Environmental Pollution 231 (2017) 997–1004

Contents lists available at ScienceDirect

Environmental Pollution

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

Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation☆

Xiang Li [a, b], Ling Peng [a], Xiaojing Yao [a, *], Shaolong Cui [a], Yuan Hu [a, b], Chengzeng You [a, b], Tianhe Chi [a]

[a] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 24 February 2017
Received in revised form
29 August 2017
Accepted 31 August 2017
Available online 25 September 2017

Keywords:
Air pollutant concentration predictions
Long short-term memory neural network (LSTM NN)
Recurrent neural network
Spatiotemporal correlation
Multiscale prediction

ABSTRACT

Air pollutant concentration forecasting is an effective method of protecting public health by providing an early warning against harmful air pollutants. However, existing methods of air pollutant concentration prediction fail to effectively model long-term dependencies, and most neglect spatial correlations. In this paper, a novel long short-term memory neural network extended (LSTME) model that inherently considers spatiotemporal correlations is proposed for air pollutant concentration prediction. Long short-term memory (LSTM) layers were used to automatically extract inherent useful features from historical air pollutant data, and auxiliary data, including meteorological data and time stamp data, were merged into the proposed model to enhance the performance. Hourly $PM_{2.5}$ (particulate matter with an aerodynamic diameter less than or equal to 2.5 μm) concentration data collected at 12 air quality monitoring stations in Beijing City from Jan/01/2014 to May/28/2016 were used to validate the effectiveness of the proposed LSTME model. Experiments were performed using the spatiotemporal deep learning (STDL) model, the time delay neural network (TDNN) model, the autoregressive moving average (ARMA) model, the support vector regression (SVR) model, and the traditional LSTM NN model, and a comparison of the results demonstrated that the LSTME model is superior to the other statistics-based models. Additionally, the use of auxiliary data improved model performance. For the one-hour prediction tasks, the proposed model performed well and exhibited a mean absolute percentage error (MAPE) of 11.93%. In addition, we conducted multiscale predictions over different time spans and achieved satisfactory performance, even for 13–24 h prediction tasks (MAPE = 31.47%).

1. Introduction

Air pollution is a serious environmental problem that has attracted increasing attention worldwide (Kurt and Oktay, 2010). Certain air pollutants, such as $PM_{2.5}$ (particulate matter with an aerodynamic diameter less than or equal to 2.5 μm), can traverse the nasal passages during inhalation and reach the throat and even the lungs. Long-term exposure to ambient fine particulate matter can negatively affect human health (Dockery et al., 1993; Iii et al., 2002; Krewski et al., 2009) and cause respiratory and cardiovascular diseases and some other illnesses (Kappos et al., 2004; Neuberger et al., 2004; Wilson et al., 2005; Bravo and Bell, 2011). Therefore, obtaining air pollutant concentration information in real time is significant for air pollution control and the prevention of health issues due to air pollution (Zheng et al., 2013).

In recent years, many research efforts have focused on enriching approaches to predicting air pollutant concentrations. In general, methods of predicting air pollutant concentrations fall into two major categories: deterministic and statistical methods.

Deterministic methods adopt meteorological principles and statistical methods to model the emission, dispersion, transformation, diffusion and removal processes of pollutants based on atmospheric physics and chemical reactions; thus, the spatiotemporal distributions of air pollutants are simulated at different scales and orientations (Bruckman, 1993; Coats, 1996; Lurmann, 2000; Guocai, 2004; Baklanov et al., 2008; Kim et al., 2010; Jeong et al., 2011). These methods are viewed as model-based methods because their structures are predefined based on certain theoretical

☆ This paper has been recommended for acceptance by Dr. Hageman Kimberly Jill.
* Corresponding author.
E-mail address: yaoxj@radi.ac.cn (X. Yao).

http://dx.doi.org/10.1016/j.envpol.2017.08.114
0269-7491/

hypotheses, and the parameters can be calculated via specific priori knowledge. Many air quality models have been developed to simulate the complicated process of air pollutant diffusion. Representative methods, such as the Community Multiscale Air Quality (CMAQ) model (Chen et al., 2014), Nested Air Quality Prediction Modeling System (NAQPMS) (Wang et al., 2001) and WRF-Chem model (Saide et al., 2011), are commonly adopted for air pollutant concentration forecasting in urban areas. Although developed theories provide valuable insights for understanding pollutant diffusion mechanisms, most of these theoretical models are relevant to sophisticated priori knowledge, unreliable and limited data, and various usage constraints (Vautard et al., 2007; Stern et al., 2008).

Statistical methods, however, avoid sophisticated theoretical models and simply apply statistics-based models to predict air quality. Widely used methods include the multiple linear regression (MLR) method (Li et al., 2011), the autoregressive moving average (ARMA) method (Box and Jenkins, 1976), the support vector regression (SVR) method (Nieto et al., 2013), the artificial neural network (ANN) method (Hooyberghs et al., 2005), and hybrid methods (Díaz-Robles et al., 2008; Chen et al., 2013). Among these models, the ANN method, which can perform nonlinear mapping and is self-adaptive and robust, generally provides satisfying performance; therefore, it has been widely used in time series forecasting fields (Yoon et al., 2011). Recently, various ANN structures have been developed to improve predictions of air pollutant concentrations. Typical examples include the widely employed multilayer perceptron (MLP) (Paschalidou et al., 2011), back propagation neural network (BPNN) (Kolehmainen et al., 2001), radial basis function neural network (RBF NN) (Lu et al., 2002), neuro-fuzzy neural network (NFNN) (Mishra and Goyal, 2016), general regression neural network (GRNN) (Antanasijević et al., 2013), and recurrent neural network (RNN) (Feng et al., 2011). Due to the dynamic nature of relevant atmospheric environments, RNNs are particularly suited to capturing the spatiotemporal evolution of air pollutant distributions because RNNs can handle arbitrary sequences of inputs, thereby guaranteeing the capacity to learn temporal sequences (Ma et al., 2015). Certain RNNs, such as the time delay neural network (TDNN) (Ong et al., 2016) and Elman neural network (Prakash et al., 2011), have been used for air pollutant prediction in previous studies. However, these RNN models face two issues: 1) in the RNN structure, the time lag must be determined in advance, which requires a considerable number of experiments to identify the optimum time lag; and 2) traditional RNNs fail to capture long time dependencies in input sequences, and training RNNs with long time lags is difficult because vanishing gradient and exploding gradient problems may be encountered (Hochreiter and Schmidhuber 1997).

To resolve these issues, a special RNN architecture referred to as a long short-term memory neural network (LSTM NN) was developed by Hochreiter and Schmidhuber (1997). Unlike traditional RNNs, LSTM NNs are capable of learning long time series and are not affected by the vanishing gradient problem. These features are especially important for modeling spatiotemporal air pollutant processes in which the air pollutant concentration of one station is related to the previous status and those at nearby stations because of pollutant transport processes.

In recent years, the LSTM NN has been successfully applied to many studies involving time series prediction, such as traffic flow prediction (Lv et al., 2015), wind power prediction (Felder et al., 2010), human trajectory prediction (Alahi et al., 2016), etc. Recently, Sak et al. (2016) (Sak et al., 2016) adopted the LSTM NN for pollution risk prediction, but they only classified the pollution risk ranking without conducting real-value predictions of air pollutant concentrations. Moreover, they made predictions separately for individual cities without considering the spatial correlations between monitoring stations. To the best of our knowledge, the LSTM NN has not been applied in the domain of air pollutant concentration prediction. This paper aims to extend the LSTM NN to spatiotemporal correlation modeling and air pollutant concentration prediction.

The contributions of this paper are as follows: (1) an LSTM NN is extended to capture the long-term spatiotemporal dependency of air pollutant concentrations, and a multiscale prediction framework which can forecast the air pollutant concentration over the next 24 h is presented; (2) the proposed method can effectively and automatically extract the spatiotemporal correlations within air pollutant concentration data; and (3) auxiliary data are integrated into a traditional LSTM NN model, and the integrated model exhibits better performance than traditional methods.

2. Data and methods

2.1. Data description

Hourly $PM_{2.5}$ concentration data from 12 air quality monitoring stations in downtown Beijing collected from Jan/01/2014 to May/28/2016 were obtained from the Ministry of Environmental Protection of China (http://datacenter.mep.gov.cn/). Concentrations were measured using a Thermo Fisher 1405F detector and calculated based on the tapered element oscillating microbalance (TEOM) method. Meteorological data from the same period were downloaded from The National Oceanic and Atmospheric Administration's (NOAA's) national climate data center (https://www.climate.gov/). In successive experiments, we chose four main factors from the meteorological dataset that are highly related to $PM_{2.5}$ concentrations: temperature, humidity, wind speed and visibility (Díaz-Robles et al., 2008; Saide et al., 2011; Guocai, 2004). Fig. 1 shows the distribution of the air quality monitoring stations (blue triangles) and the location of the meteorological station (green triangle). Simple linear interpolation was performed to fill in the missing values in both datasets. Our dataset contained 20196 records for each station. In our experiment, we randomly selected 80

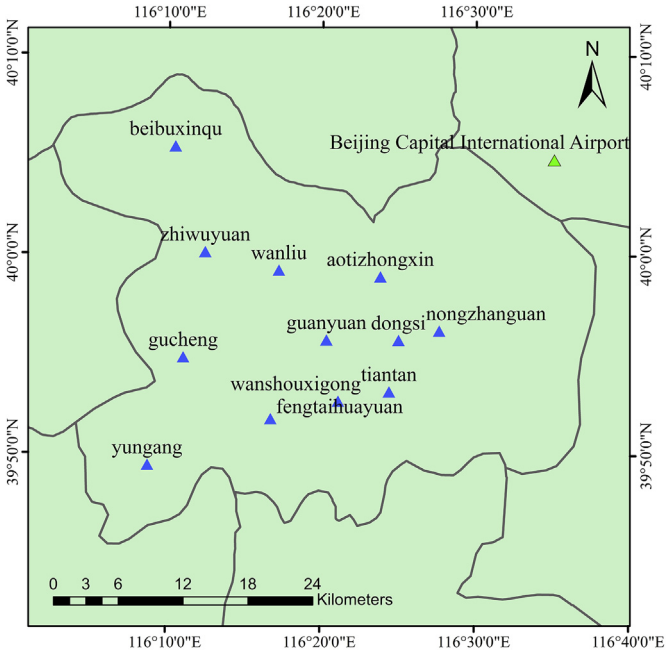

Fig. 1. Distribution of air quality monitoring stations in Beijing City.

percent of the data as the training set, and the remaining 20 percent was used as the test set.

2.2. Spatiotemporal correlation analysis

First, we analyzed the spatial correlation of $PM_{2.5}$ concentrations among the 12 stations. Pearson's correlation coefficient (Pearson, 1895) was used to measure the correlations, and the results are shown in Table 1. Notably, all correlation values are above 0.82 (p-value<0.05), which demonstrates that the $PM_{2.5}$ concentrations are highly correlated among the stations. This strong spatial correlation supports the use of a single model to predict the $PM_{2.5}$ concentrations at all stations rather than separate models for each station because nearby related inputs can improve the prediction performance.

Then, autocorrelation functions (Box and Jenkins, 1976) were used to measure the temporal correlations among the $PM_{2.5}$ concentration time series at each site. For time delay k, the autocorrelation coefficients can be calculated as follows:

$$\rho_k = \frac{Cov(y(t), y(t+k))}{\sigma_{y(t)}\sigma_{y(t+k)}} \tag{1}$$

where $y(t)$ and $y(t+k)$ denote the air pollutant concentrations at time t and time $t+k$, respectively, $Cov(\cdot)$ is the covariance and $\sigma(\cdot)$ is the standard deviation.

Fig. 2 shows the autocorrelation coefficients of each site, and the 12 curves represent the different sites. An obvious descending trend is observed with increasing time lag, confirming the fact that earlier events have a weaker influence on the current status. In addition, when the time lag is less than 17, the correlation function is higher than 0.5, indicating a high temporal correlation. These findings can be used to select the appropriate time lags for our prediction tasks.

Considering the high spatiotemporal correlations among sites and the historical status of sites, we applied the time-delayed $PM_{2.5}$ concentrations from all 12 stations as inputs and used the LSTM NN to automatically determine the spatiotemporal correlations.

2.3. LSTME model

An LSTM NN is extended in this study to predict the regional air pollutant concentration. The LSTM NN is a gated RNN that can effectively account for long time dependencies. A detailed introduction to this method is given in Appendix A. In our LSTM NN extended (LSTME) model, the LSTM layers were used to extract representative features from historical air pollutant concentration data. The overall prediction framework is shown in Fig. 3.

Time-delayed historical data from all monitoring stations were stacked to construct an input tensor for the LSTM layers (see "Main Inputs" in Fig. 3), and features in the spatially correlated data with long time dependencies were automatically extracted layer-by-layer (see "LSTMs" part in Fig. 3). A recursive arrow indicates that the layer extraction process can be repeated several times for optimum performance.

In addition to modeling spatiotemporal correlations using the LSTM NN, as discussed above, the use of auxiliary data represents another promising method of improving prediction performance. Previous studies have proved that meteorological factors play vital roles in the daily variability of pollutants (He et al., 2013; Bai et al., 2016). Here, current meteorological data along with one-hot encoded time stamp data were used to enhance the LSTM NN model. One-hot encoding is a common operation of changing categorical data into binarized codes. For example, monthly index data have 12 unique categorical values, and by using one-hot encoding, each monthly index can be transferred to a 12-dimensional vector (e.g., March is given as [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]). Specifically, we concatenated these auxiliary data into an LSTM output feature vector to provide additional input features for our model to learn (see "Auxiliary Inputs" in Fig. 3, the numbers in brackets indicate the data dimensions).

The first two or more LSTM layers were initially used to extract inherent features from the historical air pollutant data by learning over a long time span. Then, the month of year and hour of day data were encoded using the one-hot encoding method and merged with the extracted features, along with current meteorological data. Next, one or more fully connected layers ("FCs" in Fig. 3) were used to obtain further representations of the merged features. Finally, a fully connected layer was used to generate the prediction output.

To evaluate the effectiveness of the proposed method, three indicators, including the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), were used in our experiments. These indicators can be formulated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^*| \tag{3}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - y_i^*|}{y_i^*} \tag{4}$$

where y_i^* is the observed air pollutant concentration, y_i is the

Table 1
Pearson's coefficients between stations.

R	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12
s1	1.00	0.96	0.97	0.94	0.95	0.96	0.94	0.84	0.87	0.93	0.90	0.90
s2	0.96	1.00	0.96	0.96	0.93	0.95	0.93	0.83	0.86	0.94	0.90	0.89
s3	0.97	0.96	1.00	0.94	0.96	0.95	0.96	0.85	0.90	0.94	0.92	0.93
s4	0.94	0.96	0.94	1.00	0.91	0.93	0.91	0.82	0.84	0.94	0.90	0.89
s5	0.95	0.93	0.96	0.91	1.00	0.94	0.94	0.85	0.89	0.90	0.89	0.90
s6	0.96	0.95	0.95	0.93	0.94	1.00	0.92	0.84	0.87	0.92	0.89	0.88
s7	0.94	0.93	0.96	0.91	0.94	0.92	1.00	0.88	0.92	0.92	0.92	0.93
s8	0.84	0.83	0.85	0.82	0.85	0.84	0.88	1.00	0.87	0.83	0.85	0.85
s9	0.87	0.86	0.90	0.84	0.89	0.87	0.92	0.87	1.00	0.85	0.89	0.90
s10	0.93	0.94	0.94	0.94	0.90	0.92	0.92	0.83	0.85	1.00	0.93	0.91
s11	0.90	0.90	0.92	0.90	0.89	0.89	0.92	0.85	0.89	0.93	1.00	0.94
s12	0.90	0.89	0.93	0.89	0.90	0.88	0.93	0.85	0.90	0.91	0.94	1.00

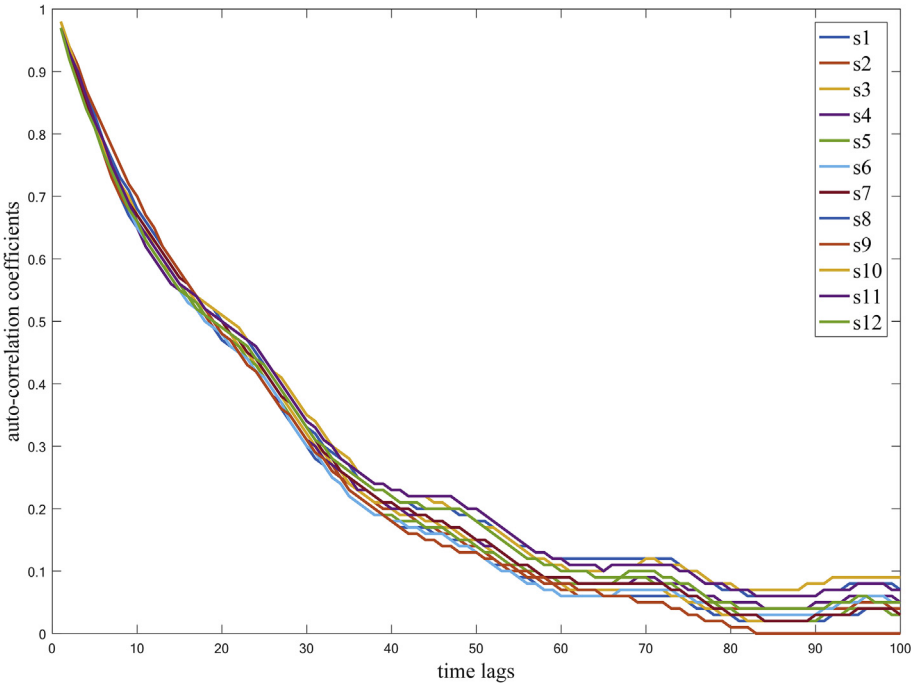

Fig. 2. Variations among the autocorrelation coefficients with respect to different time lags.

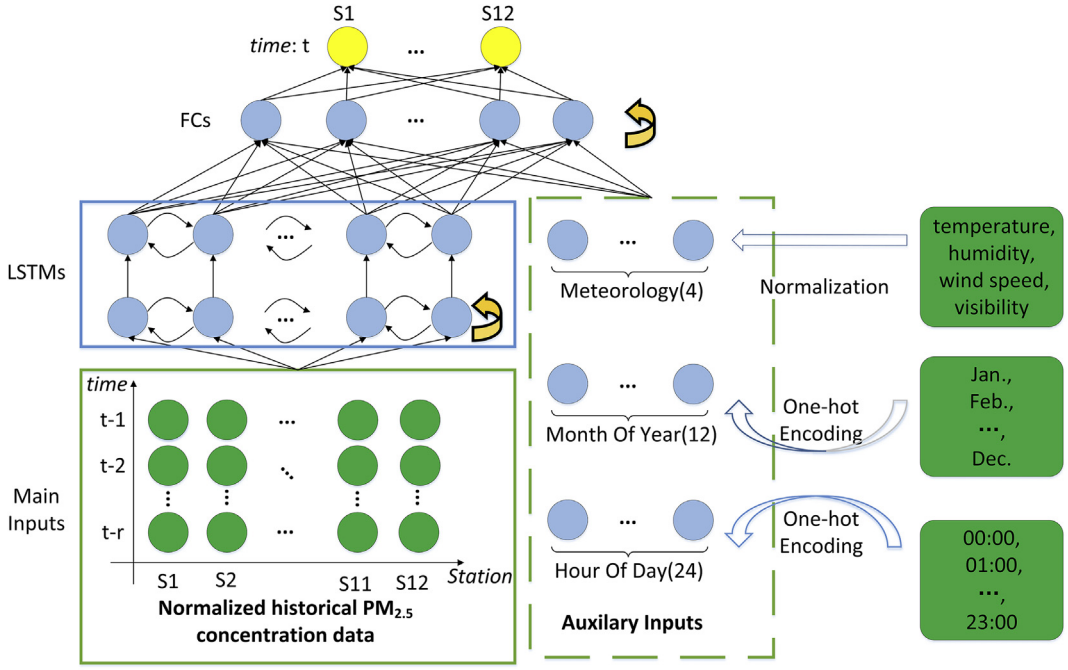

Fig. 3. Network framework of the LSTME model for air pollutant concentration prediction. The main inputs (historical $PM_{2.5}$ concentration data) are included in green solid boxes; auxiliary inputs (meteorological data, month of year and hour of day) are included in green dashed boxes; *r* indicates the time lag; and the numbers in the brackets indicate the data dimensions. A recursive arrow indicates that the processing of this layer can be repeated. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

predicted air pollutant concentration, and *n* is the number of test samples.

2.4. Network architecture

Several hyperparameters should be preset before building the LSTME prediction architecture (shown in Fig. 3), including the number of LSTM layers, the number of nodes in each LSTM layer, the number of fully connected layers, the number of nodes in each fully connected layer, the time lags, and the learning rate. We investigated the effect of each parameter while keeping the other parameters fixed, and a random search method with 5-fold cross-validation was applied to find the optimum hyperparameters.

To do so, our model requires some basic settings. For simplicity,

the number of nodes in each neuron layer was set to an equivalent value chosen from a candidate set of {600, 800, 1000, 1400, 2000}. We built our model using two LSTM layers and one fully connected layer. We chose this setting based on the results of several comparative experiments, which showed that this configuration yielded the best performance.

First, we examined the influences of different time lags. The prediction performance (shown in Table 2) showed that with a time lag r equal to 8, our LSTME model achieved optimal performance, as indicated by the RMSE, MAE and MAPE. According to previous studies, a small time lag cannot guarantee enough long-term memory inputs for our LSTME model; thus, the model cannot fully exploit the LSTM NN for long-term memory modeling. Large time lags permit an increased number of unrelated inputs, which increase the model's complexity and the difficulty of learning useful features. As a compromise, the time lag was set to 8, which was the most appropriate setting for our model.

Next, we investigated the effect of the number of nodes in each neuron layer, we fixed the time lag to 8 based on the above results. The results (shown in Table 3) show that with the increase in the number of neuron nodes, the prediction performance improves slightly. Thus, we set the number of nodes to 1000 in the successive experiments to optimize both the accuracy and time efficiency.

3. Results and discussion

3.1. Prediction performance

After determining the best network architecture for our prediction task, the training set was utilized to train our LSTME model until convergence. Evaluations were conducted using the test set, and Fig. 4 shows the predicted and observed $PM_{2.5}$ concentrations. Fig. 4 shows that the predicted data are generally consistent with the observed data. The R^2 value between the observed and predicted data indicated that 98% of the explained variance was captured by the model.

In addition, we evaluated the rank prediction performance of the proposed model. According to the National Technical Regulation of the Ambient Air Quality Index, we generated predicted and ground (true) rankings and calculated the overall ranking accuracy. The overall rank prediction accuracy was 90.3%, indicating satisfactory performance in rank prediction. Only 2.9% of the test samples received a predicted ranking that was overestimated, and only 6.8% of the test samples received a predicted ranking that was underestimated.

3.2. Comparison of experiments

We compared the performance of the proposed LSTME model with the performances of the STDL (Li et al., 2016), TDNN, SVR, and ARMA models. These models were trained and tested using the same training and test sets applied for the LSTME model; however, the input data differed slightly in each model. The TDNN and STDL models use the same inputs as our LSTME model, but the network architecture differs. The TDNN model adopts a traditional neural network for feature representation, while the STDL model uses stacked autoencoders in a dimension reduction-like manner. These three models have the capability of simultaneously predicting the air pollutant concentrations of all stations. The SVR and ARMA models, which are merely time series prediction models, were used to conduct prediction experiments for each station separately using input data from a single station. We integrated the prediction performance after each prediction. In addition, to evaluate the importance of auxiliary data, we conducted an additional experiment using the LSTM NN model without auxiliary data (dashed box in Fig. 3). The prediction performance of all models is shown in Table 4.

Table 2
Effect of time lags.

Time lag	RMSE	MAE	MAPE (%)
4	13.78	6.39	11.00
6	10.64	4.27	8.38
8	10.82	4.29	8.06
12	11.15	4.41	8.17
16	11.62	4.52	8.51

Table 3
Effect of the number of neuron nodes in each layer.

No. Nodes	RMSE	MAE	MAPE (%)
600	11.43	4.74	8.66
800	11.21	4.50	8.35
1000	10.82	4.29	8.06
1400	10.65	4.12	7.91
2000	10.40	4.02	7.73

Three useful findings can be extracted from Table 4. First, compared with the three "shallow" models (the SVR, ARMA and TDNN models), the three deep learning-based models (LSTME, LSTM NN and STDL models) exhibited better prediction performance due to their capability to represent high-level spatiotemporal features. This finding is consistent with those of previous studies (Li et al., 2016), in which deep architecture was found most suitable for modeling complex spatiotemporal processes. Second, compared with the TDNN model, the LSTM-based models, including the LSTME model and the traditional LSTM NN model, exhibited higher prediction precision, as indicated by the RMSE, MAE and MAPE values. This result suggests that the LSTM NN can more efficiently capture spatiotemporal correlations. Table 4 also indicates that the LSTME performs better than the traditional LSTM NN model, which suggests that auxiliary data can improve the prediction performance.

3.3. Multiscale predictions

Intuitively, historical data from different periods have different effects on future time lags. Therefore, as shown in Fig. 5, we grouped the air pollutant concentration data within particular time lags to formulate inputs (shown in the solid rectangle) for multiscale prediction tasks, which were used to train separate models. Each blue dashed arrow shown in Fig. 5 represents a predictor. Over the next 3 h, we trained a separate model in each hour. Then, we divided the next 4–24 h into three time lags (4–6, 7–12, and 13–24 h) and trained separate models to predict the mean air pollutant concentration of each time lag.

We separately selected the appropriate hyperparameter settings for each multiscale prediction task to yield the best overall performance. However, to simplify the experimental calculations, a fixed network structure with two LSTM layers and one fully connected layer was used in all tasks. Here, we only tested the effects of different time lags. After performing grid searches, as previously described, we obtained the best architecture for the different prediction tasks. These results are listed in Table 5.

Table 5 shows that as the prediction time lag increased, the optimum time lag increased and the prediction performance rapidly decreases, as indicated by the MAPE values, which varied from 11.93% to 31.47%. Notably, long-term prediction tasks are instinctively more difficult; thus, they require more relevant historical input data, such as large optimum time lags, than do short-

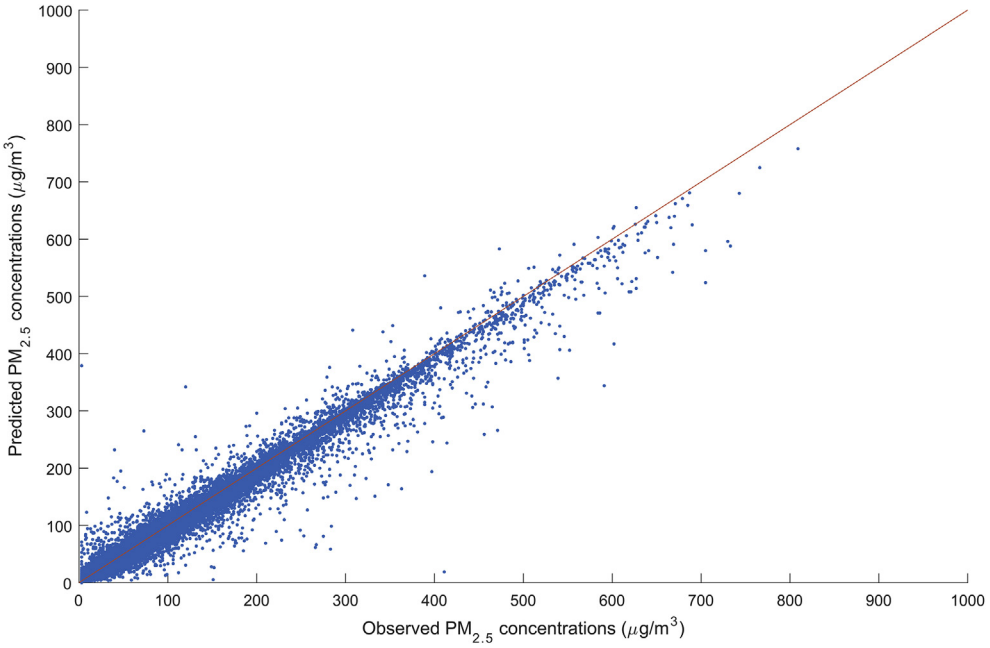

Fig. 4. Predicted and observed values of the test set.

Table 4
Comparison of the performances of the different methods.

Method	RMSE	MAE	MAPE (%)
LSTME	12.60	5.46	11.93
LSTM NN	17.94	7.81	15.84
STDL	14.96	9.00	21.75
TDNN	16.19	10.04	26.87
ARMA	24.40	13.05	27.54
SVR	22.04	11.14	28.45

Table 5
Structure and prediction accuracy of multiscale air pollutant concentration prediction.

Task	Time lag	RMSE	MAE	MAPE (%)
1-h prediction	8	12.60	5.46	11.93
2-h prediction	8	19.12	8.51	17.48
3-h prediction	10	18.58	8.61	17.53
4 to 6-h prediction	10	20.57	10.62	20.28
7 to 12-h prediction	16	33.96	16.88	29.73
13 to 24-h prediction	28	41.94	14.68	31.47

term prediction tasks. The proposed model exhibited satisfactory performance, even for the 13–24 h prediction task (MAPE = 31.47%).

4. Conclusions

This paper presents an LSTME model to predict air pollutant concentrations based on historical air pollutant concentration data, meteorological data, and time stamp data. The LSTME model is capable of modeling time series with long time dependencies and can automatically determine the optimum time lags. To evaluate the performance of our proposed model, hourly $PM_{2.5}$ concentrations in Beijing City were collected from 12 air quality monitoring stations. Six different models, including our LSTME, the traditional LSTM NN, the STDL, the TDNN, the ARMA and the SVR models, were compared using the same dataset. Experiments demonstrated that the proposed LSTM outperformed other algorithms, as indicated by the RMSE, MAE and MAPE values. Several useful findings can be concluded from this study.

1) Compared with traditional shallow models, such as the SVR, ARMA and TDNN models, deep learning-based models exhibited better prediction performance.

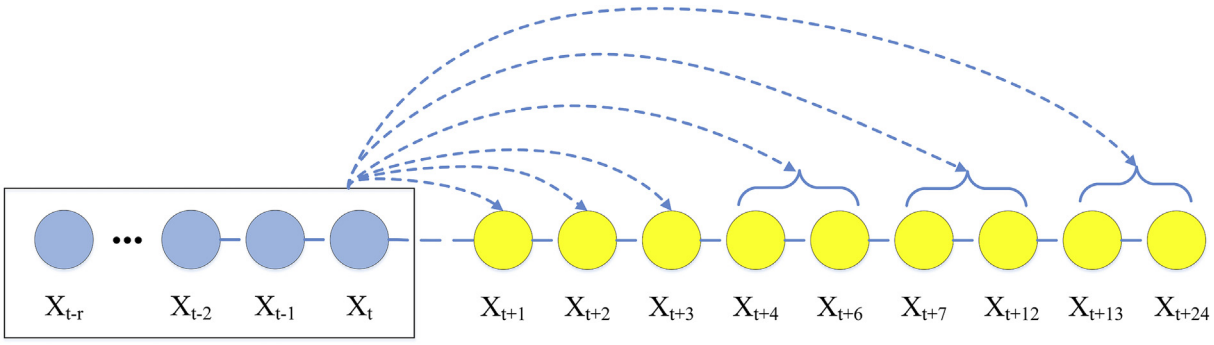

Fig. 5. Illustration of the multiscale predictors.

2) Compared with an RNN model, the TDNN model, our LSTME model and the traditional LSTM NN can more effectively capture spatiotemporal correlations, and they exhibited better prediction performance, as indicated by the RMSE, MAE and MAPE values.
3) The use of auxiliary data, such as meteorological data and time stamp data, can notably improve prediction performance.
4) Our model provides a multiscale method of predicting air pollutant concentrations. Although the prediction performance of long-term prediction tasks was reduced, the performance of the proposed model was suitable for long-term prediction tasks.

Acknowledgements

This research was financially supported by the National Science-technology Support Plan Project of China (2015BAJ02B00). The authors also thank the China Scholarship Council (CSC) (No. 201704910704) for supporting this work.

Appendix A. Related work

The LSTM NN is a special type of RNN that consists of one input layer, one output layer, and a series of recurrently connected hidden layers known as memory blocks. Each block is composed of one or more self-recurrent (connected to itself) memory cells and three multiplicative units (input, output and forget gates) that provide continuous analogues of read, write and reset operations for the cells. Fig. 1 gives an example of an LSTM memory block with a single cell. Each memory block has a recurrently self-connected linear unit-constant error carousel (CEC) at its core, and the activation of the CEC indicates the cell state. The self-recurrent memory cell can obstruct any outside interference; thus, the status can remain unchanged from one time step to another, which further allows the LSTM NN to solve the vanishing gradient problem. The forget gate was designed to learn to reset memory blocks once their status is out of date, thereby preventing the cell status from growing without bounds and causing saturation of the squashing function. Furthermore, the input gate enables incoming signals to modify the cell state, whereas the output gate permits or impedes the cell state from affecting other neurons.

In detail, the model input is given as $x = (x_1, x_2 \ldots, x_n)$, where each $x_i \in R^T$, $i = 1, 2, \ldots, n$; n denotes the number of input dimensions; T denotes the time lag; and the output sequence is given as $y = (y_1, y_2, \ldots, y_n)$. In our air pollutant concentration prediction tasks, x_i and y_i denote the input and predicted air pollutant concentrations, respectively, at station i. The forward training process of the LSTM NN can be formulated with the following equations:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$C_t = f_t * C_{t-1} + i_t * tanh(W_C \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4}$$

$$h_t = o_t * tanh(C_t) \tag{5}$$

where i_t, o_t, and f_t denote the activation of the input gate, output gate and forget gate, respectively; C_t and h_t denote the activation vector for each cell and memory block, respectively; and W and b denote the weight matrix and bias vector, respectively. In addition, $\sigma(\cdot)$ denotes the sigmoid function, which is defined in equation (6), and $tanh(\cdot)$ denotes the tanh function, which is defined in equation (7).

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{7}$$

Training the LSTM NN is typically based on truncated back propagation through time (BPTT) (Williams and Peng, 1990) and a customized version of real-time recurrent learning (RTRL) (Robinson and Fallside, 1987).

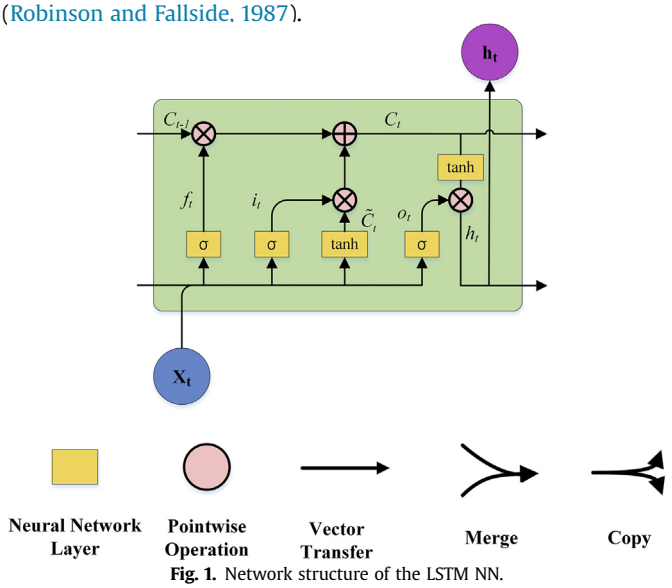

Fig. 1. Network structure of the LSTM NN.

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S., 2016. Social lstm: human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–971.

Antanasijević, D.Z., Pocajt, V.V., Povrenović, D.S., Ristić, M., Perić-Grujić, A.A., 2013. PM 10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci. Total Environ. 443, 511–519.

Bai, Y., Li, Y., Wang, X., Xie, J., Li, C., 2016. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmos. Pollut. Res. 7, 557–566.

Baklanov, A., Mestayer, P.G., Clappier, A., Zilitinkevich, S., Joffre, S., Mahura, A., Nielsen, N.W., 2008. Towards improving the simulation of meteorological fields in urban areas through updated/advanced surface fluxes description. Atmos. Chem. Phys. 8, 523–543.

Box, G.E.P., Jenkins, G.M., 1976. Time series analysis: forecasting and control. J. Operational Res. Soc. 22, 199–201.

Bravo, M.A., Bell, M.L., 2011. Spatial heterogeneity of PM10 and O3 in São Paulo, Brazil, and implications for human health studies. J. Air & Waste Manag. Assoc. 61, 69–77.

Bruckman, L., 1993. Overview of the enhanced geocoded emissions modeling and projection (Enhanced GEMAP) system. In: Proceeding of the Air & Waste Management Association's Regional Photochemical Measurements and Modeling Studies Conference, p. 562. San Diego, CA.

Chen, J., Lu, J., Avise, J.C., DaMassa, J.A., Kleeman, M.J., Kaduwela, A.P., 2014. Seasonal modeling of PM 2.5 in California's san Joaquin valley. Atmos. Environ. 92, 182–190.

Chen, Y., Shi, R., Shu, S., Gao, W., 2013. Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis. Atmos. Environ. 74, 346–359.

Coats Jr., C.J., 1996. High-performance algorithms in the sparse matrix operator kernel emissions (SMOKE) modeling system. In: Proc. Ninth AMS Joint Conference on Applications of Air Pollution Meteorology with A&WMA, Amer. Meteor. Soc. Citeseer, Atlanta, GA, pp. 584–588.

Díaz-Robles, L.A., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G., Moncada-Herrera, J.A., 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos. Environ. 42, 8331–8340.

Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, Jr., Benjamin, G., Speizer, F.E., 1993. An association between air pollution and

mortality in six U.S. cities. N. Engl. J. Med. 329, 1753–1759.
Felder, M., Kaifel, A., Graves, A., 2010. Wind power prediction using mixture density recurrent neural networks. In: Poster Presentation gehalten auf der European Wind Energy Conference.
Feng, Y., Zhang, W., Sun, D., Zhang, L., 2011. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. Atmos. Environ. 45, 1979–1985.
Guocai, Z., 2004. Progress of weather research and forecast (WRF) model and application in the United States. Meteorol. Mon. 12, 5.
He, J., Yu, Y., Liu, N., Zhao, S., 2013. Numerical model-based relationship between meteorological conditions and air quality and its implication for urban air quality management. Int. J. Environ. Pollut. 53, 265–286.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.
Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., 2005. A neural network forecast for daily average PM 10 concentrations in Belgium. Atmos. Environ. 39, 3279–3289.
Iii, C.A.P., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama J. Am. Med. Assoc. 287, 1132–1141.
Jeong, J.I., Park, R.J., Woo, J., Han, Y., Yi, S., 2011. Source contributions to carbonaceous aerosol concentrations in Korea. Atmos. Environ. 45, 1116–1125.
Kappos, A.D., Bruckmann, P., Eikmann, T., Englert, N., Heinrich, U., Höppe, P., Koch, E., Krause, G.H., Kreyling, W.G., Rauchfuss, K., 2004. Health effects of particles in ambient air. Int. J. Hyg. Environ. Health 207, 399–407.
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 model. Atmos. Environ. 44, 523–532.
Kolehmainen, M., Martikainen, H., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. Atmos. Environ. 35, 815–825.
Krewski, D., Jerrett, M., Burnett, R.T., Ma, R., Hughes, E., Shi, Y., Turner, M.C., Pope, C.A., Thurston, G., Calle, E.E., Thun, M.J., Beckerman, B., DeLuca, P., Finkelstein, N., Ito, K., Moore, D.K., Newbold, K.B., Ramsay, T., Ross, Z., Shin, H., Tempalski, B., 2009. Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. Res. Rep. Health Eff. Inst. 140, 5.
Kurt, A., Oktay, A.B., 2010. Forecasting air pollutant indicator levels with geographic models 3days in advance using neural networks. Expert Syst. Appl. 37, 7986–7992.
Li, C., Hsu, N.C., Tsay, S., 2011. A study on the potential applications of satellite data in air quality monitoring and forecasting. Atmos. Environ. 45, 3663–3675.
Li, X., Peng, L., Hu, Y., Shao, J., Chi, T., 2016. Deep learning architecture for air quality predictions. Environ. Sci. Pollut. Res. 23, 22408–22417.
Lu, W.Z., Wang, W.J., Fan, H.Y., Leung, A., Xu, Z.B., Lo, S.M., Wong, J., 2002. Prediction of pollutant levels in causeway bay area of Hong Kong using an improved neural network model. J. Environ. Eng.-ASCE 128, 1146–1157.
Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F., 2015. Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transport. Syst. 16, 1–9.
Lurmann, F.W., 2000. Simplification of the UAMAERO Model for Seasonal and Annual Modeling: the UAMAERO-LT Model (Final Report).
Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp. Res. Part C Emerg. Technol. 54, 187–197.
Mishra, D., Goyal, P., 2016. Neuro-fuzzy approach to forecast NO2 pollutants addressed to air quality dispersion model over Delhi, India. Aerosol Air Qual. Res. 16, 166–174.
Neuberger, M., Schimek, M.G., Horak, F., Moshammer, H., Kundi, M., Frischer, T., Gomiscek, B., Puxbaum, H., Hauck, H., 2004. Acute effects of particulate matter on respiratory diseases, symptoms and functions:: epidemiological results of the Austrian Project on Health Effects of Particulate Matter (AUPHEP). Atmos. Environ. 38, 3971–3981.
Nieto, P.G., Combarro, E.F., Del Coz Díaz, J.J., Montañés, E., 2013. A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): a case study. Appl. Math. Comput. 219, 8923–8937.
Ong, B.T., Sugiura, K., Zettsu, K., 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. Neural Comput. Appl. 27, 1553–1566.
Paschalidou, A.K., Karakitsios, S., Kleanthous, S., Kassomenos, P.A., 2011. Forecasting hourly PM10 concentration in Cyprus through artificial neural networks and multiple regression models: implications to local environmental management. Environ. Sci. Pollut. Res. 18, 316–327.
Pearson, K., 1895. Note on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. 58, 240–242.
Prakash, A., Kumar, U., Kumar, K., Jain, V.K., 2011. A wavelet-based neural network model to predict ambient air pollutants' concentration. Environ. Model. Assess. 16, 503–517.
Robinson, A.J., Fallside, F., 1987. The Utility Driven Dynamic Error Propagation Network. University of Cambridge Department of Engineering.
Saide, P.E., Carmichael, G.R., Spak, S.N., Gallardo, L., Osses, A.E., Mena-Carrasco, M.A., Pagowski, M., 2011. Forecasting urban PM10 and PM2. 5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model. Atmos. Environ. 45, 2769–2780.
Sak, H., Yang, G., Li, B., Li, W., 2016. Modeling Dependence Dynamics of Air Pollution: Pollution Risk Simulation and Prediction of PM $ _ {2.5} $ Levels. arXiv preprint arXiv:1602.05349.
Stern, R., Builtjes, P., Schaap, M., Timmermans, R., Vautard, R., Hodzic, A., Memmesheimer, M., Feldmann, H., Renner, E., Wolke, R., 2008. A model intercomparison study focussing on episodes with elevated PM10 concentrations. Atmos. Environ. 42, 4567–4588.
Vautard, R., Builtjes, P., Thunis, P., Cuvelier, C., Bedogni, M., Bessagnet, B., Honore, C., Moussiopoulos, N., Pirovano, G., Schaap, M., 2007. Evaluation and intercomparison of Ozone and PM10 simulations by several chemistry transport models over four European cities within the CityDelta project. Atmos. Environ. 41, 173–188.
Wang, Z., Maeda, T., Hayashi, M., Hsiao, L., Liu, K., 2001. A nested air quality prediction modeling system for urban and regional scales: application for high-ozone episode in Taiwan. Water, Air, Soil Pollut. 130, 391–396.
Williams, R.J., Peng, J., 1990. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. Neural Comput. 2, 490–501.
Wilson, J.G., Kingham, S., Pearce, J., Sturman, A.P., 2005. A review of intraurban variations in particulate air pollution: implications for epidemiological research. Atmos. Environ. 39, 6444–6462.
Yoon, H., Jun, S.C., Hyun, Y., Bae, G.O., Lee, K.K., 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J. Hydrol. 396, 128–138.
Zheng, Y., Liu, F., Hsieh, H., 2013. U-Air: when urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1436–1444.