Science of the Total Environment 664 (2019) 1–10

Contents lists available at ScienceDirect

Science of the Total Environment

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

A hybrid model for spatiotemporal forecasting of $PM_{2.5}$ based on graph convolutional neural network and long short-term memory

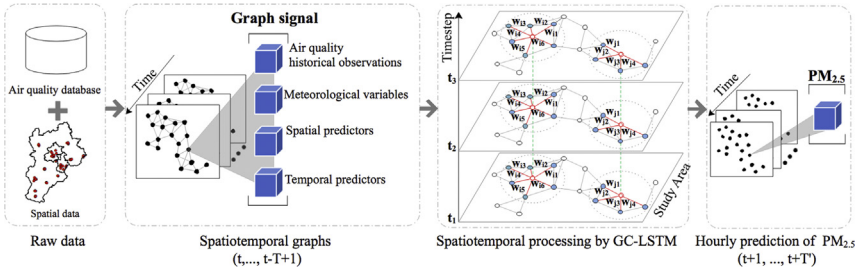
Yanlin Qi[a], Qi Li[a,*], Hamed Karimian[a], Di Liu[b]

[a] Institute of Remote Sensing and Geographic Information System, Peking University, Beijing, China
[b] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei, China

HIGHLIGHTS

- A hybrid model is proposed for hourly $PM_{2.5}$ forecasting.
- Considering spatiotemporal dependency can improve model performance.
- Our model shows better performance especially for long-term prediction.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 20 November 2018
Received in revised form 14 January 2019
Accepted 25 January 2019
Available online 1 February 2019

Editor: Pavlos Kassomenos

Keywords:
Air pollution forecasting
Spatiotemporal data modelling
Graph convolutional neural network
Long short-term memory
Deep learning

ABSTRACT

Increasing availability of data related to air quality from ground monitoring stations has provided the chance for data mining researchers to propose sophisticated models for predicting the concentrations of different air pollutants. In this paper, we proposed a hybrid model based on deep learning methods that integrates Graph Convolutional networks and Long Short-Term Memory networks (GC-LSTM) to model and forecast the spatiotemporal variation of $PM_{2.5}$ concentrations. Specifically, historical observations on different stations are constructed as spatiotemporal graph series, and historical air quality variables, meteorological factors, spatial terms and temporal attributes are defined as graph signals. To evaluate the performance of the GC-LSTM, we compared our results with several state-of-the-art methods in different time intervals. Based on the results, our GC-LSTM model achieved the best performance for predictions. Moreover, evaluations of recall rate (68.45%), false alarm rate (4.65%) (both of threshold: 115 μg/m^3) and correlation coefficient R^2 (0.72) for 72-hour predictions also verify the feasibility of our proposed model. This methodology can be used for concentration forecasting of different air pollutants in future.

1. Introduction

High concentration of $PM_{2.5}$ (Particulate matter with diameter less than 2.5 μm) in the major cities of North China Plain has gained remarkable public concern specially due to its adverse environmental issues and health effects. Thus, accurate forecasting of $PM_{2.5}$ concentration is demanded to aid decision makers and subpopulations in these regions to decrease the effect of exposure to it.

Previous studies in predictions of air pollutant concentrations can be roughly classified into two major categories as simulation-based methods and observation-based methods. Simulation-based studies such as CMAQ (Byun, 1999), GEOS-Chem (Bey et al., 2001) and WRF/Chem (Grell et al., 2005), usually apply the current understandings of atmospheric physics and chemical processes to predict mass concentrations of different pollutants

* Corresponding author.
E-mail address: liqi@pku.edu.cn (Q. Li).

https://doi.org/10.1016/j.scitotenv.2019.01.333
0048-9697/

(Geng et al., 2015; Li et al., 2013). However, factors such as lack of accurate input data (e.g. emission sources and amount) may cause discrepancies in the result of these models for China.

With increasing availability of historical data that are collected by different sensors, observation-based technique is another widely-used method in which statistical models connect several explanatory variables to predict $PM_{2.5}$ concentration as output (Reyes and Serre, 2014; Sun and Sun, 2016; Zhang et al., 2016). Aerosol optical depth (AOD) is one of the most typical satellite-based data products to predict the surface distribution of particle pollutants. However, remarkable missing values (low spatial coverage) and also low temporal resolution make the application of AOD for hourly forecasting limited (e.g., MODIS) (Remer et al., 2005; Wang et al., 2017b; Zhang et al., 2018). Although some geostationary satellites have been introduced to overcome this problem, they were not used for global coverage and still monitored the aerosols over limited geographical area (e.g., Himawari-8) (Wang et al., 2017a; Zang et al., 2018). Therefore, ground-based data are commonly used as explanatory variables due to their high accuracy and temporal resolution. Since installing more densely distributed air quality monitoring stations has been an irresistible trend under the public concern, spatiotemporal modelling of ground stations will play a non-substitutable role with more flexibility and accuracy for urban area.

Considering linear relationship between $PM_{2.5}$ and explanatory variables, methods like geographically weighted regression (Ma et al., 2014), geographically and temporally weighted regression (GTWR) (He and Huang, 2018) and land-use regression (LUR) (Shamsoddini et al., 2017) were proposed. However, the linearity assumption may lead to misperforming of the models. Support vector machine (SVM) (Sun and Sun, 2016), random forest (Shamsoddini et al., 2017), artificial neural networks (ANN) (Feng et al., 2015; Mao et al., 2017; Mckendry, 2002) are some of the widely-used methods in air pollution prediction which took nonlinearity into account and better result than conventional linear models were reported (Chaloulakou et al., 2003; Pérez et al., 2000). Moreover, with the rapid growth of data volume, the exploitation of deep learning methods in recent work further shows the full potential of artificial neural networks (Shi et al., 2015; Ong et al., 2016).

As for the complex temporal dependency in air pollution data series, deep learning models like recurrent neural networks (RNN) and its variant long short-term memory (LSTM) have been developed by previous work (Fan et al., 2017; Li et al., 2017) for the air pollutant forecast over different time spans, but without considering spatial dependency between monitoring stations.

In geo-statistical problems, the non-linear spatial dependency is one of the important factors and considering different effects locally may improve the model performance. In recent years, neural networks have shown their ability to discover non-linearity and can commonly overcome the problem of big dataset, hence embedding spatial feature processing into these neural-network-based methods is another sensible approach. With the non-Euclidean distribution of geo-objects like ground monitoring stations, this embedding work has been a knotty problem in previous work. Some compromising approaches tried to apply the convolutional neural network (CNN) in Euclidean space to excavate spatial dependency in air quality data on monitoring sites by manually rearranging stations to 2D arrays (Wu et al., 2018). However in this approach, the original spatial information is destructed and it ignores the temporal dependency. The work by Huang and Kuo (2018) stepped further by applying ConvLSTM model which integrates CNN and LSTM structure to simulate spatiotemporal relationships of station-observing data. But this method is also CNN-based and most appropriate for spatial relationships in the Euclidean space.

Since these CNN-based methods are tailored to data in form of 2-D or 3-D matrices like images and videos, generalizing the convolution operator to arbitrary graph structure has become a recent area of interest. Bruna et al. (2014) proposed the first graph convolution kernel for the spectral graph convolutional neural networks, but somehow limited by computational defects with big graphs. It was soon improved by the following work (Defferrard et al., 2016; Henaff et al., 2015; Kipf and Welling, 2017), which significantly decreased the computing complexity and made it a competitive method for non-linear spatial dependency modelling.

In this paper, our objective is to address the aforementioned limitations and propose a hybrid model to improve the forecast of $PM_{2.5}$ mass concentration. We applied graph convolutional networks (GCN) to extract the spatial dependency between different stations and LSTM to capture the temporal dependency among observations at different times.

2. Data and methods

2.1. Study area and available data

Our study area is Jing-Jin-Ji (Beijing, Tianjin and Hebei), the major part of North China Plain (113°E–120°E and 36°N–43°N, Fig. 1). This area has experienced rapid growth in urbanization, industrial production and energy consumption over the past few decades, and suffers from frequent and sever air pollution scenarios annually (Karimian et al., 2018).

Open air quality observation data from ground monitoring stations in China have been published by the Environmental Protection Administration (EPA) of China since early 2013. We collected an hourly scaled dataset of pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, CO) from 76 stations over Jing-Jin-Ji area for the time period of January 1, 2015 to April 1, 2016 (totally 10,944 h). We used the latest valid observation to fix missing data for each kind of the observed air pollutants. Besides, from the China Weather Website Platform which is maturely maintained by China Meteorological Administration (CMA), we downloaded hourly meteorological observation data including wind speed, wind direction, temperature, pressure and relative humidity for the same period. The meteorological data were matched to air quality monitoring stations by searching the nearest stations of them. In addition to concentrations of air pollutants and meteorological properties, we added spatial information including the longitude and latitude of each station, as well as temporal information such as year, month, day and day of the week properties to each piece of data. Table 1 summarizes the statistics of related variables.

2.2. Spatial dependency modelling

2.2.1. Graph construction

Constructing a graph structure is usually preferred to represent the spatial relationships of geospatial data. Whereas prior knowledge of graph structure may be naturally existed in some situations like transportation network, such knowledge is not directly available in our problem. According to the definition of graph structure, the input features of N monitoring stations at each time t can be translated to graph signals as a feature matrix $\boldsymbol{X}_t \in \mathbb{R}^{N\times M}$ where M is the number of features associated with each node. Spatial relationships existed among different locations can be represented as an undirected graph $G = (V, E, \boldsymbol{A})$ with N nodes $v_i \in V$ (including all the stations in study area) and each edge $(v_i, v_j) \in E$. $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is the spatial weight matrix, where each element $A_{i,j}$ represents the quantitative spatial correlation between v_i and v_j ($A_{i,j} = 0$).

We established the spatial weight matrix $\boldsymbol{A}$ based on the spatial distances between air quality stations in Jing-Jin-Ji area. With the geographic coordinate set $C = \{(x_1, y_1), \cdots (x_N, y_N)\}$

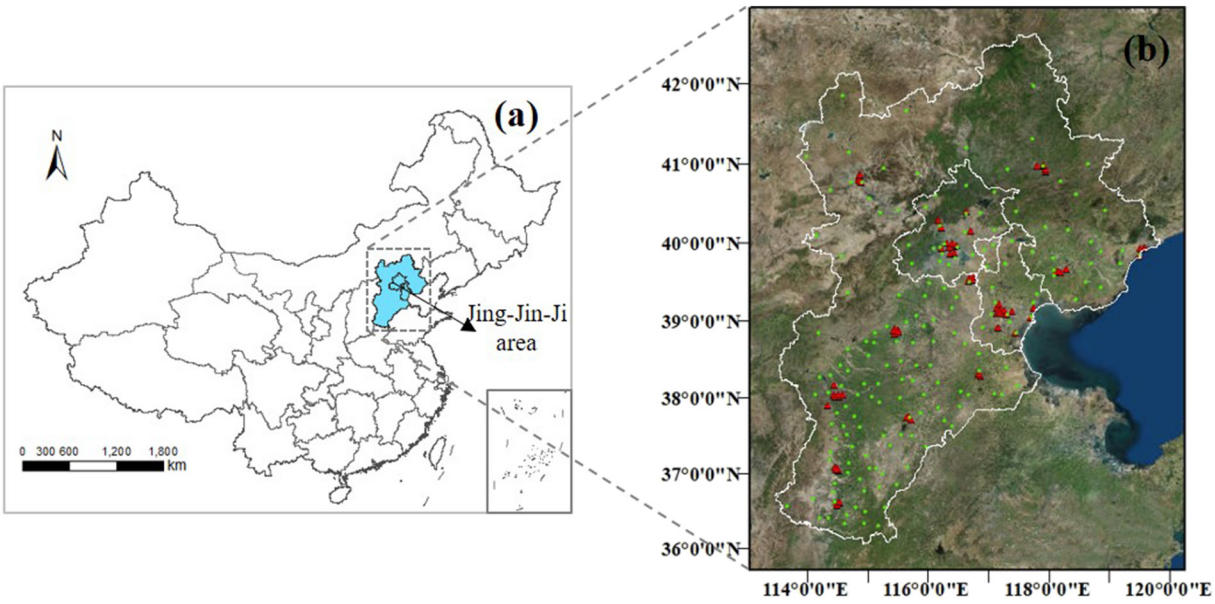


Fig. 1. (a) The location of Beijing, Tianjin, and Hebei province (Jing-Jin-Ji area). (b) The spatial distribution of air pollution monitoring stations (marked by red triangles) and meteorological stations (green dots).

where x_i and y_i are the latitude and longitude of vertex i, the spatial distance between site i and site j is given by

$$d_{i,j} = d_{geo}\left((x_i, y_i), (x_j, y_j)\right). \tag{1}$$

Then the distance $d_{i,j}$ is applied to calculate the element of matrix $\boldsymbol{A}$

$$A_{i,j} = \begin{cases} \frac{1}{d_{i,j}}, & i \neq j \wedge d_{i,j} < R \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $R = 200$ km is the threshold value we adopted. Only stations located within distance of R is considered to be connected with the target. In addition to inverse distance weighted method, more complex functions like Gaussian kernel or other definition approaches are also applicable.

The graph structure generated by our definitions is shown in Fig. 2. More details of nodes and edges by administrative division in this graph are summarized in Table 2. In general, sites with similar locations correspond to the same level of edge numbers. Through modelling the spatial correlation of stations, feature extraction on each site is then based on an integration of its neighborhood rather than the target own.

Table 1
Statistics of collected features. Unit, range, mean, and standard deviation values of related data from 1 January 2015 to 1 April 2016.

Variable	Unit	Range	Mean	St. dev.
$PM_{2.5}$	$\mu g/m^3$	[1,500]	74.01	74.51
PM_{10}	$\mu g/m^3$	[1,2714]	127.96	106.36
SO_2	ppb	[1,914]	35.38	41.86
NO_2	ppb	[1,467]	46.33	32.97
CO	ppb	[0.00,90]	1.42	1.47
O_3	ppb	[1,869]	52.95	48.58
Wind_ x	m/s	[−15.06,12.91]	0.63	2.78
Wind_ y	m/s	[−21.89,14.96]	−0.41	3.55
Pressure	hPa	[877.28,1044.79]	996.04	34.47
Relative humidity	%	[0.07,1.00]	0.47	0.21
Temperature	°C	[−12.31,41.26]	10.81	12.09
Latitude	(°)	[36.58,41.01]	39.14	1.17
Longitude	(°)	[114.35,119.61]	116.39	1.44
Year	NA	[2015,2016]	–	–
Month	NA	[1,12]	–	–
Day	NA	[1,31]	–	–
Day of the week	NA	[1,7]	–	–

2.2.2. Spectral graph convolutional neural networks

A key point of graph convolutional neural networks is to define spectral graph convolution based on graph Fourier transform and Laplacian matrix $\boldsymbol{L}$. Let $\boldsymbol{x} \in \mathbb{R}^N$ be a signal on the nodes of graph G, where x_i denotes a scalar signal for the i-th node. Spectral graph convolution is then considered as the multiplication of a filter g_θ with signal $\boldsymbol{x}$ in the Fourier domain. The formulation of convolution operator * for signal $\boldsymbol{x}$ can be defined as

$$\boldsymbol{y} = g_\theta * \boldsymbol{x} = g_\theta(\boldsymbol{L})\boldsymbol{x} = g_\theta(\boldsymbol{U}\Lambda\boldsymbol{U}^T)\boldsymbol{x} = \boldsymbol{U}g_\theta(\boldsymbol{\Lambda})\boldsymbol{U}^T\boldsymbol{x}, \tag{3}$$

where $\boldsymbol{\theta} \in \mathbb{R}^N$ is a vector of Fourier coefficients to be learned, and $\boldsymbol{L} = \boldsymbol{I_N} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{U}\Lambda\boldsymbol{U}^{\boldsymbol{T}} \in \mathbb{R}^{N\times N}$ is the normalized graph Laplacian. In the formula of $\boldsymbol{L}$, $\boldsymbol{I_N} \in \mathbb{R}^{N\times N}$ is the identity matrix and $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix with $D_{ii} = \sum_j A_{i,j}$ (Fan, 1997). $\boldsymbol{U} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{N\times N}$ are the matrix of eigenvectors and the diagonal matrix of eigenvalues of $\boldsymbol{L}$, respectively. Filter g_θ can be understood as a function of the eigenvalues of $\boldsymbol{L}$, i.e. $g_\theta(\boldsymbol{\Lambda})$. $\boldsymbol{U}^T\boldsymbol{x}$ denotes the graph Fourier transform of $\boldsymbol{x}$.

In practice, the truncated expansion of $g_\theta(\boldsymbol{\Lambda})$ by Chebyshev polynomials is often suggested to decrease the computational complexity, and meanwhile we can also generalize the definition in Eq. (3) to signal matrix $\boldsymbol{X_t} \in \mathbb{R}^{N\times M}$ and W filters as

$$\boldsymbol{H} = \sum_{k=0}^{K} T_k(\tilde{\boldsymbol{L}})\boldsymbol{X}_t\boldsymbol{\Theta}_k, \tag{4}$$

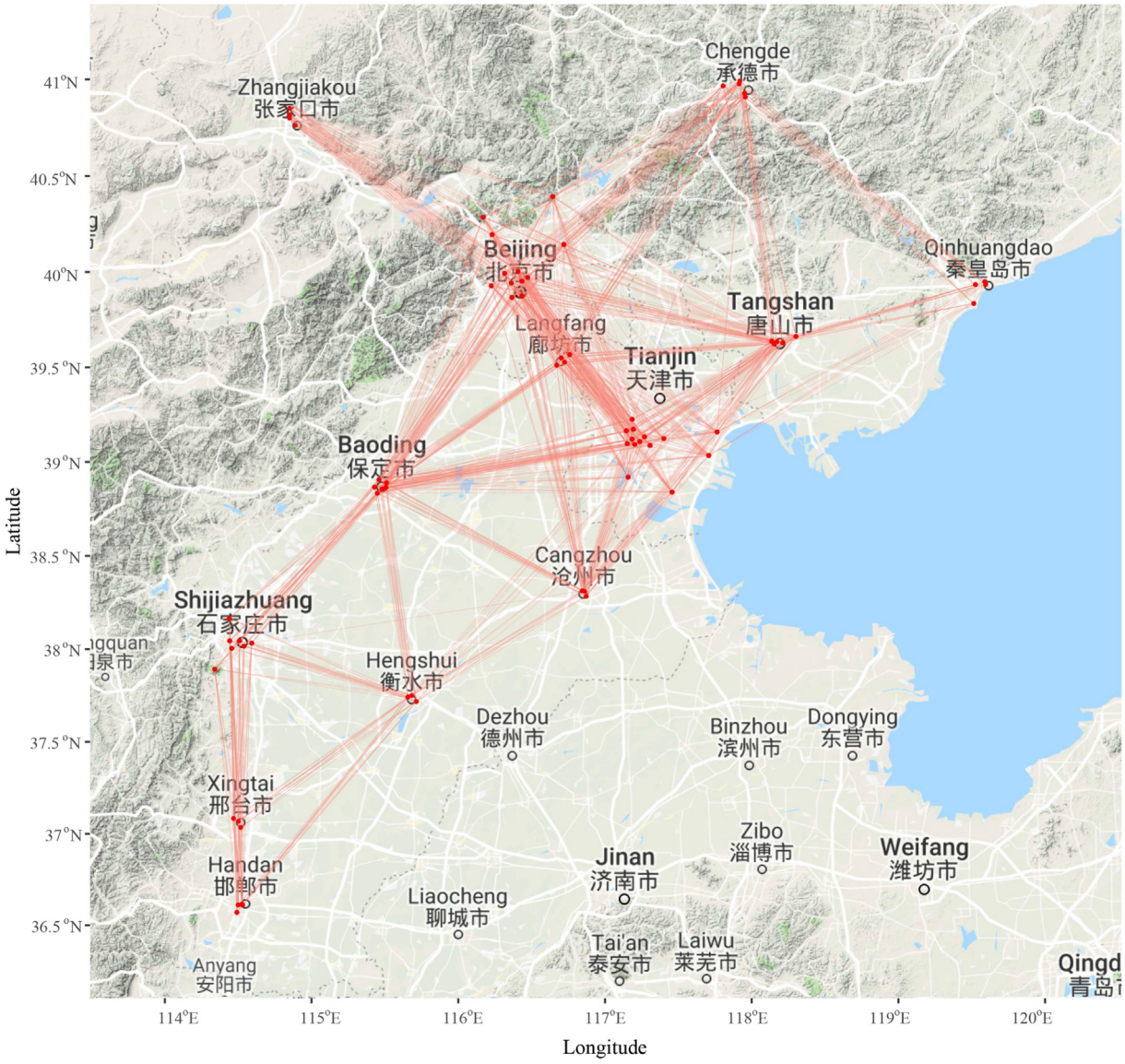


Fig. 2. Graph structure based on air quality monitoring stations (red dots) in Jing-Jin-Ji area. Spatial dependency between two stations is illustrated by the red line.

where $\boldsymbol{\Theta}_k \in \mathbb{R}^{M\times W}$ is now the matrix of filter parameters and $\boldsymbol{H} \in \mathbb{R}^{N\times W}$ is the convolved spatial feature matrix. $T_k(\tilde{\boldsymbol{L}}) \in \mathbb{R}^{N\times N}$ is the kth-order Chebyshev polynomial with the scaled Laplacian $\tilde{\boldsymbol{L}} = \frac{2\boldsymbol{L}}{\lambda_{max}} - \boldsymbol{I}_N$. λ_{max} denotes the largest eigenvalue of $\boldsymbol{L}$. For any input variable $\boldsymbol{s}$, T_k is calculated using the stable recurrence relation $T_k(\boldsymbol{s}) = 2\boldsymbol{s}T_{k-1}(\boldsymbol{s}) - T_{k-2}(\boldsymbol{s})$ with $T_0 = 1$ and $T_1 = \boldsymbol{s}$ according to its definition. The reader is referred to Defferrard et al. (2016) for an in-depth discussion of this approximation.

Note that as the filtering operation is a K-order approximation of the Laplacian, it is K-localized and depends only on nodes that are at maximum K hops away from the central node (the K-neighborhood). Fig. 3 is an illustration of graph convolution operation at 2 hops.

2.3. Temporal dependency modelling

Since the air pollution data from ground monitoring sites are usually in time series, air pollution can be modelled considering time-dependent pattern. Feed-forward neural networks (FNN) are commonly used in previous studies to predict $PM_{2.5}$. However, these models are unable to take the time dependency of parameters into account. Sequence modelling facilitates the excavation of temporal

Table 2
Statistics of graph structure of air quality stations in Jing-Jin-Ji area.

City	Station ID	Count of stations	Average count of edges
Beijing	1001A–1012A	12	51 ± 5
Tianjin	1013A–1027A	14	43 ± 7
Shijiazhuang	1029A–1035A	7	23 ± 1
Tangshan	1036A–040A	5	46 ± 0
Qinhuangdao	1042A, 1044A–1046A	4	15 ± 1
Handan	1047A–1050A	4	17 ± 0
Baoding	1051A–1056A	6	47 ± 1
Zhangjiakou	1057A–1061A	5	16 ± 0
Chengde	1062A–1066A	5	30 ± 1
Langfang	1067A–1070A	4	48 ± 0
Cangzhou	1071A–1073A	3	42 ± 1
Hengshui	1074A–1076A	3	29 ± 1
Xingtai	1077A–1080A	4	17 ± 0

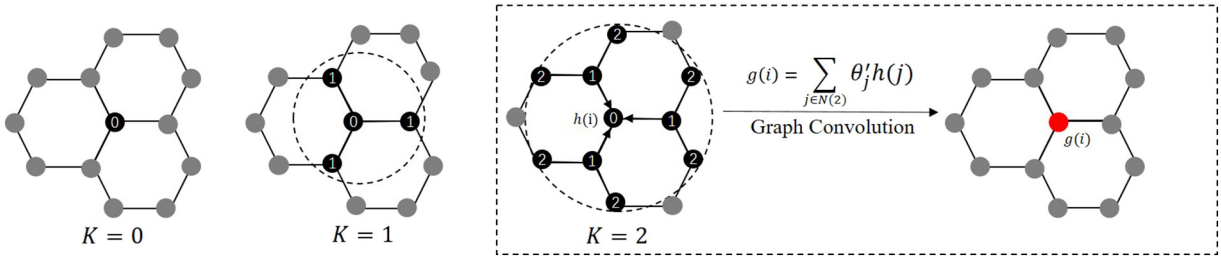


Fig. 3. Illustration of the neighborhood on a 3-nearest-neighborhood graph and graph convolution operation of support $K = 2$.

dynamic features in historical data and helps make better predictions. Compared with FNN, recurrent neural networks (RNN) are designed to deal with time series data but this technique suffers from vanishing or exploding gradient problem. To overcome this, Long Short-term Memory (LSTM) was proposed (Hochreiter and Schmidhuber, 1997). Fig. 4 shows the basic structure of a memory block of LSTM. Each memory block contains one or more memory cells and three nonlinear gates named as forget gate $\boldsymbol{f}_t$, input gate $\boldsymbol{i}_t$ and output gate $\boldsymbol{o}_t$. As for the first two gates, they control the contents of unit state $\boldsymbol{c}_t$. Output gate $\boldsymbol{o}_t$ on the other hand, is used to control the amount of cell state $\boldsymbol{c}_t$ to be mapped as the current output value $\boldsymbol{h}_t$ of LSTM block (Eq. (5)).

$$\begin{aligned}
\boldsymbol{f}_t &= \sigma_g(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \\
\boldsymbol{i}_t &= \sigma_g(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \\
\boldsymbol{o}_t &= \sigma_g(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \sigma_c(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)
\end{aligned} \tag{5}$$

where $\boldsymbol{W}_f$, $\boldsymbol{W}_i$, $\boldsymbol{W}_o$ and $\boldsymbol{W}_c$ are the weight matrices for input vector $\boldsymbol{x}_t$ at time step t and $\boldsymbol{U}_f$, $\boldsymbol{U}_i$, $\boldsymbol{U}_o$ and $\boldsymbol{U}_c$ are the weight matrices assigned to hidden state value from previous block $\boldsymbol{h}_{t-1}$. $\boldsymbol{b}_f$, $\boldsymbol{b}_i$, $\boldsymbol{b}_o$ and $\boldsymbol{b}_c$ are the bias vectors. To bring non-linearity to model, σ_g, σ_c and tanh are used as activation functions and $\odot$ stands for element-wise multiplication of the matrix.

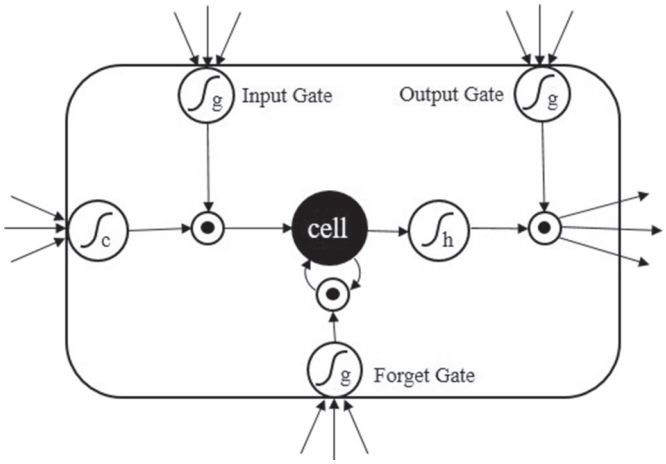


Fig. 4. LSTM memory block with one cell.

2.4. Spatiotemporal prediction architecture

Given historical data and local spatial associations of ground monitoring stations in graph structure, we proposed a hybrid model to predict the future hourly-scaled $PM_{2.5}$ mass concentrations in different time spans:

$$[PM_{2.5_{(t+1)}}, \cdots, PM_{2.5_{(t+T')}}] = h([x_{t-T+1}, \cdots, x_t]; G(V, E, A))). \tag{6}$$

The schematic of our proposed model based on graph convolution and LSTM structure (GC-LSTM) is illustrated in Fig. 5. As can be seen in the first stage, graph convolution operation is utilized to extract the spatial features. This operation is followed by a temporal feature extraction processing by LSTM. The input of LSTM in our framework is the graph convolutional features concatenated with the original signals.

Within each spatiotemporal block, graph signals $\boldsymbol{X}_t$ at each time t together with the spatial weight matrix $\boldsymbol{A}$ are extracted to compute the spatial features $\boldsymbol{H}_t$ by a graph convolution. In the next step, the graph signals $\boldsymbol{X}_t$ are concatenated with $\boldsymbol{H}_t$ to form the input of LSTM ($\boldsymbol{X}_t^{GCN}$). Finally, the output of LSTM is treated as the input of a fully-connected layer (FC) and the output of FC is prediction of $PM_{2.5}$ mass concentration at desirable time.

2.5. Experimental settings

Details of experimental settings are summarized in Table 3. Because the observations of all stations at each time are treated as a whole graph, we made random division on spatiotemporal blocks (each block consists of T graphs in time ordering), rather than original observations. The set of blocks are randomly split into 60% for training set, 20% for validation set and 20% for test set to ensure the reliability of our proposed method. For graph convolution, we considered the 2-hop neighborhood of each station and repeated 200 times of the training process on the training dataset to get the optimal state of our models. The choice of loss functions and optimization algorithms for a deep learning model can play a big role in producing optimum and faster results. As for our prediction model, root mean square error(RMSE), one of the most commonly used loss functions for regression, was selected as the loss function of our training process. Root Mean Square Prop method (RMSprop) is the gradient-based algorithm to optimize the loss function in neural networks, which has shown obvious advantages in decreasing swing range in updating and accelerating the convergence speed in regression-related problems. Therefore, we chose RMSprop as the optimizer of our models.

Moreover, multiple linear regression method (MLR) which is widely used in many forecasting problems, feed-forward neural network (FNN) architecture without considering spatial and temporal dependency and LSTM without considering spatial dependency were used as three baselines to evaluate the capability of our model in

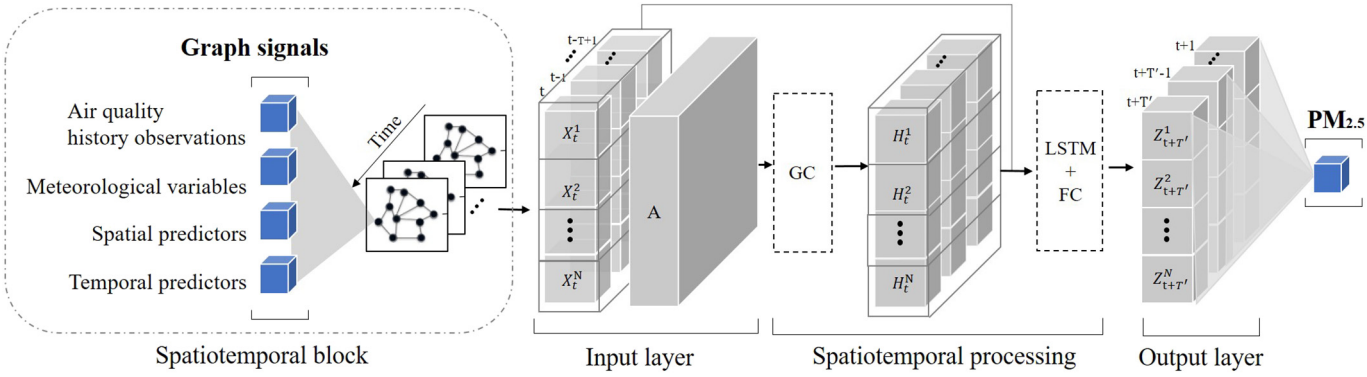


Fig. 5. The architecture of proposed graph convolutional recurrent neural network model in this study.

$PM_{2.5}$ prediction with them. It is worth mentioning that we used the same dataset for the purpose of comparison.

Three metrics including the index of agreement (IA, a nondimensional and bounded measure of the model prediction error degree with values closer to 1 indicating a better match.), the mean absolute error (MAE) and the root mean square error (RMSE) were used to measure the agreement of different model predictions and ground truth values. The definitions of them are by Eqs. (7)–(9):

$$IA = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(|p_i - \bar{o}| + |o_i - \bar{o}|)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|o_i - p_i| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \tag{9}$$

where o_i and p_i are the observed (ground truth) and predicted values respectively. $\bar{o}$ is average value of n observed sample data.

3. Results and discussions

The performances of our approach and those used as baselines are summarized in Table 4. The best results are marked with bold. As can be seen that the GC-LSTM model achieved the best results in comparison with other models. This highlights the importance of considering spatial and temporal dependency in forecasting $PM_{2.5}$ concentrations. With longer prediction length, errors in all models increased however with a considerably slower rate for our model. This reveals the capability of our model for prediction in different (short, middle and long) time length. Compared with MLR which assumes linear correlation, models with nonlinear assumption exhibited better results. Our model also showed better performance compared with other studies such as a deep recurrent neural network model developed by Fan et al. (2017) with RMSE 29.10 μg/m^3 for 1-hour and 35.74 μg/m^3 for 8-hour forecasting, a BP-neural network model with RMSE 52.76 μg/m^3 for 72-hour forecasting over east China (Mao et al., 2017) and the Conv-LSTM model for $PM_{2.5}$ forecasting over Beijing with RMSE 24.23 μg/m^3 for 1-hour prediction (Huang and Kuo, 2018).

In order to verify the prediction accuracy of the models at different stations, the temporally averaged IA and MAE values for predictions over different time intervals (intervals in Table 4) at each station are shown in Fig. 6. Generally speaking, our model showed steady performance and there was no extreme fluctuation observed spatially. For all stations, the highest IA and lowest MAE were derived by the GC-LSTM. Because GC-LSTM model treats the predictors of all sites as a whole graph, it helps to reduce variance of the prediction errors between different stations and improves the prediction accuracy. Moreover, by adding the neighborhood information, stations like 1057A–1066A (stations in Zhangjiakou and Chengde) which got relatively poor performances in three baselines, received better results with the GC-LSTM model. It can also be seen that all three artificial-network-based models exhibited lower IA values on stations with low $PM_{2.5}$ concentrations. This shows, at low concentrations of $PM_{2.5}$, the models tend to overestimate the concentrations compared with observed values.

For air pollutant prediction models, the capability to predict over-standard $PM_{2.5}$ concentrations is crucial. According to the Ambient Air Quality Standards in China (EPA, 2012), exposure to $PM_{2.5}$ greater than 115 μg/m^3 can cause adverse health effects. Considering this as the threshold, we calculated the recall rate (RR, the fraction of $PM_{2.5}$ observed values over the threshold that are also predicted as overlimit values by the model) and false alarm rate of predictions (FAR, the fraction of $PM_{2.5}$ observed values below the threshold that are falsely predicted as overlimit values by the model) to evaluate the effectiveness of different models in forecasting high concentrations (Table 5). As recall rates of MLR model and FNN model decrease rapidly to lower than 15% in 72-hour prediction, we can observe that considering temporal dependency in the GC-LSTM and LSTM can improve the model performance specially in long term forecasting. For the LSTM model, it achieved lowest false alarm rates in short term forecasting (lower than 3% in 12-hour prediction). However, it was tailored to the relatively lowest detection rates in the mean time, which indicated that LSTM tended to underestimate $PM_{2.5}$ values in short-term prediction.

Table 3
Details of the experimental settings.

Parameter	Value
Number of records	831,744
Time interval (h)	1
Training set	60%
Validation set	20%
Test set	20%
Prediction length (T', h)	[1,2,4,8,12,24,48,72]
History length (T, h)	24
Number of stations	76
Parameter update	RMSprop
Support (K)	2
Training Epochs	200
Loss function	Root mean square error

Table 4
The results (IA, MAE, RMSE) of different models for hourly forecasting values of $PM_{2.5}$.

Model	Metric	+1 h	+2 h	+4 h	+8 h	+12 h	+24 h	+48 h	+72 h
MLR	IA	0.92	0.90	0.87	0.81	0.76	0.66	0.51	0.46
	MAE	24.35	26.80	31.15	35.91	38.65	43.23	46.94	48.23
	RMSE	38.03	41.70	47.52	53.75	57.38	63.32	67.06	69.32
FNN	IA	0.95	0.94	0.91	0.88	0.85	0.80	0.76	0.73
	MAE	17.77	20.38	24.16	28.29	30.74	34.17	35.85	37.01
	RMSE	28.80	32.61	39.16	44.31	47.90	53.03	56.12	58.00
LSTM	IA	0.96	0.95	0.92	0.89	0.88	0.87	0.85	0.87
	MAE	16.82	19.16	23.12	26.46	27.47	29.34	31.69	29.98
	RMSE	27.19	31.20	37.56	42.05	43.65	45.58	48.13	46.40
GC-LSTM	IA	**0.98**	**0.97**	**0.95**	**0.94**	**0.94**	**0.94**	**0.93**	**0.92**
	MAE	**13.72**	**16.72**	**19.79**	**21.62**	**21.96**	**23.10**	**23.66**	**24.21**
	RMSE	**22.41**	**27.19**	**32.10**	**34.35**	**34.19**	**35.92**	**37.32**	**38.83**

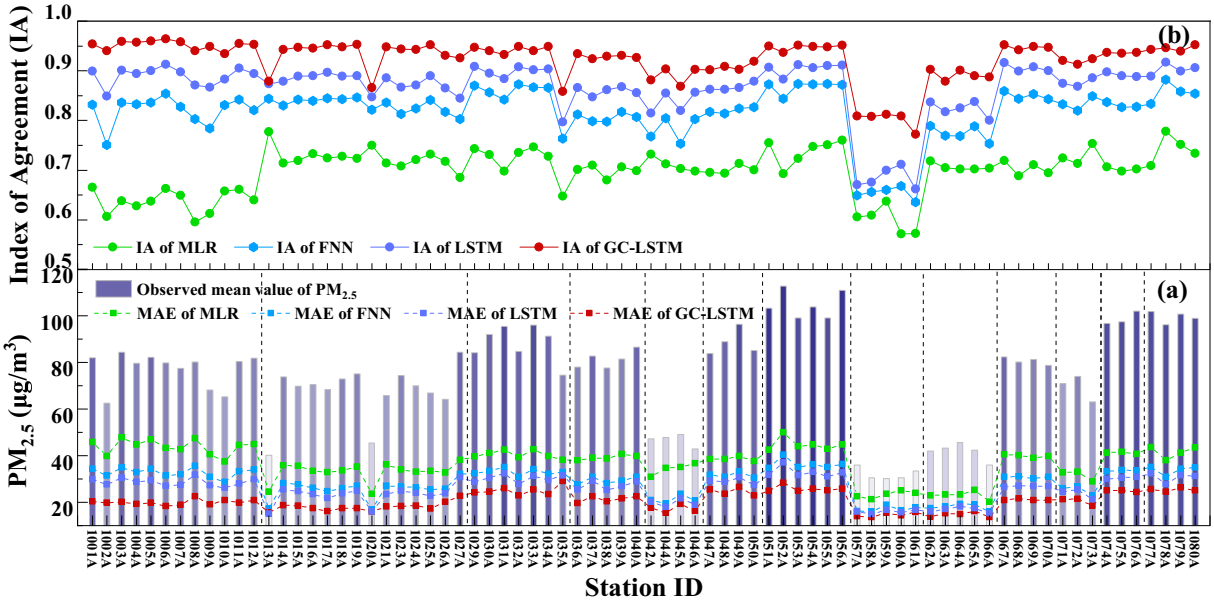


Fig. 6. (a) Distribution of average values of $PM_{2.5}$ at each station in study area (histogram) and distribution of MAE values of MLR, FNN, LSTM and GC-LSTM models (points). (b) Distribution of IA values of MLR, FNN, LSTM and GC-LSTM models.

Generally, our GC-LSTM model outperformed other baselines both on short- and long-term predictions with similar low false alarm rate.

The comparative results of linear correlation between observed $PM_{2.5}$ and corresponding predicted values for 1-hour and 72-hour time spans using the GC-LSTM and three baselines are shown in Fig. 7. For all models, the slopes of the linear regression equations are less than 1.0 and the intercepts are positive. From this, the tendency of the models to underestimate high concentrations and overestimate low concentrations is inferable. As can be seen, for prediction results based on both 1-hour (short-term) and 72-hour (long-term) time steps, the GC-LSTM obtained the highest R^2 values (0.91 and 0.72 respectively) and the minimal MAE values (13.72 μg/m^3 and 24.21 μg/m^3 respectively) among all the models, indicating strong agreement between predictions and observed values. Moreover, in 72-hour prediction, correlation improvement was more prominent with R^2 significantly increasing from 0.13 by the MLR to 0.72 by the GC-LSTM. These results suggest that applying spatiotemporal dependency and non-linear assumption can improve the model performance especially for long-term prediction.

Table 5
Average recall rates (%) and false alarm rates (%) of 76 stations (all available stations in study area) based on different approaches, with a threshold $PM_{2.5}$ value of 115 μg/m^3.

Prediction	MLR		FNN		LSTM		GC-LSTM	
	RR	FAR	RR	FAR	RR	FAR	RR	FAR
+1 h	73.84	6.01	71.55	5.40	68.12	1.38	81.25	2.29
+2 h	70.31	6.59	68.38	5.80	63.50	1.74	80.85	3.75
+4 h	63.67	7.40	62.66	6.31	58.48	2.32	77.44	4.33
+8 h	56.16	7.69	54.13	6.27	51.40	2.46	74.50	4.52
+12 h	51.98	7.90	48.46	6.13	49.09	2.46	74.16	4.56
+24 h	39.37	7.92	29.42	4.32	57.95	5.12	71.01	4.44
+48 h	20.38	4.91	13.54	4.45	56.81	5.28	69.60	4.62
+72 h	14.59	3.70	12.79	4.91	57.80	4.83	68.49	4.65

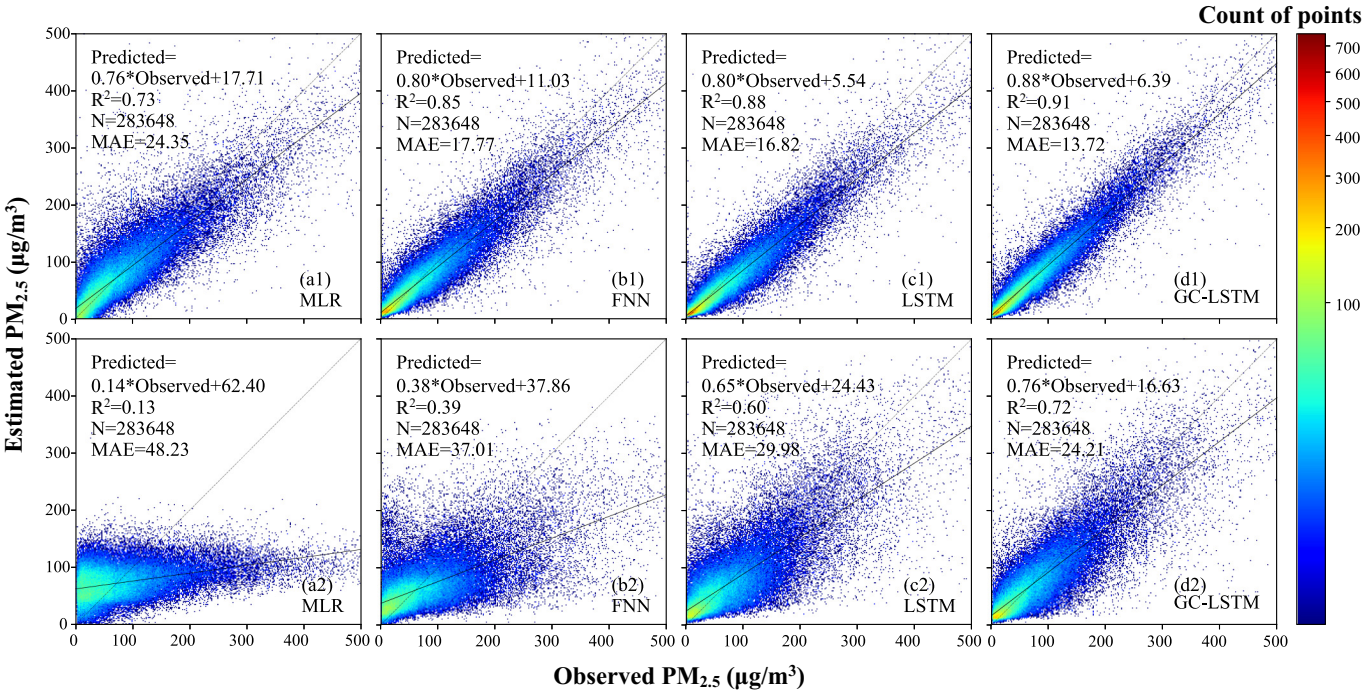


Fig. 7. Correlation between the observed and estimated $PM_{2.5}$ concentrations in 1-hour (the above) and 72-hour (the bottom) prediction by different models on test dataset: (a1) and (a2) MLR model, (b1) and (b2) FNN model, (c1) and (c2) LSTM model, (d1) and (d2) GC-LSTM model. The solid line and dashed line are the regression line and $y = x$ reference line, respectively.

Based on our proposed model, the spatial distributions of the GC-LSTM predictions of different intervals and corresponding observations using our test dataset are shown in Fig. 8. Following the results in Table 4, more uncertainty of the predictions was brought out while the time intervals increased, especially for stations located in the southwest part of our study area. This may be due to the existence of temporal emission sources in this region (e.g.,industry and construction), therefore we believe that our model will perform better by using more explanatory variables. One station that needs to be pointed out specially is 1035A, which located on the southwest corner of Shijiazhuang city and with lower average $PM_{2.5}$ concentrations than its neighborhood. This station got much higher MAE than others, with an MAE value up to 36.60 μg/m^3 in 72-hour prediction. It may be due to the occasional big difference between

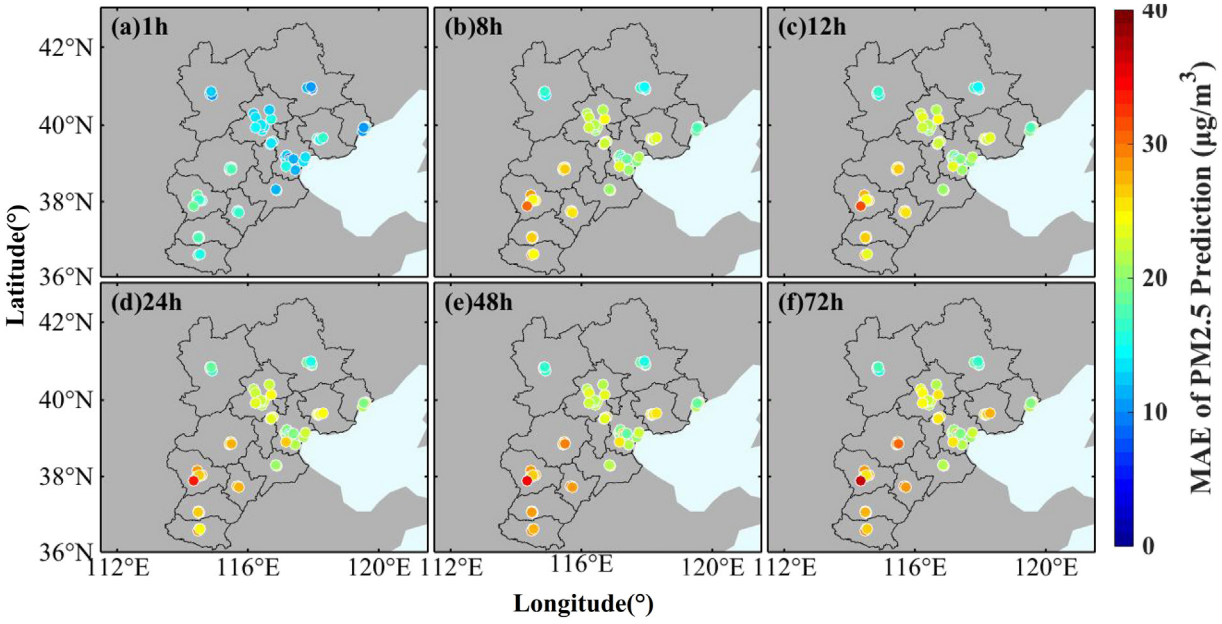


Fig. 8. Mean absolute errors (MAEs) between predicted and observed $PM_{2.5}$ concentrations for individual monitoring sites: (a–f) 1-hour, 8-hour, 12-hour, 24-hour, 48-hour and 72-hour results by GC-LSTM model.

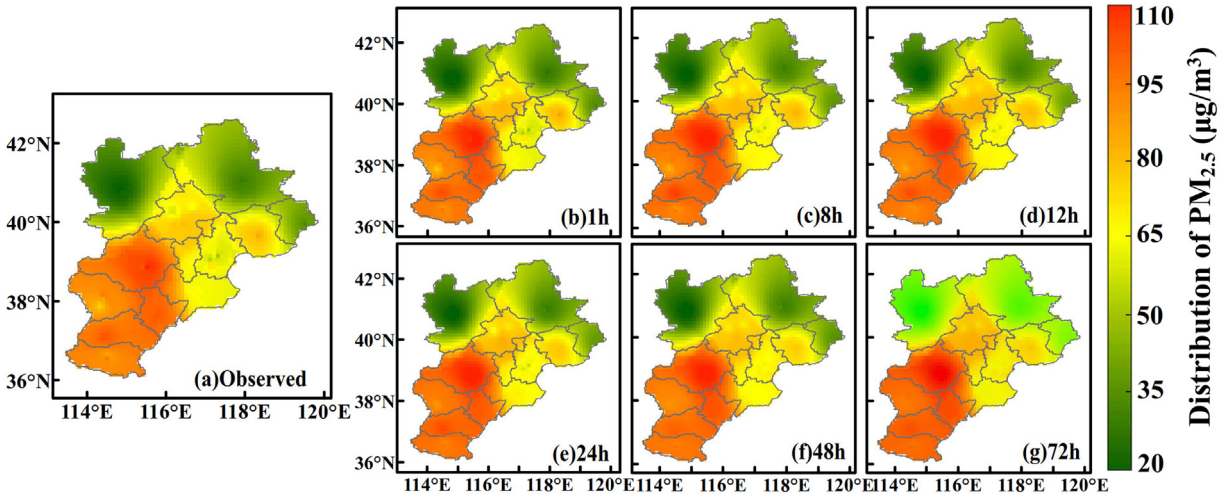


Fig. 9. Surface distribution of average $PM_{2.5}$ concentrations for (a) all observed test data and (b–g) corresponding 1-hour, 8-hour, 12-hour, 24-hour, 48-hour and 72-hour prediction by GC-LSTM model.

the target and its neighborhood that can cause more uncertainty in graph convolution process. But Considering the field-like features of $PM_{2.5}$ distribution, taking spatial dependency into account for modelling are still recommended for improving the overall forecasting accuracy.

By spatially interpolating the predicted $PM_{2.5}$ concentrations, we can get the hourly surface distribution of air pollutants over the study area. We applied the inverse distance weighted interpolation to generate averaged surface distribution of $PM_{2.5}$ concentrations of the test dataset for different time intervals (Fig. 9). As can be seen in Fig. 9a, severely-polluted areas are mainly located in southeast Hebei with the large mean $PM_{2.5}$ concentrations more than 100 μg/m^3. High anthropogenic emissions contributed by industrial emissions can be considered as the reasons for high $PM_{2.5}$ in this area. Low $PM_{2.5}$ were observed in northern areas of Hebei, which are in hilly topography and low population density. Our results are similar to those of a previous study done by Wang et al. (2017a). Compared with the interpolation in Fig. 9a, it can be observed that the predicted interpolation showed a consistent spatial distribution. But still, for the locations with $PM_{2.5}$ concentrations lower than 60 μg/m^3, our predicted surface overestimated the values, which became more obvious in 72-hour prediction.

To further explore the temporal variation of the ground-level $PM_{2.5}$ concentrations and our model predictions, we calculated the daily averaged $PM_{2.5}$ forecasted by the GC-LSTM using the test data and the corresponding daily averaged $PM_{2.5}$ observations (Fig. 10). Generally speaking, there is a consistency between the trend of the forecasted $PM_{2.5}$ and the observations, which verifies the feasibility of our model to capture the temporal variations. It can be seen that our proposed model showed a tendency of overestimating the low values (less than 100 μg/m^3) and underestimating the high values (more than 200 μg/m^3). But in general, our model made sound predictions for $PM_{2.5}$ concentration values under 300 μg/m^3. As for the abrupt extremely high $PM_{2.5}$ concentrations, especially for concen-

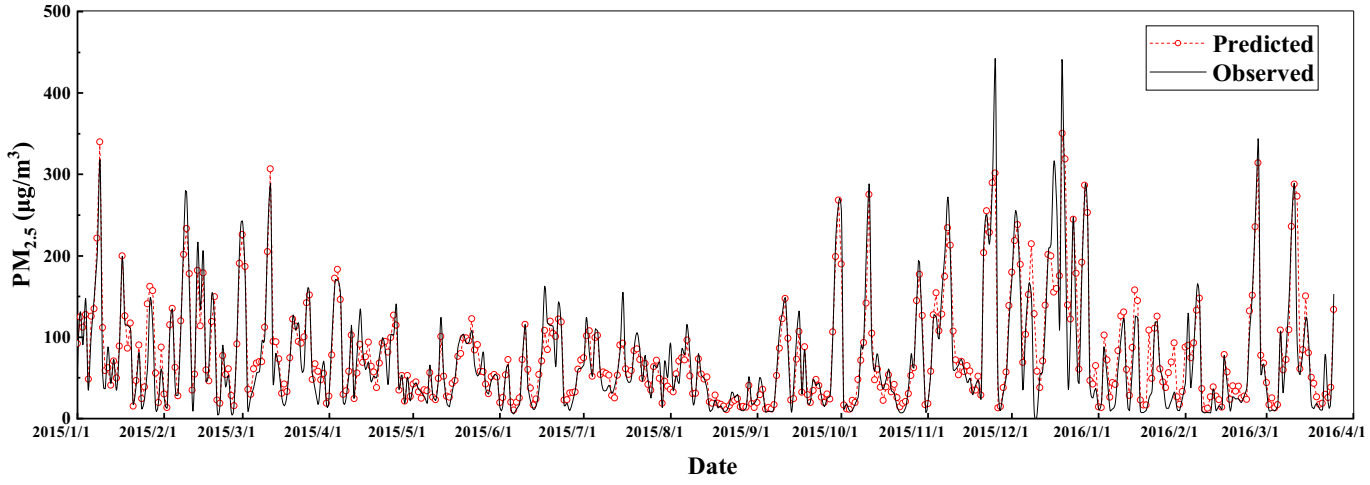


Fig. 10. Temporal distribution of daily averaged observed $PM_{2.5}$ concentrations on the test dataset (the black) and the corresponding 72-hour predictions (the red) by GC-LSTM model for Jing-Jin-Ji area from January 1, 2015 to April 1, 2016.

trations more than 400 μg/m^3, there is a relatively large gap between the predictions and observations, which may be the future directions of our work and highlights the role of more explanatory variables.

4. Conclusions

Increasing availability of historical data and computing resources has facilitated us to develop more sophisticated models for air pollution prediction. In this study, a hybrid model based on deep artificial neural networks for $PM_{2.5}$ mass concentration forecasting was proposed. Our model is based on an integration of deep learning methods of spectral graph convolution neural network and long short-term memory to take the spatiotemporal dependency among massive ground observations into account in prediction model. The hybrid model was first trained to confirm the parameters of the network, and then it was justified through the verification with test sample data. The experimental results showed that our model outperformed the state-of-the-art methods in forecasting of $PM_{2.5}$ values. The performance advantage of our proposed method is consistent both for optimal average prediction on all stations and the minimum fluctuation trend of errors among different sites. With acceptable recall rates, false alarm rates and R^2 in 1–72-hour prediction, our proposed method is verified to be an effective statistical method to estimate ground $PM_{2.5}$ concentration with historical data. Our method can be used in future studies and other regions for air pollutant prediction.

Acknowledgments

We thank Mr.Wu Chunlin for his helpful advices, as well as Dr. Hou Junxiong for providing us the data.

References

Bey, I., Jacob, D.J., Yantosca, R.M., Logan, J.A., Field, B.D., Fiore, A.M., Li, Q., Liu, H.Y., Mickley, L.J., Schultz, M.G., 2001. Global modeling of tropospheric chemistry with assimilated meteorology: model description and evaluation. J. Geophys. Res. 106, 23073–23095.

Bruna, J., Zaremba, W., Szlam, A., Lecun, Y., 2014. Spectral networks and locally connected networks on graphs. Proceedings of the International Conference on Learning Representations.

Byun, D., 1999. Science algorithms of the EPA Models-3 community multiscale air quality (CMAQ) modeling system. J. Jpn. Soc. Atmos. Environ. 43, 79–91.

Chaloulakou, A., Grivas, G., Spyrellis, N., 2003. Neural network and multiple regression models for PM_{10} prediction in athens: a comparative assessment. J. Air Waste Manage. Assoc. 53, 1183–1190.

Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering. Proceedings of the Advances in Neural Information Processing Systems. pp. 3844–3852.

EPA, 2012. Technical regulation on ambient air quality index (on trial). Technical Report, China Environmental Science Press Beijing, China.

Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H., Lin, S., 2017. A spatiotemporal prediction framework for air pollution based on deep RNN. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. IV-4/W2, 15–22.

Fan, R.K.C., 1997. Spectral Graph Theory. American Mathematical Society, pp. 212.

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J., 2015. Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. 107, 118–128.

Geng, G., Zhang, Q., Martin, R.V., van Donkelaar, A., Huo, H., Che, H., Lin, J., He, K., 2015. Estimating long-term $PM_{2.5}$ concentrations in China using satellite-based aerosol optical depth and a chemical transport model. Remote Sens. Environ. 166, 262–270.

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, 6957–6975.

He, Q., Huang, B., 2018. Satellite-based mapping of daily high-resolution ground $PM_{2.5}$ in China via space-time regression modeling. Remote Sens. Environ. 206, 72–83.

Henaff, M., Bruna, J., LeCun, Y., 2015. Deep Convolutional Networks on Graph-structured Data. arXiv preprint. arXiv:1506.05163.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Huang, C.J., Kuo, P.H., 2018. A deep CNN-LSTM model for particulate matter ($PM_{2.5}$) forecasting in smart cities. Sensors 18.

Karimian, H., Li, Q., Li, C., Chen, G., Mo, Y., Wu, C., Fan, J., 2018. Spatio-temporal variation of wind influence on distribution of fine particulate matter and its precursor gases. Atmos. Pollut. Res. https://doi.org/10.1016/j.apr.2018.06.005.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. Proceedings of the International Conference on Learning Representations.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T., 2017. Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environ. Pollut. 231, 997–1004.

Li, Z., Zang, Z., Li, Q.B., Chao, Y., 2013. A three-dimensional variational data assimilation system for multiple aerosol species with WRF/Chem and an application to $PM_{2.5}$ prediction. Atmos. Chem. Phys. Discuss. 13, 4265–4278.

Ma, Z., Hu, X., Huang, L., Bi, J., Liu, Y., 2014. Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing. Environ. Sci. Technol. 48, 7436–7444.

Mao, X., Shen, T., Feng, X., 2017. Prediction of hourly ground-level $PM_{2.5}$ concentrations 3 days in advance using neural networks with satellite data in eastern China. Atmos. Pollut. Res. 8, 1005–1015.

Mckendry, I.G., 2002. Evaluation of artificial neural networks for fine particulate pollution (PM_{10} and $PM_{2.5}$) forecasting. J. Air Waste Manage. Assoc. 52, 1096–1101.

Ong, B.T., Sugiura, K., Zettsu, K., 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting $PM_{2.5}$. Neural Comput. & Applic. 27, 1553–1566.

Pérez, P., Trier, A., Reyes, J., 2000. Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. 34, 1189–1196.

Remer, L.A., Kaufman, Y., Tanré, D., Mattoo, S., Chu, D., Martins, J.V., Li, R.-R., Ichoku, C., Levy, R., Kleidman, R., et al. 2005. The modis aerosol algorithm, products, and validation. J. Atmos. Sci. 62, 947–973.

Reyes, J.M., Serre, M.L., 2014. An LUR/BME framework to estimate $PM_{2.5}$ explained by on road mobile and stationary sources. Environ. Sci. Technol. 48, 1736–1744.

Shamsoddini, A., Aboodi, M.R., Karami, J., 2017. Tehran air pollutants prediction based on random forest feature selection method. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLII-4/W4, 483–488.

Shi, X., Chen, Z., Wang, H., Woo, W.C., Woo, W.C., Woo, W.C., 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. International Conference on Neural Information Processing Systems. pp. 802–810.

Sun, W., Sun, J., 2016. Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. J. Environ. Manag. 188, 144–152.

Wang, W., Mao, F., Du, L., Pan, Z., Gong, W., Fang, S., 2017a. Deriving hourly $PM_{2.5}$ concentrations from Himawari-8 AODs over Beijing-Tianjin-Hebei in China. Remote Sens. 9, 858.

Wang, W., Mao, F., Pan, Z., Du, L., Gong, W., 2017b. Validation of VIIRS AOD through a comparison with a sun photometer and MODIS AODs over Wuhan. Remote Sens. 9, 403.

Wu, C., Li, Q., Hou, J., Karimian, H., Chen, G., 2018. $PM_{2.5}$ concentration prediction using convolutional neural networks. Sci. Surv. Mapp. 43, 68–75.

Zang, L., Mao, F., Guo, J., Gong, W., Wang, W., Pan, Z., 2018. Estimating hourly PM_1 concentrations from Himawari-8 aerosol optical depth in China. Environ. Pollut. 241, 654–663.

Zhang, T., Gong, W., Wang, W., Ji, Y., Zhu, Z., Huang, Y., 2016. Ground level $PM_{2.5}$ estimates over China using satellite-based geographically weighted regression (GWR) models are improved by including NO_2 and enhanced vegetation index (EVI). Int. J. Environ. Res. Public Health 13, 1215. https://doi.org/10.3390/ijerph13121215.

Zhang, T., Zhu, Z., Gong, W., Zhu, Z., Sun, K., Wang, L., Huang, Y., Mao, F., Shen, H., Li, Z., et al. 2018. Estimation of ultrahigh resolution $PM_{2.5}$ concentrations in urban areas using 160 m Gaofen-1 AOD retrievals. Remote Sens. Environ. 216, 91–104.