aim to Save Millions of Lives from Air Pollution. https://www.who.int/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution.

WHO, 2021b. WHO Global Air Quality Guidelines. Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, Geneva, 2021. Licence: CC BY-NC-SA 3.0 IGO. https://www.who.int/publications/i/item/9789240034228.

Wu, C.L., He, H.D., Song, R.F., Peng, Z.R., 2022a. Prediction of air pollutants on roadside of the elevated roads with combination of pollutants periodicity and deep learning method. Build. Environ. 207, 108436 https://doi.org/10.1016/j.buildenv.2021.108436.

Wu, Z.J., Ma, C., Shi, X.C., Wu, L.B., Dong, Y., Stojmenovic, M., 2022b. Imputing missing indoor air quality data with inverse mapping generative adversarial network. Build. Environ. 215, 108896 https://doi.org/10.1016/j.buildenv.2022.108896.

Xu, C.L., Xie, L., Xiao, X., 2018. A bidirectional LSTM approach with word embeddings for sentence boundary detection. J. Signal Process. Sys. 90, 1063–1075. https://doi.org/10.1007/s11265-017-1289-8.

Xu, R., Deng, X.L., Wan, H., Cai, Y.P., Pan, X.P., 2021. A deep learning method to repair atmospheric environmental quality data based on Gaussian diffusion. J. Clean. Prod. 308, 127446 https://doi.org/10.1016/j.jclepro.2021.127446.

Xu, R., Wang, D.K., Li, J., Wan, H., Shen, S.M., Guo, X., 2023. A hybrid deep learning model for air quality prediction based on the time-frequency domain relationship. Atmosphere-basel 14, 405. https://doi.org/10.3390/atmos14020405.

Xu, X.H., Yoneda, M., 2021. Multitask air-quality prediction based on LSTM-autoencoder model. Ieee T. Cybernetics 51, 2577–2586. https://doi.org/10.1109/tcyb.2019.2945999.

Yan, R., Liao, J.Q., Yang, J., Sun, W., Nong, M.Y., Li, F.P., 2021. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Syst. Appl. 169, 114513 https://doi.org/10.1016/j.eswa.2020.114513.

Yao, Y., Pan, J.H., Liu, Z.X., Meng, X., Wang, W.D., Kan, H.D., Wang, W.B., 2021. Ambient nitrogen dioxide pollution and spreadability of COVID-19 in Chinese cities. Ecotoxicol. Environ. Saf. 208, 111421 https://doi.org/10.1016/j.ecoenv.2020.111421.

Yu, M.Z., Masrur, A., Blaszczak-Boxe, C., 2023. Predicting hourly PM2.5 concentrations in wildfire-prone areas using a spatio temporal transformer model. Sci. Total Environ. 860, 160446 https://doi.org/10.1016/j.scitotenv.2022.160446.

Zhan, Y., Luo, Y.Z., Deng, X.F., Grieneisen, M.L., Zhang, M.H., Di, B.F., 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ. Pollut. 233, 464–473. https://doi.org/10.1016/j.envpol.2017.10.029.

Zhang, B., Rong, Y., Yong, R.H., Qin, D.M., Li, M.Z., Zou, G.J., Pan, J.G., 2022a. Deep learning for air pollutant concentration prediction: a review. Atmos. Environ. 290, 119347 https://doi.org/10.1016/j.atmosenv.2022.119347.

Zhang, L.Y., Lin, J., Qiu, R.Z., Hu, X.S., Zhang, H.H., Chen, Q.Y., Tan, H.M., Lin, D.T., Wang, J.K., 2018. Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model. Ecol. Indicat. 95, 702–710. https://doi.org/10.1016/j.ecolind.2018.08.032.

Zhang, Q., Han, Y., Li, V.O.K., Lam, J.C.K., 2022b. Deep-AIR: a hybrid CNN-LSTM framework for fine-grained air pollution estimation and forecast in metropolitan cities. IEEE Access 10, 55818–55841. https://doi.org/10.1109/access.2022.3174853.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Zhang, W., 2021. Informer: beyond efficient transformer for long sequence time-series forecasting. Artif. Intell. 1–9 https://doi.org/10.48550/arXiv.2012.07436.

Zhou, Y.H., Yang, G.F., 2022. A predictive model of indoor PM2.5 considering occupancy level in a hospital outpatient hall. Sci. Total Environ. 844, 157233 https://doi.org/10.1016/j.scitotenv.2022.157233.

Zhou, Y.L., Chang, F.J., Chang, L.C., Kao, I.F., Wang, Y.S., Kang, C.C., 2019. Multi-output support vector machine for regional multi-step-ahead PM2.5 forecasting. Sci. Total Environ. 651, 230–240. https://doi.org/10.1016/j.scitotenv.2018.09.111.