Science of the Total Environment 654 (2019) 1091–1099

Contents lists available at ScienceDirect

Science of the Total Environment

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

A novel spatiotemporal convolutional long short-term neural network for air pollution prediction

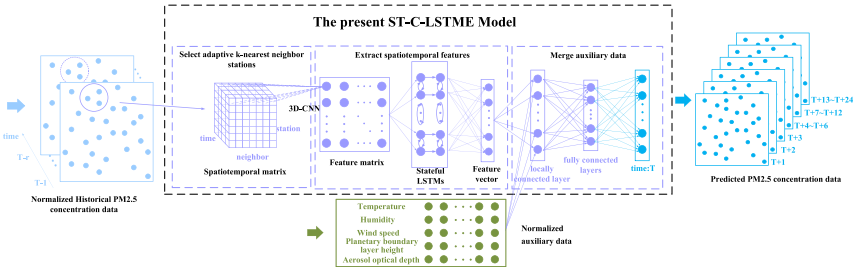
Congcong Wen [a,b], Shufu Liu [a,*], Xiaojing Yao [a], Ling Peng [a], Xiang Li [a,b], Yuan Hu [a,b], Tianhe Chi [a]

[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

HIGHLIGHTS

- PM2.5 prediction of all monitoring stations in China from 1 h to 24 h was realized.
- The addition of information from neighboring stations improves the prediction accuracy of present station.
- Aerosol data first introduced a deep learning model for PM2.5 prediction.
- The present model achieves more accurate and stable air quality prediction of different spatiotemporal scales.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 August 2018
Received in revised form 4 November 2018
Accepted 7 November 2018
Available online 9 November 2018

Editor: Jianmin Chen

Keywords:
Air pollutant concentration predictions
Spatiotemporal correlation
3D convolutional neural network (3D-CNN)
Long short-term memory neural network (LSTM NN)
Long-term prediction

ABSTRACT

Air pollution is a serious environmental problem that has drawn worldwide attention. Predicting air pollution in advance has great significance on people's daily health control and government decision-making. However, existing research methods have failed to effectively extract the spatiotemporal features of air pollutant concentration data, and exhibited low precision in long-term predictions and sudden changes in air quality. In the present study, a spatiotemporal convolutional long short-term memory neural network extended (C-LSTME) model for predicting air quality concentration was proposed. In order to encompass the spatiality and temporality of the data, the model involved the historical air pollutant concentration of the present station, as well as that of the adaptive k-nearest neighboring stations, into the model. High-level spatiotemporal features were extracted through the combination of the convolutional neural network (CNN) and long short-term memory neural network (LSTM-NN), and meteorological data and aerosol data were also integrated, in order to improve model prediction performance. Hourly PM2.5 (particulate matter with an aerodynamic diameter of ≤2.5 mm) concentration data collected at 1233 air quality monitoring stations in Beijing and the whole China from January 1, 2016 to December 31, 2017 were used to validate the effectiveness of the proposed C-LSTME model. The results show that the present model has achieved better performance than current state-of-the-art models for different time predictions at different regional scales.

1. Introduction

As a serious environmental problem, air pollution has gradually attracted worldwide attention (Kurt and Oktay, 2010). Many countries around the world are suffering from heavy air pollution. For example, the atmosphere visibility in most areas in China has shown a significant decline since the 1990s (Che et al., 2007; Zhao and Zhang, 2011; J. Deng et al., 2012; Cheng et al., 2013) and the number of hazy days every year has enormously increased (X. Deng et al., 2008; Y. Chen et al., 2013). PM2.5 (particulate matter with an aerodynamic diameter of ≤2.5 mm) is the main cause of haze. Therefore, precisely predicting its mass

* Corresponding author.
E-mail address: liusf01@radi.ac.cn (S. Liu).

https://doi.org/10.1016/j.scitotenv.2018.11.086
0048-9697/

concentration has a pivotal role in making atmospheric management decisions (Y. Zhang and Li, 2015). In addition, a large number of epidemiological studies have demonstrated that PM2.5 can cause negative health effects, such as excessive morbidity and mortality from cardiovascular and respiratory diseases (Pope et al., 2002; Dominici et al., 2006). Therefore, predicting air pollutant concentration in advance is fundamental to strengthen air pollution prevention and achieve comprehensive environmental management, which is of great significance to people's daily health and government decision-making (Zheng et al., 2015).

The existing research methods for air prediction are mainly divided into two methods: deterministic methods and statistical methods. The deterministic method is based on the aerodynamic theory and physicochemical processes, and uses mathematical methods to establish a numerical model of the dilution and diffusion of atmospheric pollution concentration. Through high-speed calculation and simulation, the dynamic changes of the atmospheric pollutant concentration are predicted (Coats, 1996; G. Zhang, 2004; Jeong et al., 2011; Woody et al., 2016; Bray et al., 2017; Zhou et al., 2017). The commonly used models are the Community Multi-scale Air Quality (CMAQ) model (J. Chen et al., 2014), the Nested Air Quality Prediction Modeling System (NAQPMS) (Z. Wang et al., 2001), and the WRFChem model (Saide et al., 2011). However, these models require very informative source data that are difficult to obtain in practice (Vautard et al., 2007; Stern et al., 2008). In addition, various types of parameters that need to be determined, and in many cases, these are estimated by experience, resulting in limited accuracy (Pan et al., 2011; Ridder et al., 2012; T. Wang et al., 2012; Xu et al., 2017).

Nowadays, statistical prediction methods have received increasing attention by virtue of the obvious advantages (Neto et al., 2014; Donnelly et al., 2015; Prasad et al., 2016; D. Wang et al., 2016). Commonly used statistical prediction methods include the multiple linear regression (MLR) method (C. Li et al., 2011), the autoregressive moving average (ARMA) method (Box and Jenkins, 1976), the support vector regression (SVR) method (Combarro, 2013), the artificial neural network (ANN) method (Hooyberghs et al., 2005), and hybrid methods (Díaz-Robles et al., 2008; Y. Chen et al., 2013). However, these models fail to effectively integrate and analyze multi-source heterogeneous data, such as traffic flow, meteorological conditions, and land use, which has considerable impact on air quality (Vardoulakis et al., 2003; Zheng et al., 2013).Besides, Zheng et al. (2015) incorporated meteorological data and weather forecast into the model, but instead of using a unified model to extract features, he divided the model into four parts: temporal predictor, spatial predictor, prediction aggregator, and infection predictor.

Deep learning, which is a novel machine learning method proposed in the field of artificial intelligence in recent years, can learn effective feature representation from a large amount of input data, thereby providing new ideas for solving the problems mentioned in the last two paragraphs (Hinton and Salakhutdinov, 2006; Yoshua Bengio, 2009). The Recurrent Neural Network (RNN) (Feng et al., 2011), Elman Neural Network (Prakash et al., 2011), Time Delay Neural Network (TDNN) (Ong et al., 2016), Geographical Deep Belief Network (Geoi-DBN) (T. Li et al., 2017) and other deep learning models have been applied to air pollution predictions in previous studies. However, these models only take either the temporal or spatial correlations of the air pollution concentration into account. X. Li et al. (2017) proposed the long short-term memory neural network extended (LSTME) model, which is based on the long short-term memory (LSTM) model and integrates meteorological data as auxiliary data at the same time to extract the spatiotemporal correlations of air quality. Nonetheless, they input data from all neighboring stations into the model, conducing to the fact that the interference from unrelated stations caused negative impacts on the accuracy of the model. Soh et al. (2018) proposed the spatial-temporal deep neural network (ST-DNN) model which integrates the long short term memory (LSTM) model to extract temporal features, involves k-nearest neighbor (KNN) and artificial neural network (ANN) to extract spatial features, and combines convolutional neural network (CNN) to extract terrain features. However, the ST-DNN model extracts temporal and spatial features separately, which undermines the inherent regularity of the data. In addition, PM2.5 is more relevant to aerosol data compared to meteorological data and other data (Chu et al., 2003; Wang and Christopher, 2003; Engel-Cox et al., 2004; Gupta et al., 2006), but none of the above models have considered aerosol data.

To remove these restrictions, the present study proposes a novel spatiotemporal convolutional long short-term memory neural network for air quality prediction. In order to take into account the spatiotemporal correlations of the air pollution concentration, the historical concentration of the present station and the adaptive k-nearest neighboring stations with high correlation were entered into the model. The convolutional neural network (CNN) and long short-term memory neural network (LSTM-NN) were used to extract the spatiotemporal features, in which the 3D-CNN that can extract high-level spatiotemporal features and stateful LSTM-NN that can maintain the state information for a long time were selected, achieving a more stable long-term prediction. In addition, the auxiliary data, including meteorological data and aerosol optical depth data, were added into the model. A lot of previous works have proven that these data have a strong correlation with PM2.5 (Chu et al., 2003; J. Wang and Christopher, 2003; Engel-Cox et al., 2004; G. Zhang, 2004; P. Gupta et al., 2006; Koelemeijer et al., 2006; Saide et al., 2011). On the other hand, since these data can reflect weather conditions, these auxiliary data were added to improve the model prediction accuracy for sudden changes in air pollution concentration. The experiment results presented in Section 3.2 demonstrate this finding.

The main contributions of the present study are as follows: (1) The neighboring distribution of each station was fully considered, which means that k-nearest neighboring stations with high correlation were adaptively selected for each station. (2) The spatio-temporality of air pollutant concentration data was taken into consideration as a whole and was processed by the present unified model. (3) The model integrated auxiliary data, including meteorological data and aerosol optical depth data. To a certain extent, auxiliary data can improve prediction accuracy, and at the same time, help the model to better predict the sudden changes in air quality. (4) The proposed method can efficiently extract better spatiotemporal correlation features and achieve high accuracy and stability for air quality prediction of different spatiotemporal scales.

2. Data and methods

2.1. Data description

Hourly PM2.5 concentration data from 1233 air quality monitoring stations in China collected from January 1, 2016 to December 31, 2017 were acquired from the Ministry of Environmental Protection of China (http://datacenter.mep.gov.cn/). 12 monitoring stations in Beijing district at the same time period were selected from the China dataset as Beijing dataset. Fig. 1 shows the distribution of China's air quality stations and the grading of PM2.5 values for each station on January 1, 2016. Auxiliary data, including meteorological data (hourly humidity, temperature and wind speed), planetary boundary layer height (PBLH), and aerosol optical depth (AOD) data for the same period, were selected. These have been previously proven to be highly correlated to PM2.5 (Chu et al., 2003; Engel-Cox et al., 2004; P. Gupta et al., 2006; Koelemeijer et al., 2006; Díaz-Robles et al., 2008; Saide et al., 2011). These data were downloaded from MERRA2 data (https://goldsmr4.gesdisc.eosdis.nasa.gov/data/MERRA2), which is the best integration of various types of observations and short-term forecast products issued by the National Aeronautics and Space Administration (NASA). It is a reanalysis data that is presently and widely used, and

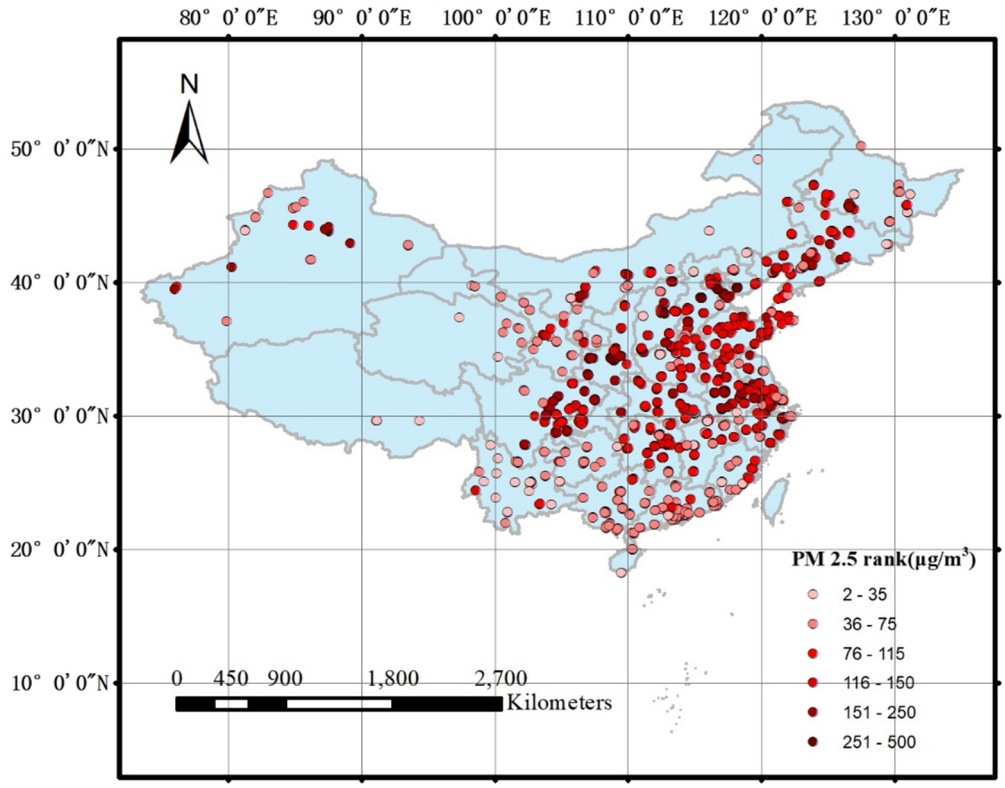


Fig. 1. Distribution of air quality monitoring stations in China. (The color of each station represents the rank of daily average PM2.5 on January 1, 2016, as described in the bottom right of the figure.)

has high precision (Trenberth and Olson, 1988; Rienecker et al., 2011; Jian and Jiang, 2015; Koster et al., 2015). The present dataset contained 17,544 records for each station. In the present experiment, 60% of the data were randomly selected as the training set, 20% as the validation set, and the remaining 20% as the test set.

2.2. Spatiotemporal analysis of data

According to Tobler's First Law of Geography (Tobler, 1970), near things have higher correlation with each other than distant things. In order to illustrate the spatiality of the PM2.5 distribution, the Euclidean distance and Pearson correlation coefficient (Pearson, 2006) between each two stations were calculated. The calculation formulas are as follows:

$$D(s_i, s_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (1)$$

$$r(s_i, s_j) = \frac{Cov(s_i, s_j)}{\sigma(s_i)\sigma(s_j)} \quad (2)$$

where $D(\cdot)$ and $r(\cdot)$ denote the distance and correlation coefficient between stations, respectively; x, y denote the latitude and longitude of the corresponding station; $Cov(\cdot)$ is the covariance and $\sigma(\cdot)$ is the standard deviation.

Fig. 2 shows the correlation coefficients of 100 randomly selected stations and their nearest 20 stations. It can be observed that all correlation values were above 0.56 (P-value < 0.01) and the correlation coefficient decreases with the increase in distance in the overall trend. Therefore, neighboring stations can be used to improve the prediction accuracy of the present station. However, for each station, the correlation values with their neighboring stations are different. In order to ensure that the neighboring stations are as relevant as possible, the number of neighboring stations selected for each station should be different. As to the temporality of the PM2.5 distribution, previous studies have pointed out that the present moment of the station has a good correlation with a certain moment in the past (X. Li et al., 2017).

To reflect the spatiotemporal correlation between stations, adaptive k-nearest neighboring stations were selected for each station, and the delayed PM2.5 concentration were entered into the model. Then, the convolutional neural network and long short-term memory neural network were applied to extract its spatiotemporal correlations.

2.3. C-LSTME model

Fig. 3 shows the overall prediction framework. The model inputs comprised of two parts: the time-delayed PM2.5 concentrations from all stations and their neighboring stations ("Light blue arrow" in

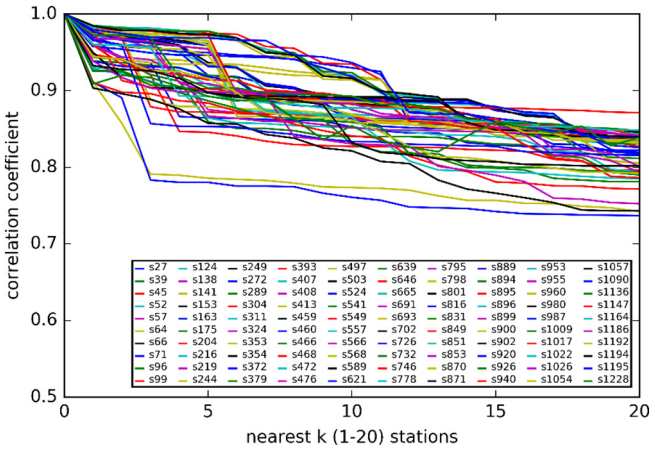


Fig. 2. Pearson's correlation coefficient between 100 randomly selected stations and their nearest 20 stations, where all correlation values were above 0.56 (P-value < 0.01).

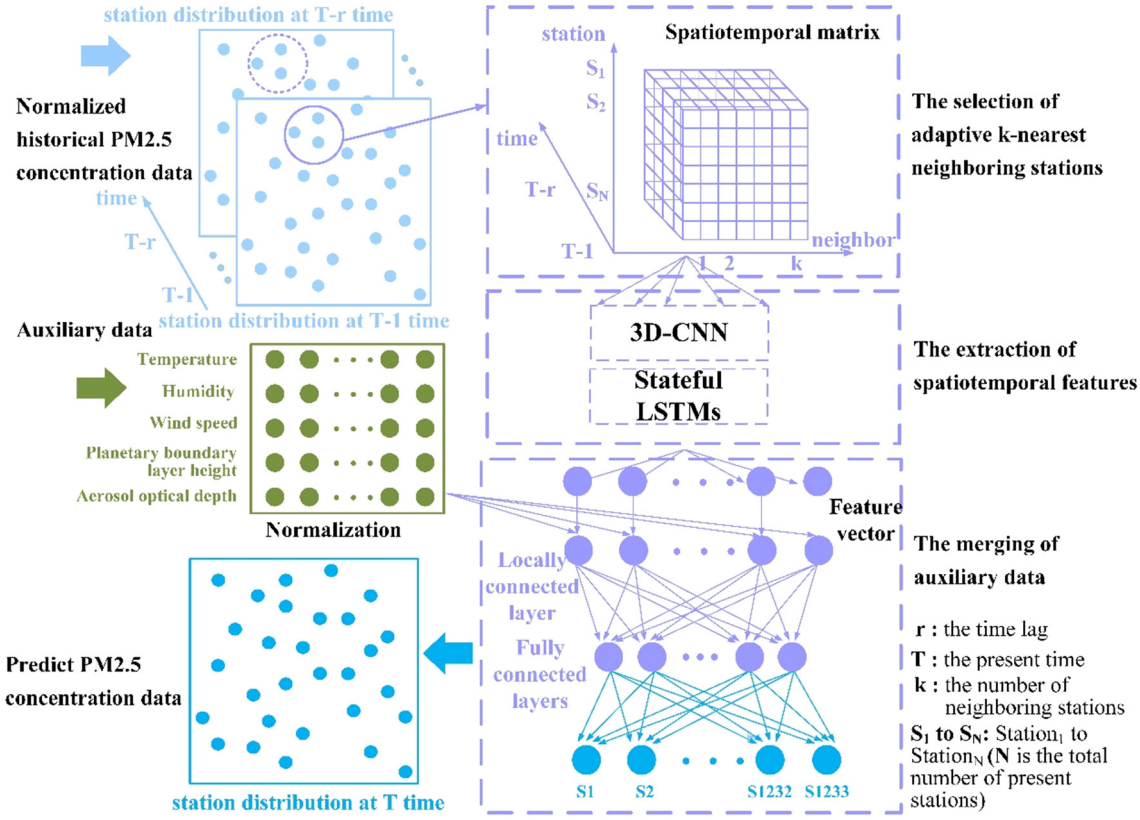


Fig. 3. Network framework of the C-LSTME model for air pollutant concentration prediction. Model inputs are pointed by light blue arrow (historical PM2.5 concentration data) and green arrow (auxiliary data), and model outputs are indicated by the blue arrow (predict PM2.5 concentration data). The variable r indicates the time lag, the variable T indicates the present time, the variable k is the number of neighboring stations and the variable N is the total number of present stations.

Fig. 3); the auxiliary data (humidity, temperature, wind speed, planetary boundary layer height, and AOD data) of all stations ("Green arrow" in Fig. 3). In addition, the model outputs were the PM2.5 prediction values of all stations ("Blue arrow" in Fig. 3). The model can be divided into three parts: the selection of adaptive k-nearest neighboring stations, the extraction of spatiotemporal features, and the merging of auxiliary data.

First, the adaptive k-nearest neighboring stations for each station were selected. Prior research on spatiality has suggested that the spatial correlation between a certain station and its neighboring station are different. Therefore, to select more related stations as the neighboring stations for present model, a correlation threshold value was set, and k-nearest neighboring stations with correlation values higher than the threshold value from near to distant were searched. The adaptive method was used to select the neighboring stations. That is, the number of neighboring stations, which is represented by k_i, is different for each station. It was assumed that the lower quartile of all the k_i is k_{Q1}, signifying that 75% of k_i is greater than k_{Q1}. To ensure the same dimension of all the input data, we used same number of neighboring stations for every station: when $k_i < k_{Q1}$, the $k_{Q1} - k_i$ present station values were added as supplements; when $k_i > k_{Q1}$, the values of k_{Q1} neighboring stations were used, which are most relevant to the present station. The historical values at r time points before time T were selected, and the adaptive k-nearest neighboring stations were searched according to the threshold value of the correlation. As a result, the r*k*N spatiotemporal matrix was obtained (see "Spatiotemporal matrix" in Fig. 3), where r is the time lag, k is the number of neighboring stations, and N is the total number of stations.

Second, the CNN and LSTM-NN model were used to extract the spatiotemporal features of air pollutant concentration data from the previous spatiotemporal matrix mentioned above. The detailed introduction to CNN and LSTM are provided in Appendices A and B. In the present, 3D-CNN was adopted, which performs better in extracting more in-depth spatiotemporal features. In addition, stateful LSTM was applied, which uses the state of each batch of LSTM samples as the initial state of the next batch of samples. The state of a memory cell is passed from the previous batch to the next batch, which is suitable for the prediction of some long sequences (S. Gupta and Dinesh, 2017). The spatiotemporal matrix was processed by 3D-CNN and stateful LSTMs, and the N * 1 feature vector was obtained, providing a more efficient representation of the spatiotemporal features (see "The extraction of spatiotemporal features" in Fig. 3).

Finally, the auxiliary data was added to improve the prediction performance of the model. A locally connected layer was used to integrate the spatiotemporal eigenvectors of the air pollution values as obtained above and the normalized auxiliary data (see "The merging of auxiliary data" in Fig. 3). After one or more fully connected layers, the PM2.5 value of the station at time T was obtained.

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 the present experiments. These indicators can be formulated, as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \tag{3}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \tag{4}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|P_i - O_i|}{O_i} \quad (5)$$

where O_i is the observed air pollutant concentration, P_i is the predicted air pollutant concentration, and n is the number of test samples.

2.4. Network architecture

Several hyper parameters should be preset before building the C-LSTME prediction architecture (shown in Fig. 3), including the number of CNN layers, 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 correlation threshold, and the time lags. The effect of each parameter was examined while keeping the other parameters fixed, and a random search method with a 5-fold cross validation was employed to determine the optimum hyper parameters.

Through a series of contrast experiments, the basic structure of the model was first determined. For the CNN, a two-layer structure was used. The first layer was a 3D convolutional layer with the k*1*r convolution kernel. The second layer was the activation layer where Rectified-Linear Unit (ReLU) was used as the activation function. In addition, as to LSTM, two LSTM layers with 1000 nodes for each layer and one fully connected layer with 600 nodes for this layer were used. The above structure configuration achieved the best prediction performance in this experiment.

Next, RSME, MAE, and MAPE were used as indicators to determine the impact of different correlation thresholds on prediction accuracy. Table 1 shows the predicted performance results obtained using different correlation thresholds when the time lag r was set to 6. It was found that excessively high or low correlation thresholds would induce a decline in the predicted performance, which is mainly because when the threshold is high, the number of neighboring stations will be small, and when the threshold is low, some of the less high related stations would be entered, causing interference. When the correlation threshold was 0.9, the model yielded better prediction performance. At this time, for >75% of the stations, the number of the neighboring stations was greater than four. That is, the value of k is 4. Hence, the model's correlation threshold was set to 0.9, which was the most appropriate setting for the proposed model.

After determining the correlation threshold, the effects of different time lags were investigated. The correlation threshold was fixed to 0.9 based on the above results, and the prediction performance of the model was obtained at different time lags (shown in Table 1). According to previous studies, a small time lag cannot guarantee enough long-term memory inputs for the model, and large time lags permit an increased number of unrelated inputs, which is consistent with the present experimental results. Therefore, the time lag was finally set to 8.

Table 1
Effect of the correlation threshold and time lag on the present model, where correlation refers to the Pearson correlation coefficient between stations, time lag refers to the number of historical moments that were selected, RMSE is the root mean square error, MAE is the mean absolute error, and MAPE is the mean absolute percentage error.

Parameters	Parameters value	RMSE (μg/m^3)	MAE (μg/m^3)	MAPE (%)
Correlation threshold	0.8	13.61	4.70	15.98
	0.85	12.25	4.52	15.85
	0.9	**11.83**	**4.46**	**15.79**
	0.95	12.25	4.52	15.87
	1	13.61	5.16	16.13
Time lag	2	11.93	4.54	16.22
	4	12.07	4.59	16.01
	6	11.83	4.46	15.79
	8	**11.48**	**4.38**	**15.29**
	10	11.77	4.41	15.80

Bold emphasis indicates the best model parameters value and the smallest RSME, MAE, MAPE.

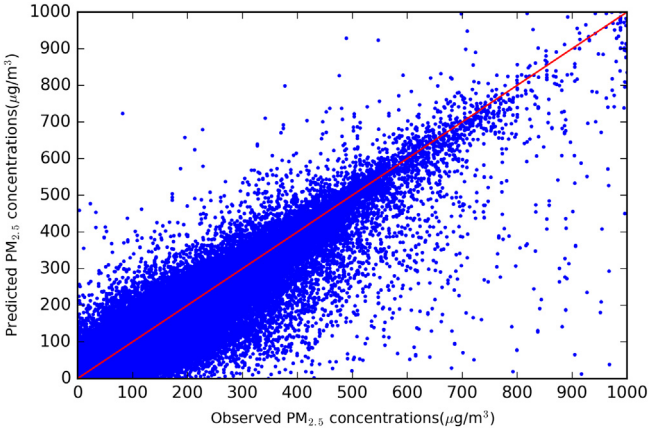


Fig. 4. Predicted and observed values of the test set.

3. Results and discussion

3.1. Prediction performance

After determining the best network architecture for the present prediction task, the training set was utilized to train the present C-LSTME model until convergence. The evaluations were conducted using the test set, and the predicted and observed PM2.5 concentrations are presented in Fig. 4. It can be observed from the figure that the predicted value was generally consistent with the observed value. The R^2 value between the observed and predicted data demonstrated that the model can capture 92% of the explained variance.

Furthermore, the prediction performance of the proposed model was evaluated after dividing the air quality into six ranks, according to the National Technical Regulation of the Ambient Air Quality Index. In addition, the predicted and observed rankings were generated, and the overall ranking accuracy was calculated. The overall rank prediction accuracy was 87.6%, indicating satisfactory performance in rank prediction. In addition, merely 7.5% of the test samples received an overestimated predicted ranking, while merely 4.9% of the test samples received an underestimated predicted ranking.

3.2. Comparison of experiments

In order to test the performance of the present model on different spatial scales and temporal scales, we compared the prediction results of it with other state-of-the-art models for different future hours using Beijing dataset and the whole China dataset respectively.

Table 2
Comparison of the mean absolute error (MAE) of different models for the next 1st to the 6th hour predictions.

MAE (μg/m3)	1st hour	2nd hour	3rd hour	4th hour	5th hour	6th hour
C-LSTME	**5.77**	**8.85**	**10.31**	**11.60**	**12.25**	**14.00**
LSTME	7.01	9.37	10.44	12.65	15.94	17.96
ST-DNN	7.36	10.83	11.15	14.17	15.35	16.43
LSTM	7.86	10.05	11.59	14.49	15.82	16.22
Zheng	12.73	14.67	18.27	22.94	26.25	28.37
ARMA	18.54	19.47	21.85	23.64	25.81	28.51
SVR	26.79	30.77	33.69	35.74	37.37	38.21
LR	34.27	35.59	37.15	38.41	39.68	40.25

Bold emphasis indicates the best model and smallest MAEs for 1st to 6th predictions.

Table 3
Comparison of the performance of different methods using three indicators: the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE).

Method	RSME (μg/m^3)	MAE (μg/m^3)	MAPE (%)
C-LSTME	**12.08**	**5.82**	**17.09**
ST-C-LSTM	17.76	8.02	21.93
LSTME	18.25	8.48	22.82
LSTME (AOD)	21.17	10.57	25.38
LSTME (meteorological)	22.22	11.09	29.93
LSTM NN	25.95	12.48	30.89
ARMA	34.40	28.05	37.54
SVR	39.92	29.98	37.23

Bold emphasis indicates the best model (C-LSTME should be bold) and smallest RSME, MAE, MAPE.

3.2.1. Beijing dataset

For the Beijing dataset, we carried out the comparison experiment using four non-deep learning baseline models, including the linear regression (LR) model, the support vector regression (SVR) model, the autoregressive moving average (ARMA) model, and the Zheng model (Zheng et al., 2015), and four deep learning baseline models, including the LSTM model, the STDNN model, the LSTME model, and the present model. Mean absolute error (MAE) is used as an evaluation indicator for accuracy, and Table 2 shows MAE for the above eight models in predicting air quality from the 1st hour to 6th hour.

3.2.2. China dataset

In order to further reflect the superior performance of the present model for large-area air quality prediction, data from 1233 station in China were input into the model at the same time. Due to data limitations, we use the following baseline models:

Non-deep learning models: the support vector regression (SVR) model; the autoregressive moving average (ARMA) model.

Deep learning models: the LSTM model; the LSTME model using three different kinds of auxiliary data, including LSTME model using only meteorological data, LSTME model using only AOD data, and LSTME model using both meteorological data and AOD data; and the C-LSTM model which didn't involve any auxiliary data but integrated CNN to extract high-related neighboring stations features.

Based on the performance evaluation indicators including RMSE, MAE, and MAPE, Table 3 shows the prediction performance for the next 1st hour of present model and the above seven baseline models using the China dataset. Fig. 5 presents the 3500-hour prediction results of one randomly selected station using the present C-LSTME model and C-LSTM model which didn't integrate any auxiliary data.

For the next 2nd hour to 24th hour, we divided them into five groups (the 2nd hour, the 3rd hour, the 4th–6th hour, the 7th–12th hour, and the 13th–24th hour) and trained separate models to predict the average air quality for each interval. The previously well-performing LSTME model was used as a baseline model, and we calculated the RMSE, MAE, MAPE, and the standard deviation of the corresponding errors for the LSTME model and the present model, respectively.

The Fig. 6 presents the basic framework of the model for long-term prediction. The adaptive method was used with the above-mentioned correlation threshold to select the neighboring stations. The time-delayed PM2.5 concentrations of all stations, as well as the neighboring stations, were used as the model input. Then, 3D-CNN and stateful LSTM were used to extract the spatiotemporal features of the data. The basic structure of the model was kept unchanged, and the time lags to predict the different time points were adjusted. The optimal time lag for the different time predictions was determined through a series of comparative experiments, as shown in Table 4. The best prediction results of the LSTME model with the same time interval as the C-LSTME model are also presented in Table 4.

3.3. Discussion

Through using different regional datasets, the comparison between the performance of present model and other baseline models on predicting PM2.5 concentration at different future hours, we can get the following conclusions:

First, as shown in Tables 2 and 3, the deep learning methods have less error than the non-deep learning methods for any regional scale. Although the errors of all models increase to different extent as the regional scale increases, the deep learning methods are more adaptable than non-deep learning methods.

Second, as we can see in Table 2, our model has the smallest MAE compared to other three models including Zheng, STDNN and LSTME models. It's mainly because that compared to LSTME model, we consider the influence of high-related surrounding stations; compared to Zheng and STDNN models, we use one unified model to extract spatial-temporal features instead of extracting temporal and spatial features separately through different models.

Third, from Tables 2 and 3, it can be found that C-LSTM model and LSTME model have smaller errors than the LSTM model on both datasets. It can be mainly attributed to that LSTME model integrates auxiliary data compared to LSTM model, and C-LSTM model takes features of neighboring stations into consideration

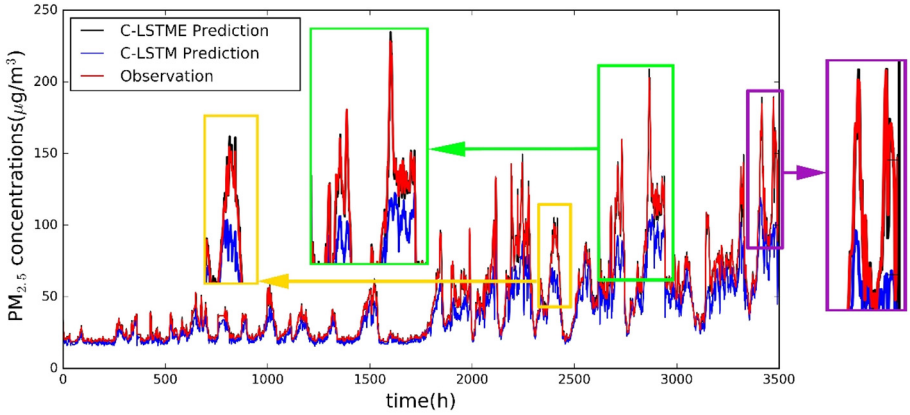


Fig. 5. Comparison of C-LSTM model and C-LSTME model prediction results. The boxes in the figure represent the partial enlargement for details of abrupt changes in air pollution.

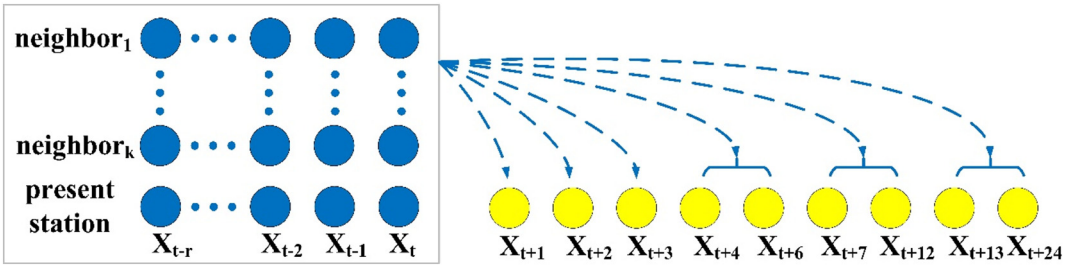


Fig. 6. Basic framework of the model for long-term prediction, where neighbor$_1$ to neighbor$_k$ refer to the k nearest neighboring stations, and X_{t-r} to X_{t+24} refer to PM2.5 concentration data at time t − r to t + 24 respectively (r refers to the time lag).

compared to LSTM model. And C-LSTME model proposed in this paper combines C-LSTM model and LSTME model to achieve the best results.

Fourth, as demonstrated in Tables 2 and 3, by comparing LSTME model with LSTM model, and C-LSTME with C-LSTM, it can be concluded that adding auxiliary data contributes to the improvement of model performance. Furthermore, it can be observed from Fig. 5 that the addition of the auxiliary data can help identify and predict the abrupt changes in air quality. Through the comparison among three LSTME models with different auxiliary data in Table 3, we can see that AOD is more effective than meteorological data in improving the error accuracy of the model.

Fifth, as illustrated in Table 4, by comparing the prediction accuracy of the LSTME model and C-LSTME model on long-term prediction, it can be found that the accuracy of both models decreases as the prediction time extends. However, the RMSE standard deviation of the C-LSTME model (3.96) is much lower than that of LSTME (10.44), demonstrating that C-LSTME achieves higher accuracy and stability on long-term prediction. This can be mainly attributed to the combination of 3D-CNN and stateful LSTM, which makes it possible to extract high-level spatiotemporal features, and maintain the transfer of state information at the same time.

4. Conclusion

The present study proposes a spatiotemporal convolutional long short-term memory neural network extended (C-LSTME) model for air quality prediction. Through the integration of the PM2.5 data obtained from high-related neighboring stations into the model, and the use of historical data of PM2.5 concentration, meteorological data and aerosol data which is more related to PM2.5 but was neglected by other models, the accurate and stable prediction was realized. Furthermore, through comparison experiments with other state-of-the-art models on the same dataset, it was found that the C-LSTME model has higher prediction accuracy, taking the RMSE, MAE and MAPE as indicators. The main findings of the present study are as follows:

(1) The addition of PM2.5 information from neighboring stations, which contributes to the spatiality of the data, can considerably improve the prediction accuracy of the model.
(2) The supplement of auxiliary data can help predict sudden changes in air quality, thereby improving the prediction performance of the model. Moreover, compared to meteorological data, the AOD data contributes more to the accuracy of the model.
(3) The present model can efficiently extract more essential spatiotemporal correlation features through the combination of 3D-CNN and stateful LSTM, thereby yielding higher accuracy and stability for air quality prediction of different spatiotemporal scales.

Table 4
Comparison of the long-term prediction performance of C-LSTME model and LSTME model, where time lag refers to the number of historical moments that were selected, RMSE is the root mean square error, MAE is the mean absolute error, and MAPE is the mean absolute percentage error.

Prediction time	Time lag	C-LSTME			LSTME		
		RMSE (μg/m^3)	MAE (μg/m^3)	MAPE (%)	RMSE (μg/m^3)	MAE (μg/m^3)	MAPE (%)
1st hour	8	12.08	5.82	17.09	18.25	8.48	22.82
2nd hour	10	14.27	6.49	19.01	21.77	11.53	28.37
3rd hour	10	16.22	6.99	23.83	24.23	11.93	29.42
4th–6th hour	12	19.71	10.78	24.46	26.22	13.64	31.17
7th–12th hour	20	21.55	13.85	27.34	39.61	19.90	40.62
13th–24th hour	28	23.18	14.17	28.26	47.59	19.70	42.36
Standard deviation		**3.96**	3.44	**4.07**	**10.44**	4.24	**6.89**

Bold emphasis indicates the larger difference of two model in long-term prediction's standard deviation.

Acknowledgements

This research was financially supported by the National Science technology Support Plan Project of China (2015BAJ02B00), and the National Natural Science Foundation of China (Grant No. 41701438 and 41471430).

Appendix A. CNN related work

The Convolutional Neural Network (CNN) is a deep neural network model (Lecun et al., 1998), which has been widely used in speech analysis (Sukittanon et al., 2004) and image recognition (Y. N. Chen et al., 2007).The CNN model has two characteristics: local receptive field and weight sharing. Local receptive field means that a single neuron in each layer of the network is only connected to neurons in a corresponding neighborhood of its input layer. Weight sharing can greatly reduce the training parameters of the network model and requires relatively few training samples. The 2D-CNN directly processes two-dimensional matrices through weight sharing and convolution operations, avoiding complex feature extraction and data reconstruction processes in traditional pattern recognition algorithms. However, 2D-CNN does not consider the characteristics of the time dimension. For this reason, Du et al. (2015) proposed the use of 3D ConvNets on RGB video datasets to train deep three-dimensional convolutional networks for learning spatiotemporal features. The three-dimensional convolution formula is as follows:

$$v^{xyz} = \sum_{p=0}^{P-1}\sum_{q=0}^{Q-1}\sum_{r=0}^{R-1} w^{pqr} u^{(x+p)(y+q)(z+r)} \quad (1)$$

where w is the convolution kernel, u is the input feature map, v is the output feature map, the superscript represents the element value of

the corresponding position, and p, q and r are the sizes of the three dimensions, respectively.

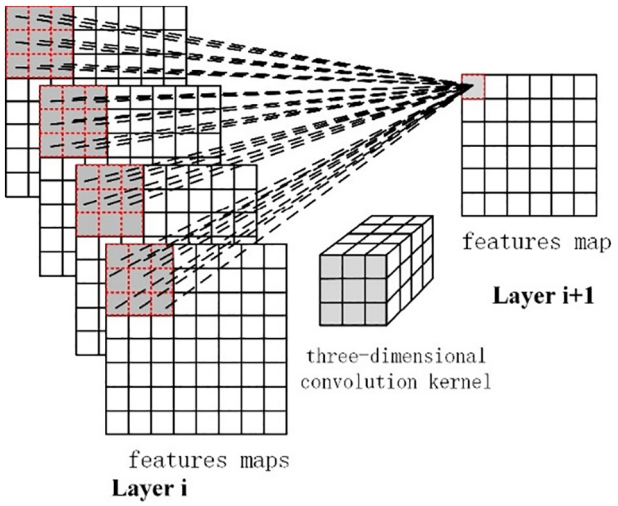


Fig. 1. Schematic diagram of the 3D convolution operation.

Appendix B. LSTM related work

The Recurrent Neural Network (RNN) has achieved great success in sequence learning tasks. However, the main known problem with RNN is the difficulty of modeling long-term dependencies due to gradient disappearance or explosion problems (Sepp, 1998; Y. Bengio et al., 2002). One of the most effective ways to solve this problem is to use the LSTM network (Hochreiter and Schmidhuber, 1997). The LSTM network introduces a new structure called memory cell, which stores long-term dependencies. The storage unit has three main elements: input gate, forget gate and output gate. Fig. 2 presents the network structure of the LSTM.

The input gate is designed to control the writing of input information to the memory, while the forget gate and output gate determines whether the information is saved or released from memory at each decision point. Since the variants of the LSTM do not show significant differences in performance (Greff et al., 2017), the common LSTM described in a reference (Zaremba and Sutskever, 2014) was used, where each gate, memory cell and output of the hidden layer are calculated, as follows:.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

where W and b denote the corresponding weights and bias vectors; x_t, h_t and C_t represent the input, output and memory cells at time t; h_{t-1} and C_{t-1} represent the output and memory cells at time $t-1$; i_t, o_t and f_t are the input, output and forget gates. In addition, $\sigma(\cdot)$ denotes the sigmoid function, which is defined in Eq. (8), and $\tan h(\cdot)$ denotes the tanh function, which is defined in Eq. (9).

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{8}$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{9}$$

References

Bengio, Y., 2009. Learning deep architectures for AI. Found. Trends® Mach. Learn. 2 (1), 1–127.

Bengio, Y., Simard, P., Frasconi, P., 2002. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5 (2), 157–166.

Box, G.E.P., Jenkins, G.M., 1976. Time series analysis, forecasting and control, Holden-day. J. R. Stat. Soc. 134 (3).

Bray, C.D., Battye, W., Aneja, V.P., et al., 2017. Evaluating ammonia (NH 3) predictions in the NOAA National Air Quality Forecast Capability (NAQFC) using in-situ aircraft and satellite measurements from the CalNex2010 campaign. Atmos. Environ. 163.

Che, H., Zhang, X., Li, Y., et al., 2007. Horizontal visibility trends in China 1981–2005. Geophys. Res. Lett. 34 (24), 497–507.

Chen, Y.N., Han, C.C., Wang, C.T., et al., 2007. A CNN-based face detector with a simple feature map and a coarse-to-fine classifier - withdrawn. IEEE Trans. Pattern Anal. Mach. Intell. PP (99), 1.

Chen, Y., Shi, R., Shu, S., et al., 2013. Ensemble and enhanced PM 10 concentration forecast model based on stepwise regression and wavelet analysis. Atmos. Environ. 74 (74), 346–359.

Chen, J., Lu, J., Avise, J.C., et al., 2014. Seasonal modeling of PM 2.5 in California's San Joaquin Valley. Atmos. Environ. 92, 182–190.

Cheng, Z., Wang, S., Jiang, J., et al., 2013. Long-term trend of haze pollution and impact of particulate matter in the Yangtze River Delta, China. Environ. Pollut. 182C (6), 101.

Chu, D.A., Kaufman, Y.J., Zibordi, G., et al., 2003. Global monitoring of air pollution over land from the Earth Observing System-Terra Moderate Resolution Imaging Spectroradiometer (MODIS). J. Geophys. Res.-Atmos. 108 (D21), 4661.

Coats, C.J., 1996. High Performance Algorithms. The Sparse Matrix Operator Kernel Emissions (smoke) Modeling System vol. 13(1) pp. 584–588.

Combarro, E.F., 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 (17), 8923–8937.

Deng, X., Tie, X., Wu, D., et al., 2008. Long-term trend of visibility and its characterizations in the Pearl River Delta (PRD) region, China. Atmos. Environ. 42 (7), 1424–1435.

Deng, J., Du, K., Wang, K., et al., 2012. Long-term atmospheric visibility trend in Southeast China, 1973–2010. Atmos. Environ. 59 (59), 11–21.

Díaz-Robles, L.A., Ortega, J.C., Fu, J.S., et al., 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos. Environ. 42 (35), 8331–8340.

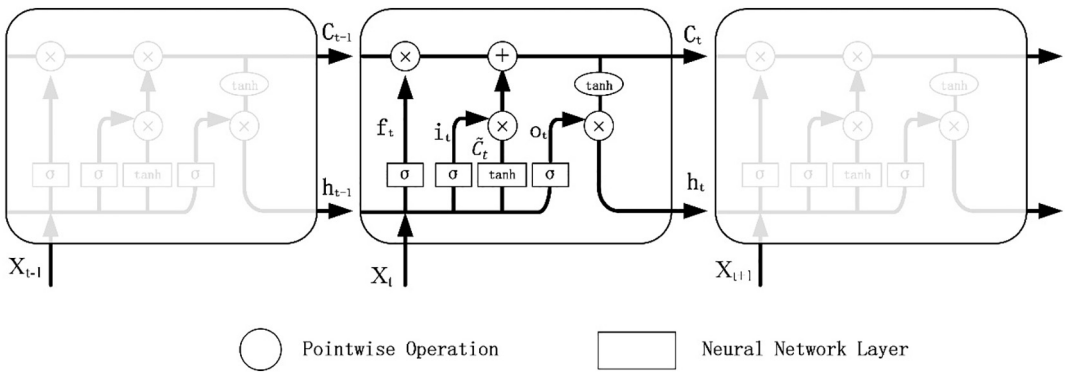


Fig. 2. The network structure of the LSTM.

Dominici, F., Peng, R.D., Bell, M.L., et al., 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA, J. Am. Med. Assoc. 295 (10), 1127–1134.
Donnelly, A., Misstear, B., Broderick, B., 2015. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. Atmos. Environ. 103 (103), 53–65.
Du, T., Bourdev, L., Fergus, R., et al., 2015. Learning spatiotemporal features with 3D convolutional networks. Proc. ICCV, 2015, pp. 4489–4497.
Engel-Cox, J.A., Holloman, C.H., Coutant, B.W., et al., 2004. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. Atmos. Environ. 38 (16), 2495–2509.
Feng, Y., Zhang, W., Sun, D., et al., 2011. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. Atmos. Environ. 45 (11), 1979–1985.
Greff, K., Srivastava, R.K., Koutnik, J., et al., 2017. LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. 28 (10), 2222–2232.
Gupta, S., Dinesh, D.A., 2017. Resource usage prediction of cloud workloads using deep bidirectional long short term memory networks. Proceedings of the 11th International Conference on Advanced Networks and Telecommunications Systems (ANTS), pp. 1–6.
Gupta, P., Christopher, S.A., Wang, J., et al., 2006. Satellite remote sensing of particulate matter and air quality assessment over global cities. Atmos. Environ. 40 (30), 5880–5892.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hooyberghs, J., Mensink, C., Dumont, G., et al., 2005. A neural network forecast for daily average PM concentrations in Belgium. Atmos. Environ. 39 (18), 3279–3289.
Jeong, J.I., Park, R.J., Woo, J.H., et al., 2011. Source contributions to carbonaceous aerosol concentrations in Korea. Atmos. Environ. 45 (5), 1116–1125.
Jian, D., Jiang, Y., 2015. A preliminary evaluation of global and east Asian cloud radiative effects in reanalyses. Atmos. Oceanic Sci. Lett. 8 (2), 100–106.
Koelemeijer, R., Homan, C., Matthijsen, J., 2006. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. Atmos. Environ. 40 (27), 5304–5315.
Koster, R.D., Bosilovich, M.G., Akella, S., et al., 2015. Technical report series on global modeling and data assimilation. MERRA-2; Initial Evaluation of the Climate. vol. 43.
Kurt, A., Oktay, A.B., 2010. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst. Appl. 37 (12), 7986–7992.
Lecun, Y., Bottou, L., Bengio, Y., et al., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.
Li, C., Hsu, N.C., Tsay, S.C., 2011. A study on the potential applications of satellite data in air quality monitoring and forecasting. Atmos. Environ. 45 (22), 3663–3675.
Li, X., Peng, L., Yao, X., et al., 2017. Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environ. Pollut. 231 (Pt 1), 997–1004.
Li, T., Shen, H., Yuan, Q., et al., 2017. Estimating ground-level PM2.5 by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophys. Res. Lett. 44 (23).
Neto, P.S.G.D.M., Madeiro, F., Ferreira, T.A.E., et al., 2014. Hybrid intelligent system for air quality forecasting using phase adjustment. Eng. Appl. Artif. Intell. 32 (6), 185–191.
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 (6), 1553–1566.
Pan, L., Sun, B., Wang, W., 2011. City air quality forecasting and impact factors analysis based on Grey model. Procedia Eng. 12, 74–79.
Pearson, K., 2006. Note on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. 58, 240–242.
Pope, C., Burnett, R.T., Thun, M.J., et al., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. J. Am. Med. Assoc. 287 (9), 1132–1141.
Prakash, A., Kumar, U., Kumar, K., et al., 2011. A wavelet-based neural network model to predict ambient air pollutants' concentration. Environ. Model. Assess. 16 (5), 503–517.
Prasad, K., Gorai, A.K., Goyal, P., 2016. Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time. Atmos. Environ. 128 (1), 246–262.
Ridder, K.D., Kumar, U., Lauwaet, D., et al., 2012. Kalman filter-based air quality forecast adjustment. Atmos. Environ. 50 (4), 381–384.
Rienecker, M.M., Suarez, M.J., Gelaro, R., et al., 2011. MERRA: NASA's modern-era retrospective analysis for research and applications. J. Clim. 24 (14), 3624–3648.
Saide, P.E., Carmichael, G.R., Spak, S.N., et al., 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 (16), 2769–2780.
Sepp, H., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertainty Fuzziness Knowledge Based Syst. 6 (02), 107–116.
Soh, P.-W., Chang, J.-W., Huang, J.-W., 2018. Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access 6, 38186–38199.
Stern, R., Builtjes, P., Schaap, M., et al., 2008. A model inter-comparison study focussing on episodes with elevated PM10 concentrations. Atmos. Environ. 42 (19), 4567–4588.
Sukittanon, S., Surendran, A.C., Platt, J.C., et al., 2004. Convolutional networks for speech detection. Interspeech, pp. 1077–1080.
Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234–240.
Trenberth, K.E., Olson, J.G., 1988. An evaluation and intercomparison of global analyses from the National Meteorological Center and the European Centre for Medium Range Weather Forecasts. Bull. Am. Meteorol. Soc. 69 (9), 1047–1056.
Vardoulakis, S., Fisher, B.E.A., Pericleous, K., et al., 2003. Modelling air quality in street canyons: a review. Atmos. Environ. 37 (2), 155–182.
Vautard, R., Builtjes, P.H.J., Thunis, P., et al., 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 (1), 173–188.
Wang, J., Christopher, S.A., 2003. Intercomparison between satellite-derived aerosol optical thickness and PM2.5 mass: implications for air quality studies. Geophys. Res. Lett. 30 (21), 267–283.
Wang, Z., Maeda, T., Hayashi, M., et al., 2001. A nested air quality prediction modeling system for urban and regional scales: application for high-ozone episode in Taiwan. Water Air Soil Pollut. 130 (1–4), 391–396.
Wang, T., Jiang, F., Deng, J., et al., 2012. Urban air quality and regional haze weather forecast for Yangtze River Delta region. Atmos. Environ. 58 (15), 70–83.
Wang, D., Wei, S., Luo, H., et al., 2016. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Sci. Total Environ. 580, 719–733.
Woody, M.C., Wong, H.W., West, J.J., et al., 2016. Multiscale predictions of aviation-attributable PM2.5 for U.S. airports modeled using CMAQ with plume-in-grid and an aircraft-specific 1-D emission model. Atmos. Environ. 147, 384–394.
Xu, Y., Du, P., Wang, J., 2017. Research and application of a hybrid model based on dynamic fuzzy synthetic evaluation for establishing air quality forecasting and early warning system: a case study in China. Environ. Pollut. 223, 435–448.
Zaremba, W., Sutskever, I., 2014. Learning to Execute (ArXiv preprint arXiv:1410.4615).
Zhang, G., 2004. Progress of weather research and forecast (WRF) model and application in the United States. Meteorol. Monthly 30 (12), 27–31.
Zhang, Y., Li, Z., 2015. Remote sensing of atmospheric fine particulate matter (PM 2.5) mass concentration near the ground from satellite observation. Remote Sens. Environ. 160, 252–262.
Zhao, P., Zhang, X., 2011. Long-term visibility trends and characteristics in the region of Beijing, Tianjin, and Hebei, China. Atmos. Res. 101 (3), 711–718.
Zheng, Y., Liu, F., Hsieh, H.P., 2013. U-Air:when urban air quality inference meets big data. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1436–1444.
Zheng, Y., Yi, X., Li, M., et al., 2015. Forecasting fine-grained air quality based on big data. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 2267–2276.
Zhou, G., Xu, J., Xie, Y., et al., 2017. Numerical air quality forecasting over eastern China: an operational application of WRF-Chem. Atmos. Environ. 153, 94–108.