Computers & Geosciences 155 (2021) 104869

Contents lists available at ScienceDirect

Computers and Geosciences

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

Spatiotemporal causal convolutional network for forecasting hourly $PM_{2.5}$ concentrations in Beijing, China

Lei Zhang [a,g], Jiaming Na [b,c,d,g,*], Jie Zhu [b,c], Zhikuan Shi [e], Changxin Zou [f], Lin Yang [a]

[a] School of Geography and Ocean Science, Nanjing University, Nanjing, 210023, China
[b] College of Civil Engineering, Nanjing Forestry University, Nanjing, 210037, China
[c] Key Laboratory of Virtual Geographic Environment, Ministry of Education, 210023, Nanjing, China
[d] School of Geography, Nanjing Normal University, 210023, Nanjing, China
[e] College of Land Management, Nanjing Agricultural University, Nanjing, 210095, China
[f] Nanjing Institute of Environmental Sciences, Ministry of Ecology and Environment, Nanjing, 210042, China
[g] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, 210023, China

ARTICLE INFO

Keywords:
$PM_{2.5}$ prediction
Air pollution
Causal convolutional network
Spatiotemporal correlation
Deep learning

ABSTRACT

Air pollution in Northeastern Asia is a serious environmental problem, especially in China where $PM_{2.5}$ levels are quite high. Accurate $PM_{2.5}$ predictions are significant to environmental management and human health. Recently, deep learning has received increasing attention from relevant researchers. In this work, a spatiotemporal causal convolutional neural network (ST-CausalConvNet) for short-term $PM_{2.5}$ prediction is proposed. The distinguishing characteristics of the proposed model is that the convolutions in the model architecture are causal, where an output at a certain time step is convolved only with elements from the same or earlier time steps in the previous layer. Accordingly, no information leakage is induced from the future to the past in this model. The spatial dependence between multiple monitoring stations was also considered in the model. Spatiotemporal correlation analysis was performed to select relevant information from monitoring stations that have a high relationship with the target station. The information from the target and related stations were then employed as the inputs and fed into the model. A case study from May 1, 2014 to April 30, 2015 in Beijing, China was conducted. The next hour $PM_{2.5}$ concentration was predicted by the proposed model by using historical air quality and meteorological data from 36 monitoring stations. Experimental results show that the trends of the predicted $PM_{2.5}$ concentrations and the observed values were consistent. The proposed method achieved a better prediction performance than the other three comparative models, namely artificial neural network (ANN), gated recurrent unit (GRU), and long short-term memory (LSTM). Furthermore, the effects of the important parameters and the model transferability were also conducted. We conclude that the proposed ST-CausalConvNet is a potential effective model for air pollution forecasting.

1. Introduction

Air pollution has become a particularly important social issue and received increasing attention from researchers due to rapid industrialisation and urbanisation. Atmospheric particulate matter (PM), a mixture of solid and aqueous species, enters the atmosphere via natural pathways or anthropogenically (World Health Organisation (WHO), 2003; Xing et al., 2016). $PM_{2.5}$ refers to particles with an aerodynamic diameter less of than 2.5 μm ($PM_{2.5}$), one of the dominant sources of harmful air pollution. This particle has a bad influence on our living environment and physical health. The International Agency for Research on Cancer (IARC) designated air pollution as a human carcinogen for the first time in 2013 and regards it as a general and major environmental carcinogen (Loomis et al., 2013). In 2016, the WHO reported projected that an estimated 4.1 million premature deaths worldwide are related to exposure to $PM_{2.5}$, mainly from lung cancer, heart disease, heart stroke, respiratory infections and chronic lung disease (WHO, 2016). Therefore, air quality forecasting has become an essential research topic for atmospheric environmental protection, public health and assisting government managers in scientific decisions (Di et al., 2019; Qiao et al., 2019; Chen et al., 2019a, 2020). Short-term exposure to air pollutants will increase the risk of cardiovascular and respiratory diseases,

* Corresponding author. Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, 210023, Nanjing, China.
E-mail addresses: zhanglei@smail.nju.edu.cn (L. Zhang), najiaming92@gmail.com (J. Na).

https://doi.org/10.1016/j.cageo.2021.104869
Received 10 October 2020; Received in revised form 24 May 2021; Accepted 22 June 2021
Available online 26 June 2021
0098-3004/

according to a growing body of research (Ma et al., 2017; Liu et al., 2018; Ai et al., 2019). However, the daily level measurement or prediction might not fully capture the immediate health effects of ambient air pollution (Bhaskaran et al., 2011). Thus, a more accurate time scale prediction will provide the basis for further air pollution studies.

Many initiatives have been dedicated to researching effective approaches to $PM_{2.5}$ forecasting. These initiatives can be basically categorized into deterministic and statistical approaches. Deterministic approaches (Byun, 1999; Grell et al., 2005; Kim et al., 2010) are based on the knowledge of the formation and diffusion of pollutants. The understanding of the atmospheric physics and chemical processes is used as the fundamental theory for the prediction of mass concentrations. However, these deterministic methods heavily depend on theoretical assumptions, and key knowledge of the physics process may be insufficient, making the explanation of the nonlinearity and heterogeneity of many influence factors difficult (Siwek and Osowski, 2016; Cabaneros et al., 2019; Pak et al., 2020). This problem leads to bias in the air pollution prediction result.

Meanwhile, statistical approaches based on the data-driven strategy have received extensive interest. This type of method requires a large amount of historical data from monitoring stations. Statistical methods explore the quantitative relationship between historical environmental covariates and future air quality under the support of a large quantity of historical data from the monitoring stations. Regression models in machine learning methods are the commonly used predictors for extracting this relationship, which is also termed as the statistical learning approach (James et al., 2013). The statistical learning task learns a functional relationship $f : \mathscr{X} \cdot \mathscr{Y}$ from the training dataset, where $\mathscr{X}$ is the feature space, which includes the historical values of air pollutant concentrations, weather, traffic and other relevant factors; and $\mathscr{Y}$ is the observed values of $PM_{2.5}$ concentrations at a specific time in the future. A variety of statistical learning models have been studied for air pollution prediction. Several linear statistical models have been proposed for PM prediction (Castellano et al., 2009; de Gennaro et al., 2013; Donnelly et al., 2015; Xiao et al., 2018). Given that air pollution forecasting is affected by various environmental factors, calibrating a linear model with a highly accurate predicted performance is difficult. Nonlinear models, such as random forest (Hu et al., 2017; Stafoggia et al., 2019; Wei et al., 2019), support vector machine (Sun and Liu, 2016; Liu et al., 2019) and artificial neural networks (ANNs) (Pérez et al., 2000; Perez and Reyes, 2006; Yildirim and Bayramoglu, 2006; Zhou et al., 2014), have been widely applied. In comparison with conventional linear models, these nonlinear models expand the hypothesis space of the model, which is larger than that of linear models, thereby improving the prediction performance (Pérez et al., 2000; Chaloulakou et al., 2003; Chu et al., 2016; Delavar et al., 2019).

Recently, deep learning models have shown great potential in the air pollution forecasting task (Ong et al., 2014; Ahn et al., 2017; Pak et al., 2018). The deep learning method can discover the underlying intricate nonlinear structures in high-dimensional data and is applicable to many fields of science (LeCun et al., 2015; Goodfellow et al., 2016). The multiple processing layer structure enables deep learning models to learn the data representation by using hierarchical abstractions (LeCun et al., 2015; Schmidhuber, 2015), addressing the bottleneck of traditional statistical methods. Given that air pollution forecasting is a typical multivariate time series analysis problem, deep learning models, such as recurrent neural networks (RNNs), can create and process memories of arbitrary sequences of input patterns. Hence, some variants of RNNs (e.g. long short-term memory [LSTM] (Hochreiter and Schmidhuber, 1997) and gated recurrent unit [GRU] (Cho et al., 2014) networks) have been successfully developed with gating mechanisms for predicting air quality in some previous works (Ong et al., 2014; Li et al., 2016; Ahn et al., 2017). Hybrid models that combine the convolutional neural network (CNN) and LSTM were studied, exhibiting good performance in predicting $PM_{2.5}$ concentration (Huang and Kuo, 2018; Pak et al., 2018, 2020). Given that the emission and diffusion of air pollution are spatially and temporally correlated, some recent studies have also considered the spatiotemporal correlations of time series datasets from multiple monitoring stations at different locations and achieved a more accurate and stable prediction performance (Qi et al., 2019; Pak et al., 2020).

These successful studies have inspired many deep learning works to apply recurrent architectures for time series modeling tasks. Hence, most current deep learning models for air pollution forecasting take RNNs, such as GRUs or LSTMs, as the core component units. However, some recent studies have shown that using the CNN architecture alone can also effectively deal with sequential problems (Oord et al., 2016; Bai et al., 2018). Some results indicated that certain well-designed convolutional architectures were used prominently well in various tasks, such as document understanding (Kalchbrenner et al., 2014; Kim, 2014; Johnson and Zhang 2015, 2017), audio synthesis (Oord et al., 2016), machine translation (Kalchbrenner et al., 2016; Gehring et al., 2016, 2017) and language modeling (Dauphin et al., 2017). These networks designed to deal with time-series problems can be summarised as a family of architectures and termed as temporal convolutional networks (TCNs) (Lea et al., 2017; Bai et al., 2018). In comparison with the conventional modeling architecture of convolutional or recurrent neural networks, the distinguishing characteristic of TCNs is that the convolutions in this type of architectures are causal. The predictions by TCNs at a timestep depend only on the information at previous timesteps and not on any of the future timesteps. Accordingly, the causal constraint is satisfied, and no information leakage is induced from the future to the past. Meanwhile, the causal convolutions strategy does not contain complicated recurrent structures, such as the elaborate architectures of GRU or LSTM with gating mechanisms. Consequently, the causal convolutional network (CausalConvNet), as the core structure in TCNs, is simpler and clearer than RNN models (Bai et al., 2018). Classic convolutional or recurrent deep networks are extensively applied for predicting air pollution, whilst the study of the application of the causal convolutional network in air pollution forecasting is still insufficient. We believe that the intrinsic "causal" characteristic in the architecture of CausalConvNet might help with the $PM_{2.5}$ prediction task. Hence, it is necessary to verify if this type of network can be an alternative effective model for the $PM_{2.5}$ prediction task. The current developed CausalConvNets in TCNs were originally designed for audio generation or natural language modeling in the field of computer science. The characteristics of the spatial dependence is widely considered in air pollution prediction tasks (Cabaneros et al., 2019; Wen et al., 2019) and other geospatial research problems, but not in the current developed CausalConvNets. Therefore, research into the manner by which to combine the CausalConvNet architecture with the spatial dependence between multiple monitoring stations is needed.

With the above-mentioned motivations, our research objective is to overcome the aforementioned limitations in the existing studies and establish a deep network, namely, spatiotemporal causal convolutional network (ST-CausalConvNet), for air pollution forecasting. The proposed model adopts the causal convolutions strategy without complicated RNN structures with gating mechanisms. We extended the current existing CausalConvNets to include the consideration of the spatiotemporal relationship between the historical information from multiple nearby monitoring stations. A case study in Beijing, China was conducted to validate the proposed method of predicting $PM_{2.5}$ concentrations.

2. Materials and method

2.1. Dataset

Beijing, the capital city of China, was selected as the study area in this work (Fig. 1). Beijing has a typical humid subtropical climate characterised by hot and humid summer, cold and dry winter and relatively short spring and fall. Beijing has become one of the most seriously air-polluted cities in the world due to the rapid

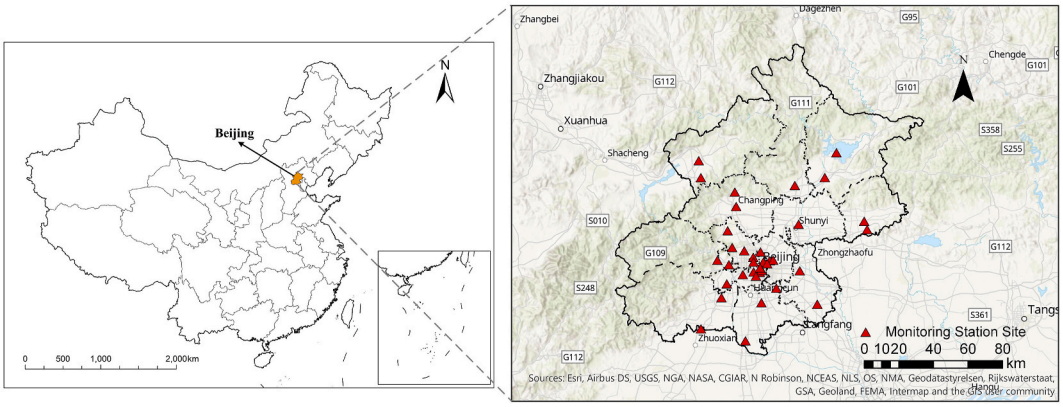

Fig. 1. Monitoring station map for air pollution in Beijing.

industrialisation and urbanisation (Zíková et al., 2016; Yan et al., 2018; Zhang et al., 2018) and attracted extensive studies in this city (e.g., Zheng et al., 2015; Chen et al., 2019c; Pak et al., 2020).

The air-quality forecast dataset from the Urban Computing Team in Microsoft Research was used in this study (Zheng et al., 2015). This dataset has numerous input variables with high quality and integrity. The dataset includes 278,023 air quality observation instances from the in-city monitoring stations and 116,867 meteorological observations instances from 17 in-city monitoring stations from May 1, 2014 to April 30, 2015. All the data are geo-coordinated with latitude and longitude, and the data description is shown in Table 1.

2.2. Spatiotemporal correlation analysis

The air pollution in an area spatially depends on that in other areas due to the spatiotemporal correlation among the monitoring stations. The model performance is limited if only the historical data at a single station is considered. Spatiotemporal correlation analysis must be performed on the data at multiple stations to overcome this problem.

Given the historical data set of N monitoring stations, the input features can be expressed as a 3D vector, $\boldsymbol{X} \in \mathbb{R}^{N\times T\times M}$, where T and Mrepresent the numbers of the time steps and the environmental features related to the $PM_{2.5}$ concentrations. $\boldsymbol{X}$ can also be expressed as a vector of N feature matrices, $[\boldsymbol{X}_{s_1}, \ldots, \boldsymbol{X}_{s_i}, \ldots, \boldsymbol{X}_{s_N}]$, where $\boldsymbol{X}_{s_i} \in \mathbb{R}^{T\times M}$ represents the feature matrix of the ith monitoring station.

Table 1
Dataset description in this study.

Variable type	Variable name	Data type	Unit
Air quality data	$PM_{2.5}$	Numeric	μg/m^3
	PM_{10}	Numeric	μg/m^3
	NO_2	Numeric	μg/m^3
	CO	Numeric	mg/m^3
	O_3	Numeric	μg/m^3
	SO_2	Numeric	μg/m^3
Meteorological data	Weather	Categorical (Sunny/Cloudy/Overcast/Foggy/Snowy/Rainy)	,,
	Temperature	Numeric	°C
	Pressure	Numeric	hPa
	Relative humidity	Numeric	%
	Wind speed (wind_speed)	Numeric	m/s
	Wind direction (wind_direction)	Categorical (No/E/W/S/N/Unstable/SE/NE/SW/NW)	,,

We firstly calculate the correlation coefficient for the time series between the target station and the other stations. We use the Pearson correlation coefficient to represent the influence degree between two stations as follows:

$$\rho(Y_*, Y_i) = \frac{\mathrm{Cov}(Y_*, Y_i)}{\sigma_{Y_*}\sigma_{Y_i}}, \tag{1}$$

where Y_* and Y_i represent the $PM_{2.5}$ concentrations in a time series at the target station and the ith station, respectively; $\mathrm{Cov}(\cdot)$ is the covariance function; and σ_{Y_*} and σ_{Y_i} are the variances of Y_* and Y_i, respectively.

The correlation of all other stations to the target station is organised into a correlation vector:

$$\boldsymbol{\rho}_* = [\rho(Y_*, Y_1), \rho(Y_*, Y_2), \ldots, \rho(Y_*, Y_i), \ldots, \rho(Y_*, Y_N)]. \tag{2}$$

Given that not all other stations have an evident impact on the target station, setting a correlation threshold to extract the data of stations, which have a relatively high influence on the target station, is rational. Thus, the final input feature vector can be determined as follows:

$$\boldsymbol{X}_* = \{\boldsymbol{X}_i \mid \rho(Y_*, Y_i) > \rho_{th},\ i \in 1, \ldots, N\}, \tag{3}$$

where ρ_{th} is a user-defined correlation threshold, and $\boldsymbol{X}_*$ is the filtered input feature vector that represents the spatiotemporal information of the target station.

After determining $\boldsymbol{X}_*$, it can be used as the input to the proposed model.

2.3. Architecture of the proposed network

The architecture of the proposed ST-CausalConvNet is illustrated in Fig. 2. The model has two parts. The first part is the integration of the spatiotemporal information of multiple monitoring stations, which is shown in Fig. 2(A); the second part is the causal convolutional network, which is shown Fig. 2(B). The details of these two parts are described in next two subsections.

2.3.1. Integration of spatiotemporal information of multiple monitoring stations

Given input feature $\boldsymbol{X}_*$ which is a 3D vector containing the spatiotemporal data of different stations, integrating the input information of $\boldsymbol{X}_*$ on the dimension of the multiple stations is intuitive. We use 1×1 convolutions to achieve this purpose. Lin et al. (2013) was the first to propose 1×1 convolutions for the deep neural network, incrementing and decrementing the dimension of the input features. In Fig. 2(A), the values of the first feature at the first time step (marked in blue) were taken as an example. These set of values can be regarded as 1×1 multi-channel information, and the number of channels is equal to the

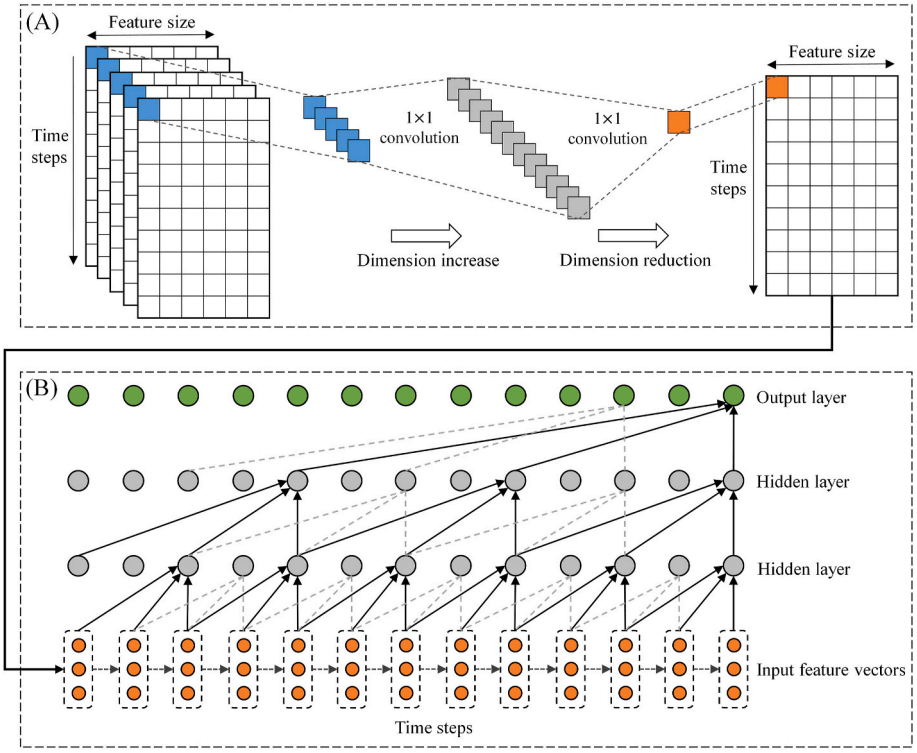

Fig. 2. Architecture of the proposed ST-CausalConvNet, which includes two main parts: (A) integration of the spatiotemporal information of multiple monitoring stations; (B) causal convolutional network with kernel size = 3 and d (dilations) = 1, 2, and 4.

stations selected in X_*. We used multiple 1 × 1 convolution kernels to expand the dimension of this multi-channel information, as shown by the feature vector marked in gray. Then, we adopted one 1 × 1 convolution kernel to reduce the dimension to one, which is marked in orange. This strategy can not only improve the feature extraction ability on the information of multiple monitoring stations by dimension increment but also aggregate the input data into a 2D feature vector, which is compatible with the input of the causal network in the second part. Other studies also proved that 1 × 1 convolution can be regarded as a dimension increment or decrement module for effectively improving the model capacity (increasing the nonlinearity) and removing computational bottlenecks (Szegedy et al., 2015; He et al., 2016).

2.3.2. Causal convolutional network

Given the integrated historical time series data, the proposed causal convolutional network predicts the future $PM_{2.5}$ concentrations of the next hour. Considering the time series problem, an ordinary CNN architecture is unsuitable because it does not consider the time dependence. Therefore, a causal convolutions architecture, a special CNN structure, was designed to deal with the sequence modeling problem. Fig. 2 shows that the causal convolutional network uses a multi-layer fully-convolutional network architecture. In contrast with the ordinary CNN, the input features from earlier than a specific time are used for the convolution in causal convolution layers. Consequently, some previous studies concluded that increasing the network depth can effectively improve the model performance (Bianchini and Scarselli, 2014; Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016) and incorporated the mechanism for multi-layer processing into the model. The potential benefit is that the abstracted representation of the input features can be effectively learned and extracted layer-wise through the layer-by-layer processing.

Although multi-layer causal convolutions can expand the coverage of the historical data, they can only look back on history with the depth of the network in a linear timescale. Thus, the information of a long history is difficult to capture. To deal with this problem, we employed dilated convolutions to allow a large receptive field following the works of Oord et al. (2016) and Bai et al. (2018) (Yu and Koltun, 2016). Thus, given filter function f with size k, the dilated convolution (F) for time series input x is defined as follows:

$$F(s) = (x *_d f)(s) = \sum_{i}^{k-1} f(i) \cdot x_{s-d \cdot i}, \tag{4}$$

where s is the element of the sequence, d is the dilation parameter, and $s - d \cdot i$ describes the direction of the past. We refer to $*_d$ as the dilated convolution operator or a d-dilated convolution operator to distinguish from the normal convolution operation. The normal convolution operator (*) is a specific version of the dilated convolution (when $d = 1$). A wide range of inputs are represented at the top level with a large dilation, effectively expanding the receptive field of CNN.

According to the illustration shown in Fig. 2(B) and the above description, the receptive field of the proposed causal network increases in two ways: First is to increase filter size k; and second is to increase dilation factor d. In this study, d is exponentially increased according to the network depth $d = 2^i$ to ensure that the long-term history can be effectively covered; specifically, $d = 2^i$ is used for the i levels of network. The number of levels (i.e. the depth of the network) and the filter size are two important model parameters. The parameter analysis is discussed in Section 3.

2.4. Experimental settings

In this section, the effectiveness of the proposed model is validated. The datasets are split into three: training, validation and test sets. We made these three sets evenly distributed over time to ensure the reliability of the validation. The training set consists of the first 60% of each month, and the validation set comprised the next 20% of each month. The remaining 20% of each month is employed as the test set. This data splitting strategy ensures the reliability of the validation method. For

illustration, monitoring station No. 1013 was used as an example. The mean square error (MSE) was selected as the loss function of the training process. We used Adam optimiser (Kingma and Ba, 2014) in the gradient-descent optimisation of the stochastic objective function of our model. We set the training epoch and the batch size to 200 and 32, respectively.

We used the RMSE, MAE and R^2 as the model error evaluation indicators to evaluate the performance of the $PM_{2.5}$ predictor. The calculations of the three metrics are presented in Equations (5)–(7).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}, \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i|, \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2}, \tag{7}$$

where O_i and P_i are the observed and predicted values, respectively; and O_i is the average value of n observed sample data. The smaller the value of RMSE and MAE and the larger the value of R^2 are, the higher the prediction accuracy.

A comparison was carried out in the experiment to verify the superiority of the model performance. Previous studies have shown that ANN can outperform conventional statistical methods, and some RNN architectures, such as GRU or LSTM, have become popular models (Cabaneros et al., 2019; Qi et al., 2019; Pak et al., 2020). We selected the ANN, GRU and LSTM architectures, which are widely used in the air pollution forecasting problem, as the three baselines for comparison with our model. A necessary task is to compare whether the model considers the spatial dependence (i.e. only the historical information at the target station or use of information from other stations that have spatial dependence). We included another simpler model by using the causal convolutional network but exclude between-monitor features, which can test the importance of using the spatially related features.

3. Experimental results

3.1. Spatial dependence analysis result

Based on the spatiotemporal correlation analysis, the spatial distribution of the influence degree of each station towards the target station is shown in Fig. 3. The map shows that the closer the spatial distance between the stations, the stronger the relationship they have. Fig. 4 quantitatively illustrates this notion. A clear negative relationship is observed between the distance from each station to the target station and the influence degree, and the influence degree decreases approximately at a rate of 0.02 per 10 km.

The performance of the prediction was tested by using different correlation threshold values (Table 2). Poor predictions were induced by extremely permissive or strict values. This phenomenon occurs because the number of correlated stations is small when the threshold is large, and the spatial information related to the target station decreases. More related information at other stations can be considered when the threshold is small. However, the noise from the data at other stations that are not highly correlated with the target station increases, and some irrelevant data are taken as the input for the model, thereby causing interference. Table 2 shows that the model achieved the best prediction performance when ρ_{th} was 0.85. Eighteen stations out of the 36 stations have an influence larger than 0.85 on the target station. The average distance of these stations to the target station is 15.0 km, and the longest distance is 37.4 km. We used this threshold value as the default for the following model test.

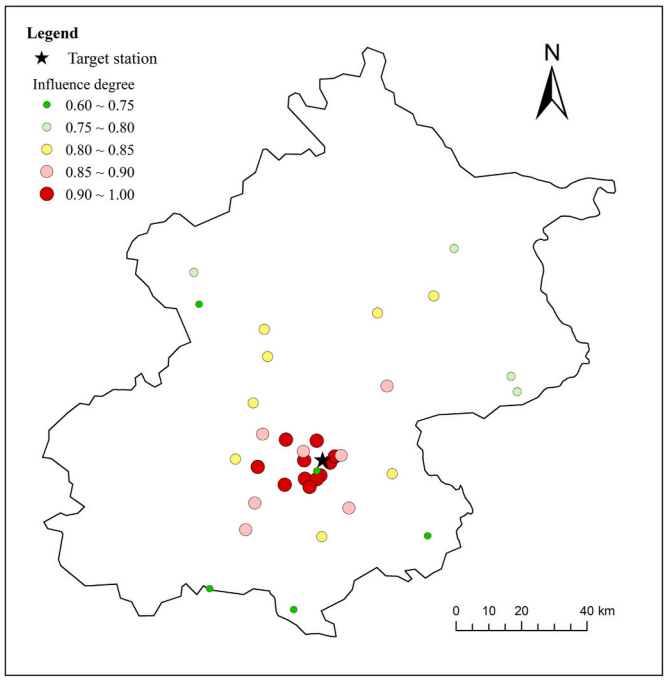

Fig. 3. Map of the influence degree (defined as the Pearson correlation coefficient) of each station toward the target station.

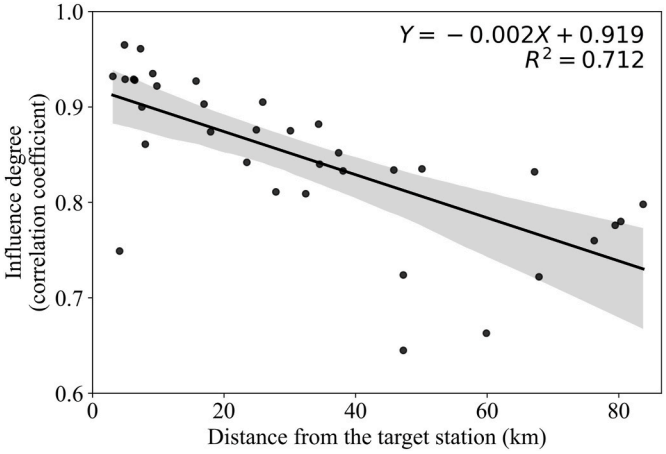

Fig. 4. Relationship between the distance from each station to the target station (km) and the influence degree (correlation coefficient).

Table 2
The performance analysis of different correlation thresholds (ρ_{th}) on ST-CausalConvNet.

ρ_{th}	RMSE	MAE	R^2
0.75	18.390	12.473	0.922
0.80	17.914	11.893	0.931
0.85	17.436	11.746	0.936
0.90	18.291	12.135	0.929
0.95	18.975	12.731	0.924

3.2. $PM_{2.5}$ concentration prediction result

The $PM_{2.5}$ concentration values were predicted by the proposed ST-CausalConvNet on the test data set. Fig. 5 shows the predicted values with the corresponding observed values. Except for some relatively large errors due to the absence of data over a period of time, the trends of the

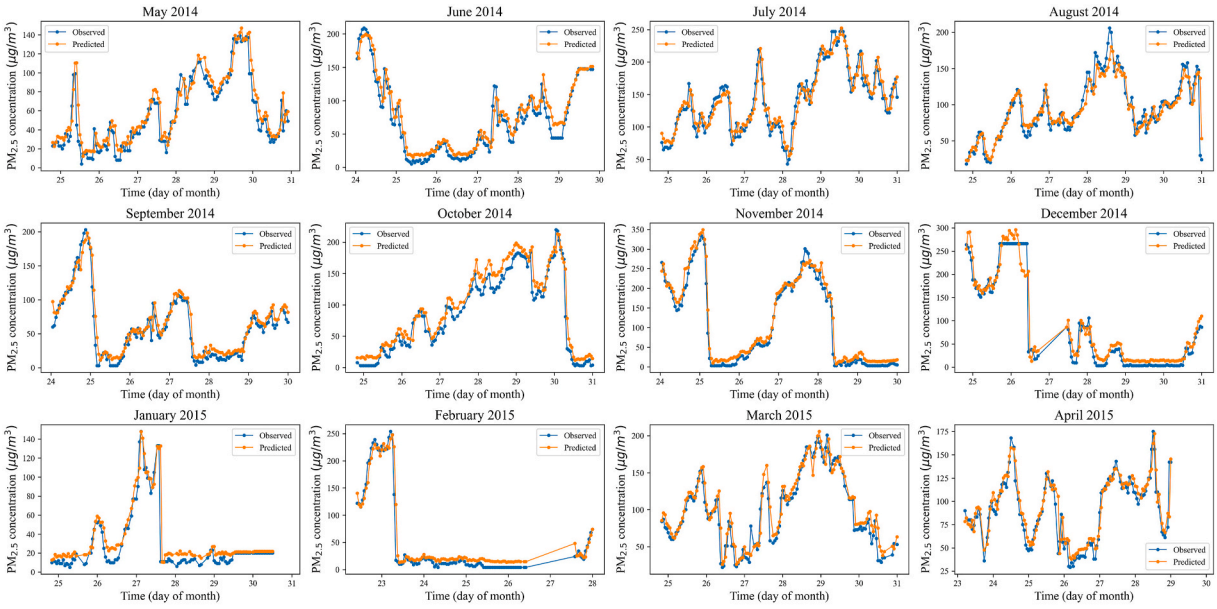

Fig. 5. Prediction results compared with observations.

predicted $PM_{2.5}$ concentrations and the observed values are consistent, verifying the feasibility of the proposed model in capturing temporal variations. The predicted values of the proposed model were sound, and the error was controlled within the tolerated range. Furthermore, the model still performed well in the period of great air pollutant change (e.g. from November 2015 to January 2016 and in February 2015). The proposed model showed a tendency of a small degree of overestimating the low values (less than 100 $\mu g/m^3$) (e.g. some periods in the last three months of 2014 and January, February 2015) and underestimating at the high values (e.g. in August 2014). The gaps between observations and predictions were slightly larger in these periods than the overall prediction result.

4. Discussion

4.1. Comparison with other methods

The result of the comparison of the different models is shown in Table 3. The proposed ST-CausalConvNet outperformed all other three models. The RMSE, MAE and R^2 of the proposed model were 17.436, 11.746 and 0.936, respectively. The proposed ST-CausalConvNet exhibited decreases of 52.7%, 38.6% and 10.5% in RMSE and 53.1%, 39.3% and 8.6% in MAE and increases of R^2 of 30.0%, 12.6% and 0.16% compared with the ANN, GRU and LSTM models, respectively. The CausalConvNet without considering the spatial dependent stations also achieved a better performance than the conventional neural networks. However, this model were slightly worse than ST-CausalConvNet with the consideration of the information from other influential stations. This result suggests that the proposed causal network that considers spatial dependence can effectively improve the model performance. Fig. 6 shows the comparison result of the linear correlation between the observed and the predicted values of $PM_{2.5}$ concentrations by using the ST-CausalConvNet and three other models. The result indicates that our model achieved the highest Pearson's correlation coefficient of 0.971, and a strong agreement between the predicted and the observed values.

Table 3
Model performance of the proposed ST-CausalConvNet and comparisons with other models.

	ANN	GRU	LSTM	CausalConvNet	ST-CausalConvNet
RMSE	36.898	28.377	19.473	18.854	**17.436**
MAE	25.045	19.362	12.853	12.109	**11.746**
R^2	0.720	0.831	0.921	0.926	**0.936**

4.2. Parameter influence on the model performance

In the ST-CausalConvNet model, two important parameters, the number of levels and kernel size, strongly affect the predictor performance. The number of levels is the same as the number of layers in the causal convolutional network architecture, which represents the depth of the network. Kernel size is the filter size (k) of the network. These details of two parameters are described in Section 2.3.

Fig. 7 shows the model performance (RMSE) under different combinations of the two parameters, which are expressed in a 2D grid. The two parameters were set from one to eight. The main reflection from the figure is that the kernel size and the levels of the network affect the model accuracy, and a balanced combination of a "moderate" value could achieve a good result. Extremely large or small values will reduce the model accuracy. The optimal combination of parameters is four for both the number of levels and kernel size.

The model performance impacted by these two parameters may be attributed to the following. With regard to the levels of the network, if the depth is extremely small, then the network will be simple and difficult to capture and extract the connotation information of the input variables through a multi-layer processing mechanism. If the network depth is extremely large, then the capacity of the model will become large, increasing the difficulty of finding the optimal solution during model training. In terms of the kernel size, if the filter size is extremely small, then the historical information of the previous period cannot be captured in the processing of the first layer of the network, resulting in an insufficient historical information input, resulting in the difficulty in improving the accuracy. If the kernel is extremely large, then it will introduce substantial historical information in the processing of the first layer, leading to the inclusion of some invalid noises and the decline of the prediction accuracy. Therefore, a moderate filter length should be adopted to control the receptive region of the effective history of a layer. Given these observations, we suggest performing a grid search over a specified range in practice. The result here suggests that the reasonable range of the levels and the filter size in our network for predicting $PM_{2.5}$ could be from three to five.

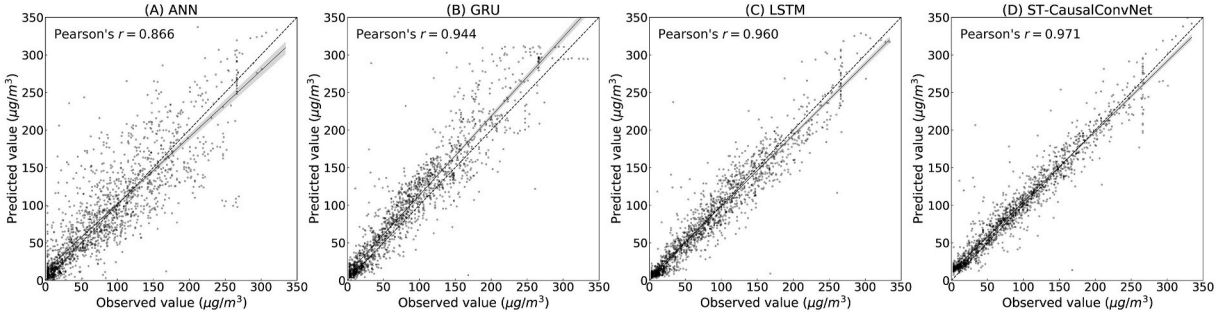

Fig. 6. Correlation analysis between the observed and predicted values of $PM_{2.5}$ concentrations by different models on the test data. The dashed line is the $y = x$ reference line, and the solid line is the regression line.

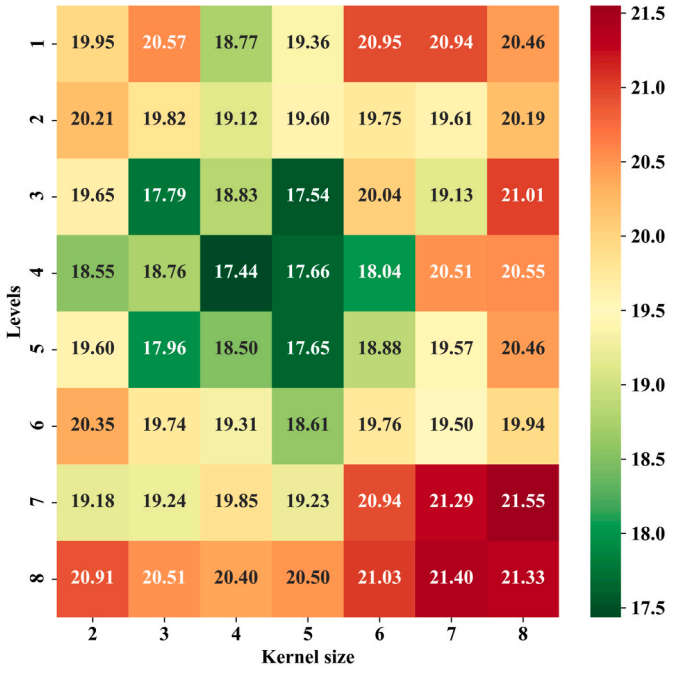

Fig. 7. RMSE with different combinations of two parameters (levels and kernel size).

4.3. Transferability and limitation

We selected other four monitoring stations around the center of Beijing city (in northwest, northeast, southwest and southeast directions) out of this dataset and performed the model to verify the transferability of our method. Fig. 8 shows the maps of the influence degree of the stations on these four target stations. The distance to the neighboring station, the higher correlation to the target station. These four maps reflected a similar pattern to that in Fig. 3. The prediction accuracies of these four target stations are shown in Table 4 which also achieved a good performance. In comparison with the accuracy of station 1013 with an R^2 of 0.936, the model also achieved a good performance at these four stations, although one station (ID: 1023) was slightly low, which may be caused by the less number of influential stations to this target station.

We tested the models in predictions for the next 4, 12 and 24 h to reveal the capability of our model for prediction in different time lengths. Table 5 shows the prediction performance of our model and other three models for these four different time lengths. Although the accuracy decrease whilst the time intervals increased, the proposed model achieved the best performance of $PM_{2.5}$ prediction for all time lengths compared with the other three models. This notion indicates that the proposed model can be effectively applied in a relatively longer prediction task and achieved an acceptable performance in the study area.

Although the proposed model can improve the prediction performance compared with other methods, this model still has some limitations. Firstly, the proposed model was designed to focus on predicting $PM_{2.5}$ at a selected target monitoring station based on an earlier analysis of the spatial dependence among multiple stations. As it is more reliable to feed the original observed data from monitoring stations into our model, the smoothed historical pollutant maps were not be utilised as the input features. However, the interpolated pollutant distribution maps can also be considered for the predictive model as some previous studies did (Delavar et al., 2019; Xiao et al., 2018). Thus, this tool is a potential way to further improve the method. Secondly, the effect of the regulatory policy for air control is an additional important guidance for the prediction (Chen et al., 2019b). Incorporating the policy information into the model will be beneficial, and figuring out the manner by which to integrate regulatory policies into the model would be an interesting research topic. Thirdly, the model was only applied in one city in China due to the limitation of open-access hourly data. In the future, the proposed model may be comprehensively evaluated by applying it in other study areas or other time periods once the more dataset could be available.

5. Conclusion

This study developed an ST-CausalConvNet for predicting air pollution. The proposed model is designed with a causal structure that prevents information leakage from future to past compared with the other deep learning models with RNN structures. Furthermore, the model structure is simpler and clearer than RNN structures. The ST-CausalConvNet not only utilises the historical feature information from the target station but also considers the information from the other stations that have a high relationship with the target station by using correlation analysis. Then, the total input vectors are fed into the causal convolutional neural network, and backpropagation is adopted to train the model. The results of our case study showed that the proposed method achieved a good performance in predicting the hourly $PM_{2.5}$ concentrations in Beijing. The comparison with other models indicated that the proposed model outperformed ANN, GRU and LSTM with 52.7%, 38.6% and 10.5% decreases in RMSE, respectively. Given that the correlation threshold values can control the amount and quality of input data, a moderate threshold value was adopted because extremely permissive or strict threshold values may lead to poor prediction performance. The results showed that optimal values of the other two important parameters, namely, the number of levels and the kernel size in the causal convolutional network, in the case study were both four. In practice, we suggest that parameter optimisation can be conducted by performing a grid search of the number of levels and kernel size and selecting the optimal parameters determined by the prediction accuracy

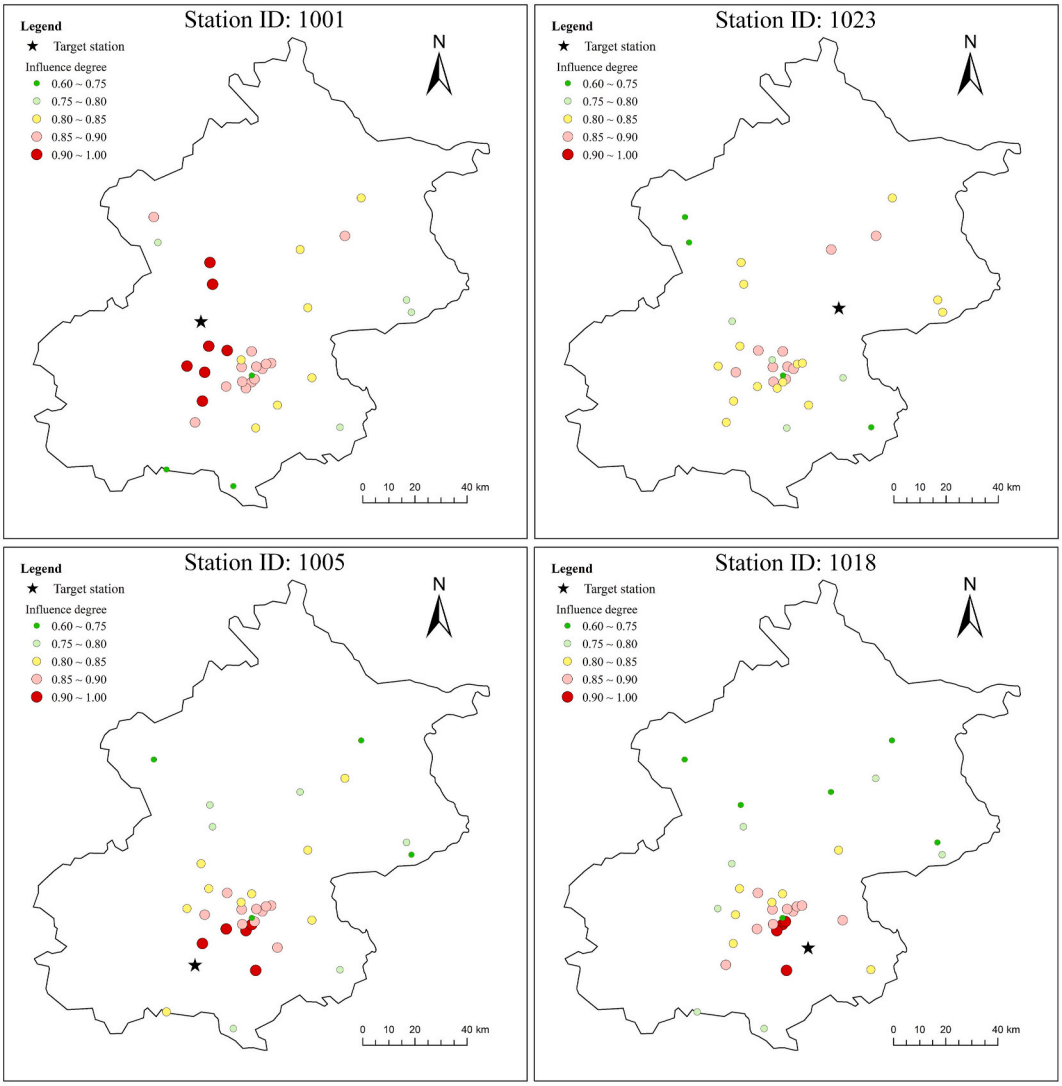

Fig. 8. Spatiotemporal relationship results of the other four monitoring stations in Beijing. The station IDs and their corresponding areas are: 1001, Haidian North New District; 1023, Shunyi New City; 1005, Fangshan Liang Xiang; and 1018, Yizhuang Development Zone.

Table 4
Prediction accuracy of these four stations.

Station ID	RMSE	MAE	R^2
1001	18.036	10.865	0.944
1023	23.741	14.037	0.931
1005	18.055	10.977	0.938
1018	22.742	13.979	0.939

when repeating the training of the model in the study area. Finally, we conclude that the proposed spatiotemporal causal convolutional neural network is a potentially effective and accurate method for air pollution forecasting.

Computer code availability

This study is executed via Python script at https://github.com/zlxy9892/ST-CausalConvNet based on PyTorch package. The dataset is http://research.microsoft.com/en-us/projects/urbanair (Urban Computing Team, Microsoft Research) and a copy is also available by the above Github link.

Table 5
Prediction accuracies of four models for different time lengths.

Model	Metric	+1 h	+4 h	+12 h	+24 h
ANN	RMSE	36.898	44.832	54.785	65.423
	MAE	25.045	29.177	37.838	45.967
	R^2	0.720	0.661	0.590	0.524
GRU	RMSE	28.377	34.867	42.496	46.819
	MAE	19.362	22.745	28.729	33.480
	R^2	0.831	0.725	0.690	0.653
LSTM	RMSE	19.473	27.471	34.130	39.516
	MAE	12.853	18.599	22.791	24.530
	R^2	0.921	0.855	0.730	0.711
ST-CausalConvNet	RMSE	**17.436**	**24.712**	**31.535**	**36.230**
	MAE	**11.746**	**14.686**	**19.922**	**22.929**
	R^2	**0.936**	**0.871**	**0.783**	**0.724**

CRediT authorship contribution statement

Lei Zhang: Conceptualization, Methodology, Data curation, Visualization, Validation, Writing – original draft. **Jiaming Na:** Conceptualization, Visualization, Validation, Supervision, Writing – original draft, Writing – review & editing. **Jie Zhu:** Visualization, Resources, Funding acquisition. **Zhikuan Shi:** Writing – review & editing. **Changxin Zou:**

Resources, Writing – review & editing. **Lin Yang:** Writing – review & editing.

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.

Acknowledgement

This work was financially supported by the Natural Science Foundation of China (No. 41971054), the talent research start-up funding project of Nanjing Forestry University (No. GXL2018049), the foundation of Key Lab of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education (No. 2020VGE04), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (No. 164320H116). We would like to express our sincere gratitude to the Editor and the two anonymous reviewers, whose valuable comments and suggestions have greatly improved the quality of the paper.

References

Ahn, J., Shin, D., Kim, K., Yang, J., 2017. Indoor air quality analysis using deep learning with sensor data. Sensors 17 (11), 2476.
Ai, S., Wang, C., Qian, Z.M., Cui, Y., Liu, Y., Acharya, B.K., et al., 2019. Hourly associations between ambient air pollution and emergency ambulance calls in one central Chinese city: implications for hourly air quality standards. Sci. Total Environ. 696, 133956.
Bai, S., Kolter, J.Z., Koltun, V., 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling arXiv preprint arXiv:1803.01271.
Bhaskaran, K., Hajat, S., Armstrong, B., Haines, A., Herrett, E., Wilkinson, P., Smeeth, L., 2011. The effects of hourly differences in air pollution on the risk of myocardial infarction: case crossover analysis of the MINAP database. BMJ 343.
Bianchini, M., Scarselli, F., 2014. On the complexity of neural network classifiers: a comparison between shallow and deep architectures. IEEE transactions on neural networks and learning systems 25 (8), 1553–1565.
Byun, D., 1999. Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. EPA/600/R-99/030.
Cabaneros, S.M., Calautit, J.K., Hughes, B.R., 2019. A review of artificial neural network models for ambient air pollution prediction. Environ. Model. Software 119, 285–304. https://doi.org/10.1016/j.envsoft.2019.06.014.
Chaloulakou, A., Grivas, G., Spyrellis, N., 2003. Neural network and multiple regression models for PM10 prediction in Athens: a comparative assessment. J. Air Waste Manag. Assoc. 53 (10), 1183–1190.
Chen, Z., Chen, D., Wen, W., Zhuang, Y., Kwan, M.P., Chen, B., et al., 2019a. Evaluating the "2+ 26" regional strategy for air quality improvement during two air pollution alerts in Beijing: variations in PM 2.5 concentrations, source apportionment, and the relative contribution of local emission and regional transport. Atmos. Chem. Phys. 19 (10), 6879–6891.
Chen, Z., Chen, D., Xie, X., Cai, J., Zhuang, Y., Cheng, N., Gao, B., 2019b. Spatial self-aggregation effects and national division of city-level PM2. 5 concentrations in China based on spatio-temporal clustering. J. Clean. Prod. 207, 875–881.
Chen, Z., Chen, D., Zhao, C., Kwan, M.P., Cai, J., Zhuang, Y., et al., 2020. Influence of meteorological conditions on PM2. 5 concentrations across China: a review of methodology and mechanism. Environ. Int. 139, 105558.
Chen, Z., Zhuang, Y., Xie, X., Chen, D., Cheng, N., Yang, L., Li, R., 2019c. Understanding long-term variations of meteorological influences on ground ozone concentrations in Beijing during 2006–2016. Environ. Pollut. 245, 29–37.
Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734. Doha, Qatar, 25–29 October 2014.
Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., Liu, F., Tian, L., Zhu, Z., Xiang, H., 2016. A review on predicting ground PM2.5 concentration using satellite aerosol optical depth. Atmosphere 7, 129.
Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017, August. Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning, 70. JMLR, pp. 933–941.
Delavar, M.R., Gholami, A., Shiran, G.R., Rashidi, Y., Nakhaeizadeh, G.R., Fedra, K., Hatefi Afshar, S., 2019. A novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of Tehran. ISPRS Int. J. Geo-Inf. 8 (2), 99.
Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., et al., 2019. An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. Environ. Int. 130, 104909.
Donnelly, A., Misstear, B., Broderick, B., 2015. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. Atmos. Environ. 103, 53–65. https://doi.org/10.1016/j.atmosenv.2014.12.011.
Gehring, J., Auli, M., Grangier, D., Dauphin, Y.N., 2016. A Convolutional Encoder Model for Neural Machine Translation arXiv preprint arXiv:1611.02344.
Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N., 2017, July. Convolutional sequence to sequence learning. In: International Conference on Machine Learning. PMLR, pp. 1243–1252.
de Gennaro, G., Trizio, L., Di Gilio, A., Pey, J., Pérez, N., Cusack, M., et al., 2013. Neural network model for the prediction of PM10 daily concentrations in two sites in the Western Mediterranean. Sci. Total Environ. 463, 875–883.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT press.
Grell, G.A., Peckham, S.E., Schmitz, R., McKeen, S.A., Frost, G., Skamarock, W.C., Eder, B., 2005. Fully coupled "online" chemistry within the WRF model. Atmos. Environ. 39 (37), 6957–6975.
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, pp. 770–778.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating PM2. 5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51 (12), 6936–6944.
Huang, C.J., Kuo, P.H., 2018. A deep cnn-lstm model for particulate matter (PM2. 5) forecasting in smart cities. Sensors 18 (7), 2220.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Statistical learning. In: James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics. Springer, New York, NY, pp. 15–57. https://doi.org/10.1007/978-1-4614-7138-7_2.
Johnson, R., Zhang, T., 2015. Effective use of word order for text categorization with convolutional neural networks. In: HLT- NAACL.
Johnson, R., Zhang, T., 2017. Deep pyramid convolutional neural networks for text categorization. ACL.
Kalchbrenner, N., Grefenstette, E., Blunsom, P., 2014. A Convolutional Neural Network for Modelling Sentences arXiv preprint arXiv:1404.2188.
Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A.V.D., Graves, A., Kavukcuoglu, K., 2016. Neural Machine Translation in Linear Time arXiv preprint arXiv:1610.10099.
Kim, Y., Fu, J.S., Miller, T.L., 2010. Improving ozone modeling in complex terrain at a fine grid resolution: Part I–examination of analysis nudging and all PBL schemes associated with LSMs in meteorological mod-el. Atmos. Environ. 44 (4), 523–532.
Kim, Yoon, 2014. Convolutional neural networks for sentence classification. In: EMNLP.
Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization arXiv preprint arXiv:1412.6980.
Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D., 2017. Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 156–165.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Li, X., Peng, L., Hu, Y., Shao, J., Chi, T., 2016. Deep learning architecture for air quality predic-tions. Environ. Sci. Pollut. Control Ser. 23 (22), 22408–22417.
Lin, M., Chen, Q., Yan, S., 2013. Network in Network arXiv preprint arXiv:1312.4400.
Liu, H., Tian, Y., Xiang, X., Li, M., Wu, Y., Cao, Y., et al., 2018. Association of short-term exposure to ambient carbon monoxide with hospital admissions in China. Sci. Rep. 8 (1), 1–7.
Liu, W., Guo, G., Chen, F., Chen, Y., 2019. Meteorological pattern analysis assisted daily PM2. 5 grades prediction using SVM optimized by PSO algorithm. Atmospheric Pollution Research 10 (5), 1482–1491.
Loomis, D., Grosse, Y., Lauby-Secretan, B., Ghissassi, F.E., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., Straif, K., 2013. The carcinogenicity of outdoor air pollution. Lancet Oncol. 14, 1262–1263.
Ma, Y., Zhang, H., Zhao, Y., Zhou, J., Yang, S., Zheng, X., Wang, S., 2017. Short-term effects of air pollution on daily hospital admissions for cardiovascular diseases in western China. Environ. Sci. Pollut. Control Ser. 24 (16), 14071–14079.
Ong, B.T., Sugiura, K., Zettsu, K., 2014. Dynamic pre-training of deep recurrent neural networks for predicting environmental monitoring data. In: 2014 IEEE International Conference on Big Data (Big Data). IEEE, pp. 760–765.
Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. WaveNet: A Generative Model for Raw Audio arXiv:1609.03499 [cs].
Pak, U., Kim, C., Ryu, U., Sok, K., Pak, S., 2018. A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction. Air Quality, Atmosphere & Health 11 (8), 883–895.
Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., Pak, C., 2020. Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: a case study of Beijing, China. Sci. Total Environ. 699, 133561.
Perez, P., Reyes, J., 2006. An integrated neural network model for PM10 forecasting. Atmos. Environ. 40 (16), 2845–2851. https://doi.org/10.1016/j.atmosenv.2006.01.010.
Pérez, P., Trier, A., Reyes, J., 2000. Prediction of PM2. 5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. 34 (8), 1189–1196.
Qiao, W., Tian, W., Tian, Y., Yang, Q., Wang, Y., Zhang, J., 2019. The forecasting of PM2. 5 using a hybrid model based on wavelet transform and an improved deep learning algorithm. IEEE Access 7, 142814–142825.
Qi, Y., 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.

Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Network. 61, 85–117.
Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv preprint arXiv:1409.1556.
Siwek, K., Osowski, S., 2016. Data mining methods for prediction of air pollution. Int. J. Appl. Math. Comput. Sci. 26, 467–478. https://doi.org/10.1515/amcs-2016-0033.
Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., De Hoogh, K., De' Donato, F., et al., 2019. Estimation of daily PM10 and PM2. 5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ. Int. 124, 170–179.
Sun, W., Liu, M., 2016. Prediction and analysis of the three major industries and residential consumption CO2 emissions based on least squares support vector machine in China. J. Clean. Prod. 122, 144–153.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.
Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., Cribb, M., 2019. Estimating 1-km-resolution PM2. 5 concentrations across China using the space-time random forest approach. Rem. Sens. Environ. 231, 111221.
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.
WHO, 2003. Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide. Tech. Rep. WHO. https://apps.who.int/iris/handle/10665/107478.
WHO, 2016. Ambient air pollution: a global assessment of exposure and burden of disease. In: WHO Library Cataloguing-In-Publication Data.
Xiao, L., Lang, Y., Christakos, G., 2018. High-resolution spatiotemporal mapping of PM2. 5 concentrations at Mainland China using a combined BME-GWR technique. Atmos. Environ. 173, 295–305.
Xing, Y., Xu, Y., Shi, M., Lian, Y., 2016. The impact of PM2. 5 on the human respiratory system. J. Thorac. Dis. 8 (1), E69.
Yan, D., Lei, Y., Shi, Y., Zhu, Q., Li, L., Zhang, Z., 2018. Evolution of the spatiotemporal pattern of PM2.5 concentrations in China–A case study from the Beijing-Tianjin-Hebei region. Atmos. Environ. 183, 225–233.
Yildirim, Y., Bayramoglu, M., 2006. Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak. Chemosphere 63 (9), 1575–1582.
Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. In: ICLR.
Zhang, Y., Lang, J., Cheng, S., Li, S., Zhou, Y., Chen, D., et al., 2018. Chemical composition and sources of PM1 and PM2.5 in Beijing in autumn. Sci. Total Environ. 630, 72–82.
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, pp. 2267–2276.
Zhou, Q., Jiang, H., Wang, J., Zhou, J., 2014. A hybrid model for PM2. 5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Sci. Total Environ. 496, 264–274.
Zíková, N., Wang, Y., Yang, F., Li, X., Tian, M., Hopke, P.K., 2016. On the source contribution to Beijing PM2. 5 concentrations. Atmos. Environ. 134, 84–95.