RCL-Learning: ResNet and Convolutional Long Short-Term Memory-based Spatiotemporal Air Pollutant Concentration Prediction Model★

Bo Zhang[a,d,e], Guojian Zou[b,c,1], Dongming Qin[f], Qin Ni[a,*], Hongwei Mao[a,*] and Maozhen Li[a,e]

[a] *College of Information, Mechanical, and Electrical Engineering, Shanghai Normal University, Shanghai 200234, PR China*
[b] *The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, PR China*
[c] *College of Transportation Engineering, Tongji University, Shanghai 201804, PR China*
[d] *Institute of Artificial Intelligence on Education, Shanghai Normal University, Shanghai 200234, PR China*
[e] *Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai 200234, PR China*
[f] *College of Electronics and Information Engineering, Tongji University, Shanghai 201804, PR China*

ARTICLE INFO

Keywords:
Deep learning
Residual network
Convolutional long short-term memory
Air pollutant concentration prediction

ABSTRACT

Predicting the concentration of air pollutants is an effective method for preventing pollution incidents by providing an early warning of harmful substances in the air. Accurate prediction of air pollutant concentration can more effectively control and prevent air pollution. In this study, a big data correlation principle and deep learning technology are used for a proposed model of predicting $PM_{2.5}$ concentration. The model comprises a deep learning network model based on a residual neural network (ResNet) and a convolutional long short-term memory (LSTM) network (ConvLSTM). ResNet is used to deeply extract the spatial distribution features of pollutant concentration and meteorological data from multiple cities. The output is used as input to ConvLSTM, which further extracts the preliminary spatial distribution features extracted from the ResNet, while extracting the spatiotemporal features of the pollutant concentration and meteorological data. The model combines the two features to achieve a spatiotemporal correlation of feature sequences, thereby accurately predicting the future $PM_{2.5}$ concentration of the target city for a period of time. Compared with other neural network models and traditional models, the proposed pollutant concentration prediction model improves the accuracy of predicting pollutant concentration. For 1- to 3-hours prediction tasks, the proposed pollutant concentration prediction model performed well and exhibited root mean square error (RMSE) between 5.478 and 13.622. In addition, we conducted multiscale predictions in the target city and achieved satisfactory performance, with the average RMSE value able to reach 22.927 even for 1- to 15-hours prediction tasks.

1. Introduction

In recent years, the increasingly serious problem of air pollution has caused widespread concerns around the world (Fong, Li, Fong, Wong & Tallon-Ballesteros, 2020). Therefore, the prediction of air pollutant concentration obtains great attentions since it plays a significant role for air pollution prevention and environment management (Maleki, Sorooshian, Goudarzi, Baboli, Birgani & Rahmati, 2019). Various organizations have recognized that the air pollutant concentration prediction technology is currently a key challenge for environment management research field.

China's $PM_{2.5}$ level is one of the highest in the world. In China, studies have shown that $PM_{2.5}$ was the cause of nearly 1.4 million deaths in 2015 due to stroke ($PM_{2.5}$ is responsible for 40% of stroke deaths), lung cancer (24%), acute pulmonary infection (33%), and ischemic heart disease (27%) (Liu, Han, Tang, Zhu & Zhu, 2016; Song, He, Wu, Jin, Chen, Li, Ren, Zhang & Mao, 2017). As for the Yangtze River Delta region, more than 13,000 people died from short-term exposure to $PM_{2.5}$ in 2010, causing economic losses of 22 billion RMB (Wang, Wang, Voorhees, Zhao, Jang, Jiang, Fu, Ding, Zhu & Hao, 2015). Effective control of $PM_{2.5}$ can not only protect people's health but also reduce social and economic losses. Therefore, accurate prediction of $PM_{2.5}$ concentration can provide people with timely warnings

*Corresponding author
zhangbo@shnu.edu.cn (B. Zhang); 2010768@tongji.edu.cn (G. Zou); qindm@3clear.com (D. Qin); niqin@shnu.edu.cn (Q. Ni); maohw2007@shnu.edu.cn (H. Mao); Maozhen.li@gmail.com (M. Li)
ORCID(s): 0000-0002-6490-1583 (G. Zou)

[1] This author contributed equally to this work and should be considered co-first author.

and enable the government to take timely actions for the environment. $PM_{2.5}$ concentration prediction can be seen as a time series processing problem that can be predicted based on past historical related data, e.g., meteorological factors such as humidity and temperature, and other pollutant factors, such as SO_2, and CO. Many works have proven that there is a complex interaction between these factors for air pollutant (Chang-Hoi, Park, Oh, Gim, Hur, Kim & Choi, 2021; Feng, Li, Zhu, Hou, Jin & Wang, 2015; Zhu, Gao, Liu, Wang, Zhu, Xu, Wang, Ding, Li & Duan, 2019; Saide, Carmichael, Spak, Gallardo, Osses, Mena-Carrasco & Pagowski, 2011; Huang & Kuo, 2018). Therefore, the features among such complex interaction relationships must be extracted and learned for further air pollutant prediction. In addition, air pollution is a regional diffusion problem that causes a spatial dimension consideration. This means that there is a spatial correlation air pollution impact among neighboring cities (Mayer, 1999; Akimoto, 2003; McKinley, Zuk, Höjer, Avalos, González, Iniestra, Laguna, Martínez, Osnaya, Reynales et al., 2005; Zhu, Sun & Li, 2017b).

In many existing works on the prediction of air pollution concentration (Zhu et al., 2019; Saide et al., 2011; Cordano & Frieze, 2000; Tian & Chen, 2010; Russell, McCue & Cass, 1988; Suleiman, Tight & Quinn, 2019; Chen, Lu, Avise, DaMassa, Kleeman & Kaduwela, 2014; Wang, Maeda, Hayashi, Hsiao & Liu, 2001; Corani & Scanagatta, 2016; Yang, Deng, Xu & Wang, 2018; Sun, Zhang, Palazoglu, Singh, Zhang & Liu, 2013; Zamani Joharestani, Cao, Ni, Bashir & Talebiesfandarani, 2019), a numerical prediction method is widely used, which can employ historical air pollutants to realize the prediction analysis of the future state of pollution. Most numerical prediction models include a deterministic model based on hypothesis theory and prior knowledge; an empirical model that only considers input and output as an independent process; a mathematical statistics model; or a traditional machine learning model with small sample data (Cordano & Frieze, 2000; Tian & Chen, 2010; Russell et al., 1988; Suleiman et al., 2019). The main advantages of these models are low computational complexity, fast calculation speed, and ease of implementation. However, by dealing with the massive amount of spatiotemporal data from multi-city sites for a spatial correlation air pollutant concentration prediction, traditional numerical analysis models have encountered three problems: (1) the complex correlation features among meteorological data and air pollution data should be extracted and learned for further prediction and performance improving; and (2) the temporal dependency feature among the historical data should be extracted accurately for prediction. This means that the redundant information or features from past long time intervals should be ignored in prediction, while the useful information or features should be taken into account in some duration for improving prediction; and (3) the spatial-related features among the neighboring cities in a region should be extracted based on their massive amounts of meteorological data and pollution data with temporal series tags. These problems have led to most traditional air pollutant prediction models performing poorly.

To date, deep learning models have shown better performance in spatiotemporal prediction, especially in the fields of image recognition, natural language processing (NLP), and historical data-based prediction (including the field of air pollutant concentration prediction) (Chang-Hoi et al., 2021; Suleiman et al., 2019; Li, Peng, Yao, Cui, Hu, You & Chi, 2017; Hossain, Rekabdar, Louis & Dascalu, 2015; Gu, Qiao & Li, 2018; Chen, An et al., 2019; Kim & Won, 2018; Luong, Pham & Manning, 2015; Yi, Wen, Tao, Ni & Liu, 2018a). In particular, many existing works in air pollution prediction, and their experimental results, have proved that the deep network structure of neural network models have better performance than traditional pollutant prediction methods, as well as traditional machine learning algorithms, because the deep features of spatial dimension and time dependence can be learned more accurately (Wang & Christopher, 2003; Mokhtari, Miri, Mohammadi, Khorsandi, Hajizadeh & Abdolahnejad, 2015; Hao & Liu, 2016; Cairncross, John & Zunckel, 2007; Zhu et al., 2017b; Zhu, Zhang, Zhang, Zhi, Li, Han & Zheng, 2017c; Lin, Li, Zheng, Cheng & Yuan, 2020; Le, Bui & Cha, 2020; Xu & Lv, 2019; Yi, Zhang, Wang, Li & Zheng, 2018b). In light of these, we propose two types of artificial neural networks to construct our prediction model: residual neural network (ResNet) and a convolutional long short-term memory network (ConvLSTM). The rationales are as follows:

(1) A deep network, e.g., a convolutional neural network (CNN), can extract and learn the spatial-related features in the fields of NLP and computer vision. However, the problems of vanishing gradients and network degradation are exacerbated as the network layer depth increases. Therefore, we introduce the ResNet framework to extract the spatial correlation features of data and avoid these two problems (He, Zhang, Ren & Sun, 2016a; Wu, Schuster, Chen, Le, Norouzi, Macherey, Krikun, Cao, Gao, Macherey et al., 2016; Ren, He, Girshick & Sun, 2015; He, Zhang, Ren & Sun, 2015, 2016b). The performance of the network will improve as the number of layers increases (He et al., 2016b; Dong, Loy, He & Tang, 2014; Sainath, Kingsbury, Saon, Soltau, Mohamed, Dahl & Ramabhadran, 2015). Similarly, the pollutant prediction method proposed in this study fully considers the prediction problems of pollutants and meteorological data in multiple cities, and we use the advantages of the deep residual neural network (ResNet) to extract the spatial features of inputs among multi-city such data.

(2) Convolutional LSTM (ConvLSTM) is proposed to extract time series features by combining the spatial convolu-

tion operation, which aims to learn the spatiotemporal association features from the high dimensional data (Sønderby, Sønderby, Nielsen & Winther, 2015; Zhu, Zhang, Shen & Song, 2017a). Compared with the recurrent neural network (RNN) models, ConvLSTM can not only avoid exploding and vanishing gradient problems, but also solve the problem of correlating the spatial and temporal features of high-dimensional data (Karim & Rafi, 2020; Xingjian, Chen, Wang, Yeung, Wong & Woo, 2015; Zhang, Zhu, Shen, Song, Afaq Shah & Bennamoun, 2017). Therefore, we can use the advantages of convolutional LSTM to perform deeper spatiotemporal correlation feature extraction on the extracted high-dimensional spatial features by ResNet.

In this study, we propose an end-to-end deep learning model-RCL-Learning that integrates ResNet and ConvLSTM. The main contributions of this work are as follows:

(1) by utilizing ResNet as the base of the proposed RCL-Learning model to avoid the problem of vanishing gradients or exploding gradients, the spatial correlation features can be extracted from the pollutant and meteorological data of multiple cities, and the problem of the degradation of the deep network is also eliminated (Srivastava, Greff & Schmidhuber, 2015);
(2) by adopting the ConvLSTM as the output prediction layer, the model obtains not only the performance advantages of time series prediction through ConvLSTM, but also avoids the problem of vanishing gradients, and thereby extracts the high-level correlation features hidden in the high-dimensional data output from the residual network layer to realize the target of the mining data spatiotemporal correlation, and;
(3) the proposed RCL-Learning model can simultaneously apply the meteorological and pollution data from multiple cities for the environmental monitoring of big data, taking into consideration changes in spatial and temporal distributions of data, as well as regulations, to achieve air pollutant concentration prediction in target city. Experiments on the data set show that our framework achieves better results than other state-of-the-art methods.

2. Related Work

According to the characteristics of the prediction methods used in related studies, air pollutant concentration prediction can be fundamentally divided into two major research methods: deterministic and statistical approaches (Feng et al., 2015; Park, Kim, Kim, Namgung, Kim, Cho & Kwon, 2018; Lee, Szpiro, Kim & Sheppard, 2015; Chen et al., 2014).

Deterministic approaches can be applied to a limited set of historical data. However, meteorological principles and statistical approaches are needed to simulate the process of real-time emission, diffusion, transformation, and removal of pollutants based on atmospheric physics and chemical reactions. The model structure based on the deterministic approaches are predefined based on certain theoretical assumptions and prior knowledge. There are several commonly used methods for air pollutant concentration prediction based on deterministic approaches: comprehensive air quality model with extensions (CAMs), the WRFChem model, nested air quality prediction modeling system (NAQPMS), and the community multiscale air quality (CMAQ) model (Zhu et al., 2019; Saide et al., 2011; Chen et al., 2014; Wang et al., 2001).

Statistical approaches can avoid the use of complex theoretical models. Compared with deterministic approaches, they can determine the correlations among complex pollutant concentration data and thus show better predictive performance. Based on the statistics, the two branches of approaches can be extended into traditional machine learning methods, and new deep learning methods. Traditional machine learning methods include a support vector machine (SVM), multi-label classifier based on Bayesian networks, the support vector regression (SVR) method, hidden Markov model (HMM), and other methods (Suleiman et al., 2019; Corani & Scanagatta, 2016; Yang et al., 2018; Sun et al., 2013; Zamani Joharestani et al., 2019).

In recent years, deep learning technology has excelled in dealing with regression problems, and various neural networks have been used to improve air pollution concentration prediction performance. Typical network models for predicting air pollution include the multilayer perceptron (MLP), recurrent neural network (RNN), LSTM neural network, the latest proposed deep CNN-LSTM model, graph convolutional neural network, and attention-based neural networks, etc (Fong et al., 2020; Maleki et al., 2019; Chang-Hoi et al., 2021; Feng et al., 2015; Huang & Kuo, 2018; Li et al., 2017; Chen et al., 2019; Park et al., 2018; Feng, Gao, Luo & Fan, 2020; Kolehmainen, Martikainen & Ruuskanen, 2001; Wang & Christopher, 2003; Mokhtari et al., 2015; Hao & Liu, 2016; Cairncross et al., 2007; Zhu et al., 2017b,c; Lin et al., 2020; Le et al., 2020; Xu & Lv, 2019; Yi et al., 2018b; Qin, Yu, Zou, Yong, Zhao & Zhang, 2019; Zhang, Zou, Qin, Lu, Jin & Wang, 2021). Because the emission, diffusion, conversion, and removal of air pollutants are a dynamic process over time, the CNN-LSTM characteristic is that it can process the time series data prediction problem and easily extract temporal and spatial features of pollutant concentrations (Huang & Kuo, 2018; Qin et al., 2019). However, the CNN-LSTM has three key problems (Xingjian et al., 2015). First of all, it is difficult for a CNN to extract the spatial features of pollutant data in depth, which can easily lead to loss of feature information and degradation of the model. Second, CNN-LSTM extracts the temporal and spatial characteristics of pollutants as an asynchronous process, so it is difficult to extract spatiotemporal correlation features of multi-city pollutants and meteorological data (Xingjian et al., 2015; Zhang et al., 2017). Third, because LSTM is mostly used to extract one-dimensional time series features, it is impossible to process high-dimensional input data (Sønderby et al., 2015; Zhu et al., 2017a).

In recent studies on pollutant concentration prediction, methods based on graphs and attention have been used to extract the spatiotemporal features of multi-site pollutant data, and the accuracy of prediction has been improved to a certain extent (Zhu et al., 2017c; Qi, Li, Karimian & Liu, 2019). In addition, progress has been made in spatiotemporal feature extraction based on attention and ConvLSTM. In references (Xue, Ji, Zhang & Cao, 2019; Lin et al., 2020; Xu & Lv, 2019), attention-based ConvLSTM (Att-ConvLSTM) is used in recognition, traffic flow prediction and pollutant concentration prediction tasks, and has achieved satisfactory performance. However, the current pollutant concentration prediction model based on Att-ConvLSTM has encountered the following challenges: First, it lacks consideration of the impact of multiple pollutants and meteorological factors on the prediction results; Second, the ConvLSTM method mainly extracts the spatiotemporal correlation features of long-term sequence data, combining the advantages of CNN and LSTM models. However, the single ConvLSTM network model has a major shortcoming. On the one hand, it is limited by the feature dimension of the input data, that is, the dimension of the hidden state is affected by the dimension of the input data. On the other hand, as the number of ConvLSTM layers increases, the model will have more problems with network degradation and training costs will increase rapidly. Therefore, it is difficult for Att-ConvLSTM to overcome the above two problems.

This paper fully considers that the prediction model should make a more accurate prediction of the $PM_{2.5}$ concentration of the target city in the future, and it should accomplish the following objectives: (1) Effectively use the historical pollutant concentration and meteorological big data from multiple cities; (2) Deep mining of the spatiotemporal correlation features of historical multiple cities pollutants and meteorological data.

3. Data Description

3.1. Data collection

The experiment used historical pollutant concentration and meteorological data from 14 cities collected from May 13, 2014 to May 30, 2018 [1]. The experimental data in this paper is based on the city level, that is, the sample data of each city every hour is a one-dimensional feature vector, and the feature elements are composed of pollutant and meteorological factors. In this paper, the selection of city sites, 14 cities (Shanghai, Nanjing, Suzhou, Nantong, Wuxi, Changzhou, Zhenjiang, Hangzhou, Ningbo, Shaoxing, Huzhou, Jiaxing, Taizhou, and Zhoushan) with rapid economic development in the Yangtze River Delta region centered on Shanghai. The geographical location of these cities is closest to Shanghai, and the spread of pollutants is more likely to affect each other. We selected 16 pollutants and meteorological factors: air quality index (AQI), $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, CO, temperature, humidity, air pressure, wind direction, wind speed, clouds, maximum temperature, minimum temperature and conditions. For non-numerical meteorological factors, including clouds and conditions, we perform a one-to-one numerical mapping. For the conditions factor, we map the 'mist' value to 1, the 'clear' value to 2, and the 'cloudy' value to 3. The missing values of the air pollutant concentration and meteorological data set are filled by spatiotemporal interpolation (Yang & Hu, 2018). Fig. 1 shows the locations of all city sites.

3.2. Particulate matter ($PM_{2.5}$) and air quality index (AQI)

We calculated the correlation coefficients between the AQI and air pollutants in the training set, validation set and test set, as shown in Table 1. Researchers have proposed that the $PM_{2.5}$ concentration can be used to evaluate the air quality (Wang & Christopher, 2003; Mokhtari et al., 2015; Hao & Liu, 2016; Cairncross et al., 2007). In Table 1, the correlation between AQI and $PM_{2.5}$ is the highest, the correlation coefficient value is 0.993 on the training set, and the correlation coefficient value on the test set is as high as 0.997, which also proves the findings of previous research (Wang & Christopher, 2003; Mokhtari et al., 2015; Hao & Liu, 2016; Cairncross et al., 2007). Therefore, this paper selects $PM_{2.5}$ with the highest correlation with AQI as the prediction target.

[1]https://github.com/zouguojian/Pollutant-concentration-and-meteorological-data

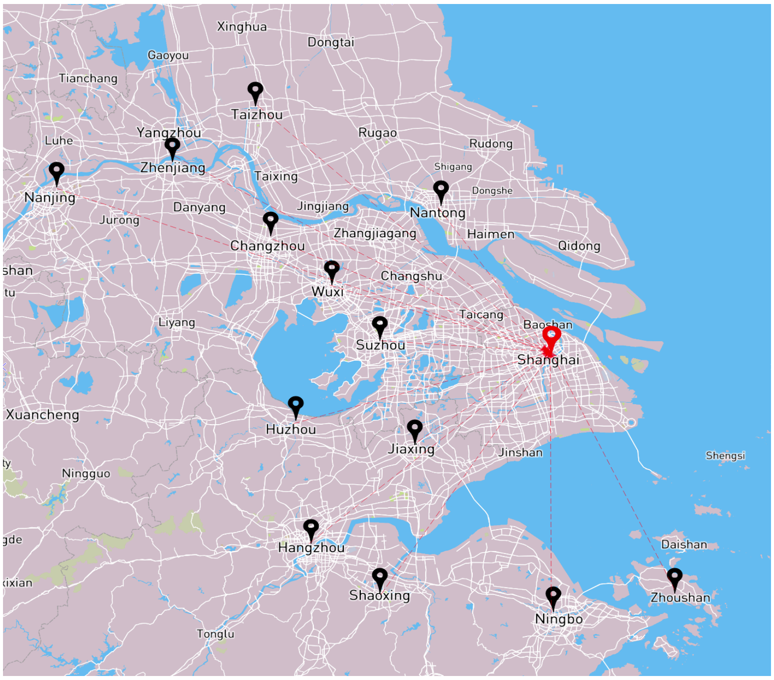


Fig. 1: The black circle indicates the surrounding city, the red circle indicates the target city, and the arrow indicates the possible impact of the surrounding city pollutants on the target city.

Table 1
Correlation coefficients between AQI and air pollutants

AQI and air pollutant	Training set	Validation set	Test set
(AQI & $PM_{2.5}$)	**0.993**	**0.996**	**0.997**
(AQI & PM_{10})	0.967	0.967	0.967
(AQI & SO_2)	0.757	0.836	0.846
(AQI & NO_2)	0.764	0.713	0.657
(AQI & O_3)	-0.506	-0.375	-0.217
(AQI & CO)	0.887	0.872	0.829

3.3. The distribution characteristics of data

3.3.1. Analysis of temporal dimension

To explore the distribution characteristics of pollutant concentration and meteorological data, we selected the 2016 annual data of the target city Shanghai as the research object. Fig. 2 shows the annual numerical changes of each pollutant concentration, including AQI. Observing the changes in the concentration of pollutants such as $PM_{2.5}$, it can be found that the trend of changes in the concentration of pollutants is generally consistent, which also reflects that there may be hidden relationships among pollutants. After statistical analysis, 49.4% of the time in 2016, the $PM_{2.5}$ concentration is greater than WHO's first interim level of 35 $\mu g/m^3$, which will have a weaker impact on the health of some abnormally sensitive people; 13.7% of the time in 2016, the $PM_{2.5}$ concentration is greater than 75 $\mu g/m^3$, which will directly affect people's daily travel and physical health (Martins & Da Graca, 2018; Zhixiang, Cai, Xiangwei, Wei & Chuanzhen, 2021). Therefore, for $PM_{2.5}$ prediction, on the one hand, we need to consider the hidden relationship between $PM_{2.5}$ and other pollutants; on the other hand, it reflects that accurate prediction can prevent the impact of $PM_{2.5}$ on people's health in advance.

Fig. 3 shows the annual numerical changes of meteorological factors. From Fig. 3, we can observe that, first, temperature, the maximum temperature, and the minimum temperature have the same changes, and the numerical change of the air pressure is precisely the opposite of the temperature; second, the numerical types and intervals of mete-

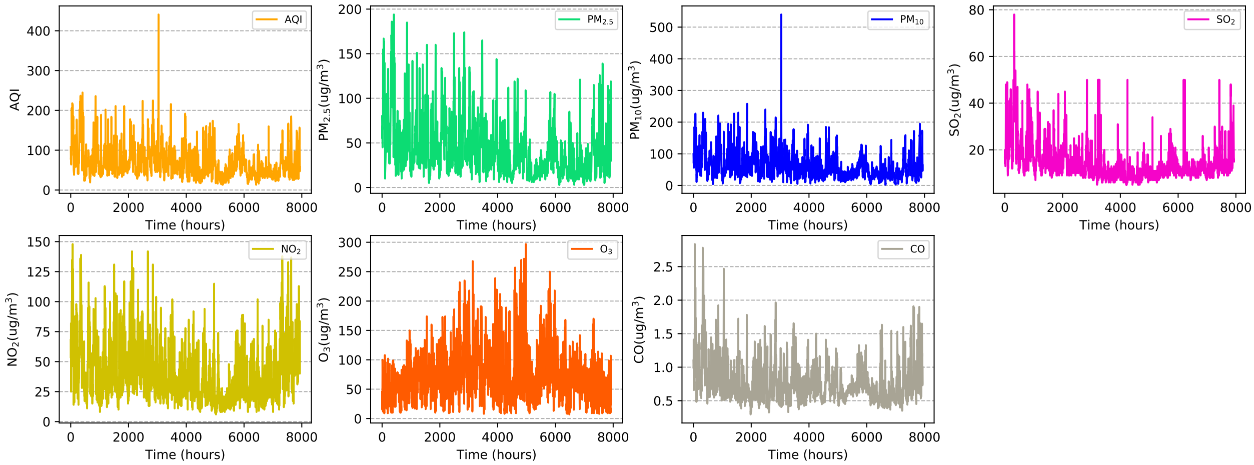


Fig. 2: Time series plots of air pollutant concentration data.

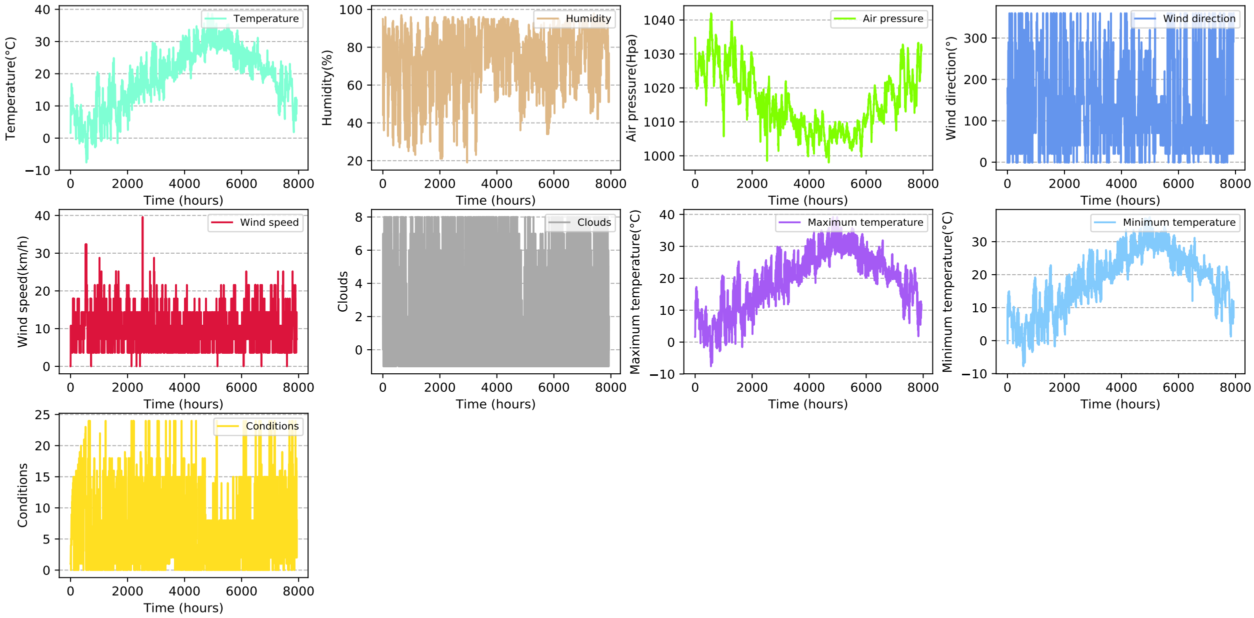


Fig. 3: Time series plots of Meteorological data.

orological elements are quite different, but the trend of change is highly similar, which means there may be mutual influences among meteorological factors. For example, as shown in Fig. 3, high temperature may result in low air pressure, and vice versa; third, the meteorological factors are consistent with the changes in $PM_{2.5}$ concentration, implying the hidden correlation between air pollutants and meteorological factors. For example, between 5000-6000 hours, the observed value fluctuates with different amplitudes. Therefore, combined with existing research results (Chang-Hoi et al., 2021; Feng et al., 2015; Zhu et al., 2019; Saide et al., 2011; Huang & Kuo, 2018), we use meteorological factors as part of the model input to extract hidden features between pollutants and meteorological factors in the $PM_{2.5}$ concentration prediction research.

Table 2
Correlation coefficients of air pollutants between Shanghai and neighboring cities

City pair	$PM_{2.5}$	PM_{10}	SO_2	NO_2	O_3	CO
(Shanghai & Nanjing)	0.569	0.599	0.554	0.611	0.690	0.491
(Shanghai & Suzhou)	**0.806**	**0.812**	**0.713**	**0.785**	**0.853**	**0.740**
(Shanghai & Nantong)	**0.785**	**0.768**	0.597	**0.697**	**0.845**	**0.713**
(Shanghai & Wuxi)	**0.764**	**0.752**	**0.743**	0.691	**0.810**	0.487
(Shanghai & Changzhou)	0.652	0.669	**0.610**	0.632	0.769	0.501
(Shanghai & Zhenjiang)	0.594	0.575	0.468	0.587	0.708	0.405
(Shanghai & Hangzhou)	0.582	0.519	0.555	0.619	0.743	0.553
(Shanghai & Ningbo)	0.709	0.636	**0.705**	**0.723**	0.777	**0.662**
(Shanghai & Shaoxing)	0.572	0.329	0.479	0.601	0.707	0.561
(Shanghai & Huzhou)	0.630	0.583	0.503	0.648	0.767	0.557
(Shanghai & Jiaxing)	**0.786**	**0.725**	0.587	**0.763**	**0.861**	**0.676**
(Shanghai & Taizhou)	0.517	0.463	0.392	0.594	0.690	0.433
(Shanghai & Zhoushan)	0.737	0.675	0.323	0.596	0.674	0.511

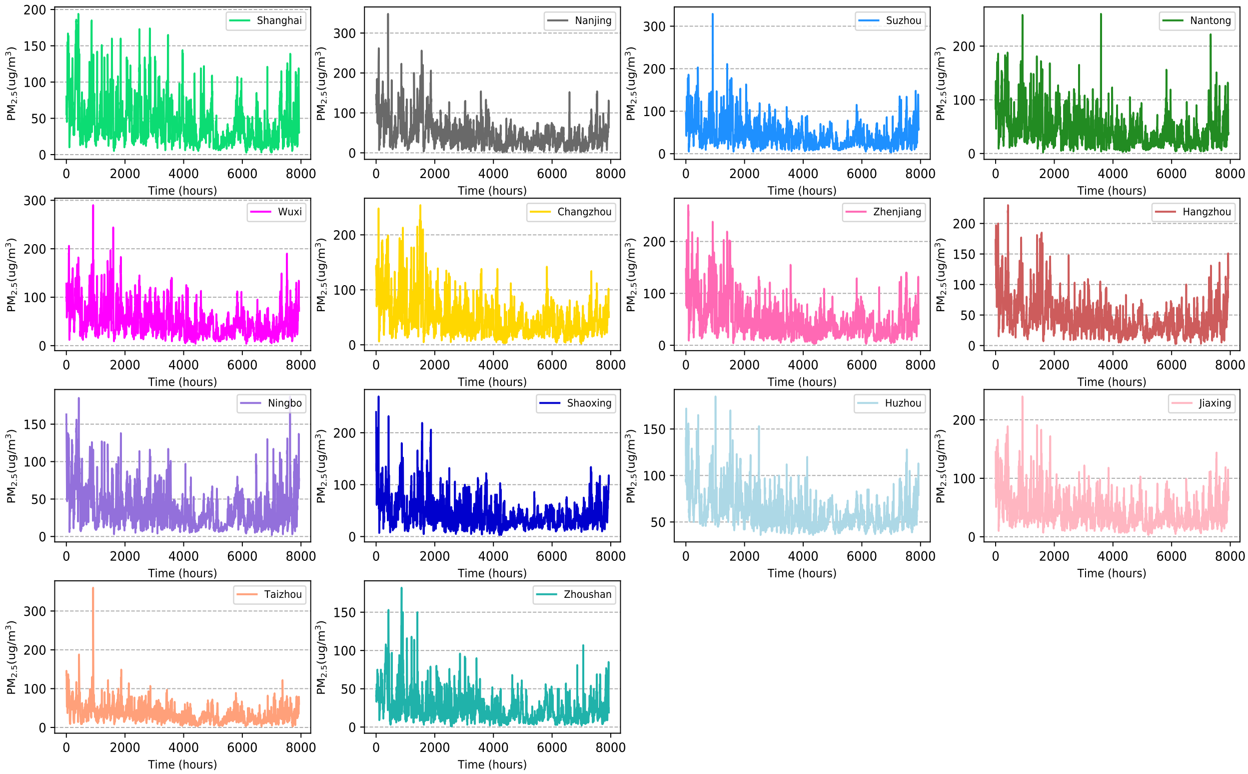


Fig. 4: The spatial distribution characteristics of $PM_{2.5}$ in Shanghai and neighboring cities.

3.3.2. Analysis of spatial dimension

The numerical changes of pollutants and meteorological factors in Fig. 2 and Fig. 3 are in the temporal dimension, and we have done a detailed analysis. As the target city of Shanghai, its $PM_{2.5}$ concentration may also have some characteristics in the spatial dimension. Similarly, we select the $PM_{2.5}$ concentration data of all cities in 2016. We calculated the correlation coefficients of air pollutants between Shanghai and surrounding cities, as shown in Table 2. Combining Table 2 and Fig. 1, first of all, we observe that cities with a shorter distance from Shanghai show higher correlation, which we indicate in bold in the table, and the correlation coefficient of $PM_{2.5}$ is generally higher than that of PM_{10}; secondly, as the distance increases, the correlation coefficients of air pollutants between Shanghai and neighboring cities gradually decreases. The influence of distance indicates that for any urban area, in addition to preventing local pollutants, it is also necessary to coordinate the prevention of regional pollutants (Hu, Wang, Ying & Zhang, 2014), reflecting the spatial relevance of air pollutants.

Next, Fig. 4 shows the changes in $PM_{2.5}$ concentration in Shanghai and neighboring cities. First, from Fig. 3 and Fig. 4, we can find that a general rule of $PM_{2.5}$ concentration in all cities is that the concentration is low when the temperature is high, and the concentration is high when the temperature is low. Second, in the spatial and temporal dimensions, we found that the change patterns of $PM_{2.5}$ in all cities are similar in Fig. 4. Third, by comparing the changes in $PM_{2.5}$ concentration between Shanghai and neighboring cities, we found that the $PM_{2.5}$ concentration in Shanghai fluctuates wildly and is more complicated. According to the spatial correlation characteristics of pollutants and the characteristics of pollutant concentration in Shanghai and neighboring cities (Wang, Ying, Hu & Zhang, 2014; Hu et al., 2014), this reflects the importance of considering the spatial correlation of pollutant concentrations in multiple cities in $PM_{2.5}$ concentration prediction research.

3.4. Data division

In our experiment, we selected 70% of the data as the training set, 15% as validation set, and the remaining 15% was used as the test set. The specific method of dividing the data in this study is as follows: first, we divide the data set uniformly according to a given window length L and a moving step size of S, and finally the total number of samples obtained is $N = ((D - D * 0.15) - L) / S$; then, we scramble the N samples, select 82% of them as the training set and 18% as the validation set. In addition, 15% of D is used as the test set, which means that we extract 15% of the data from the original data set as the test set without disturbing it; finally, we define our division method as a generalized random method. Among them, the window length L represents the sum of the time sequence length of the input model and the target prediction sequence length, and D is the size of the original data set.

4. Methodology

4.1. Framework overview

RCL-Learning is an end-to-end predictive model, and its entire training process is a mapping from the original input to the expected output. The inputs of RCL-Learning are the records of multi-city pollutant concentration and meteorological data $x = \{x_t, \cdots, x_{t-i}, \cdots, x_{t-r+1}\}, x_{t-i} \in R^{k*m}$ (k represents the number of cities, m indicates pollutants and meteorological factors), over the last r hours. The output is the $PM_{2.5}$ concentration of the future n hours $\hat{y} = \{\hat{y}_{t+1}, \cdots, \hat{y}_{t+j}, \cdots, \hat{y}_{t+n}\}, \hat{y}_{t+j} \in R$, where $\hat{y}_{t+j}$ represents the predicted value. Unlike the traditional pollutant prediction model, this study combines the advantages of ResNet and ConvLSTM networks to design a three-level architecture for RCL-Learning. The base consists of ResNet, and multiple convolution layers are used to extract deep spatial features from the pollutant and meteorological data. At the end of this layer, the ResNet extracts high-level spatial semantic features $out = \{out_t, \cdots, out_{t-i}, \cdots, out_{t-r+1}\}, out_{t-i} \in R^{h*w*c}$ (h and w represent the size of the output feature, and c represents the number of channels of the feature). The second level is the ConvLSTM layer, which combines the temporal and spatial features of the data to achieve simultaneous extraction of spatiotemporal features. The third level is comprised of a fully connected layers, which receives the output of ConvLSTM and completes the time series prediction of the final prediction result $\{\hat{y}_{t+1}, \cdots, \hat{y}_{t+j}, \cdots, \hat{y}_{t+n}\}$. The framework of the RCL-Learning model is shown in Fig. 5.

4.2. ResNet

This study uses the inherent advantages of ResNet to extract the spatial correlation features of pollutant concentration and meteorological data in multiple cities. First, air pollutant and meteorological data are input into the ResNet in time series order $\{x_t, \cdots, x_{t-i}, \cdots, x_{t-r+1}\}$ for spatial correlation feature extraction. Then, each convolutional layer in the ResNet performs feature extraction on the input data with a different convolution kernel. Finally, the features extracted by the ResNet are output in time series order $\{out_t, \cdots, out_{t-i}, \cdots, out_{t-r+1}\}$.

At the foundation of the RCL-Learning model, the ResNet constructed in this study is based on the reconstruction unit as a unit to reconstruct the traditional CNN. Each group of reconstruction units is represented on the left side of Fig. 6, which is composed of multiple convolution layers (generally no fewer than two layers) and a shortcut that uses multi-layer convolutional layers to asymptotically approach the residual function. The training process of each

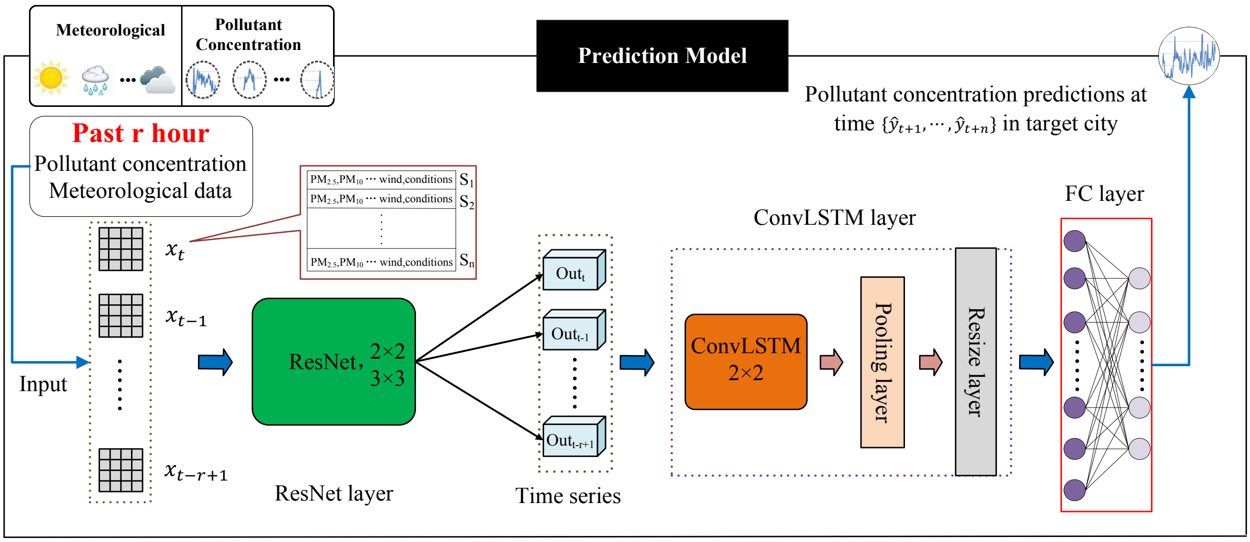


Fig. 5: Framework of the RCL-Learning model for $PM_{2.5}$ concentration prediction. x_{t-i} is the multi-city pollutant concentration and meteorological data input into the model at each moment, S_k represents the city k.

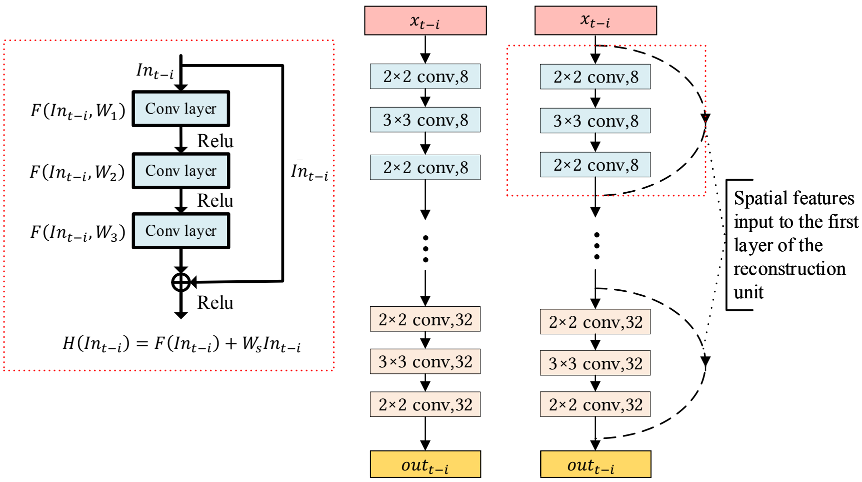


Fig. 6: Left: Reconstruction unit. Middle: Traditional CNN. Right: Residual network. 2×2 and 3×3 represents the filter size; 8 and 32 indicate the number of channels.

reconstruction unit of the residual network is shown in the following formula:

$$F\left(In_{t-i}\right) := H\left(In_{t-i}\right) - In_{t-i} \tag{1}$$

where $F(In_{t-i})$ is the residual function to learn, $:=$ is approximately equal, $H\left(In_{t-i}\right)$ (underlying mapping) is a spatial feature mapping function constructed by several convolutional layers and a shortcut connection, and In_{t-i} represents the spatial features of multi-city pollutant concentration and meteorological data entered in the first layer of the reconstruction unit at $t-i$ moment. The output of each reconstruction unit is as shown in formula (2),

$$H\left(In_{t-i}\right) = F\left(In_{t-i}\right) + W_s * In_{t-i} \tag{2}$$

where $F\left(In_{t-i}\right)$ can be represented by formula (3) ('$*$' is a convolution operation, b represents bias, δ is a $ReLU$ function, and W represents the filter for each convolutional layer). The addition of $F\left(In_{t-i}\right)$ and In_{t-i} is that of the corresponding elements of the two feature maps in each channel, and then the spatial features of the pollutant concentration and meteorological data extracted by the unit and the spatial features extracted by the previous unit are added. This can reduce the loss of important feature information, and also avoid network degradation and vanishing gradients. W_s is used to solve the dimension matching problem between In_{t-i} and $F\left(In_{t-i}\right)$.

$$F\left(In_{t-i}\right)=\delta\left(W * In_{t-i}+b\right) \tag{3}$$

By training the model, the value of the residual function $F\left(In_{t-i}\right)$ in formula (1) will asymptotically approach zero. Thus, formula (1) can be approximated as the identity mapping of $H\left(In_{t-i}\right)=In_{t-i}$ until the entire model converges. The output spatial feature out_{t-i} of the ResNet at each time series can be obtained by formula (4). Then the output value is input into the ConvLSTM in the time series.

$$out_{t-i}=\emptyset\left(H_1\left(In_{t-i}\right),\cdots,H_l\left(In_{t-i}\right),\cdots,H_h\left(In_{t-i}\right)\right) \tag{4}$$

where h represents the number of network reconstruction units, $H_l\left(In_{t-i}\right)$ denotes the spatial features of each reconstruction unit of output, and $\emptyset$ is the calculation function of the entire ResNet. Through the above calculation process, we can deeply extract the spatial correlation features $\left\{out_t,\cdots,out_{t-i},\cdots,out_{t-r+1}\right\}$ and use it as input to the ConvLSTM.

4.3. ConvLSTM

After the spatial feature extraction of the ResNet part, the high dimensional spatial feature sequence is obtained. This study uses the advantages of ConvLSTM to perform temporal and spatial association feature extraction on time series data $\left\{out_t,\cdots,out_{t-i},\cdots,out_{t-r+1}\right\}$, which can be divided into two stages: spatiotemporal feature extraction and $PM_{2.5}$ concentration prediction. In the spatiotemporal feature extraction stage, ConvLSTM performs spatiotemporal correlation feature extraction on the input data $\left\{out_t,\cdots,out_{t-i},\cdots,out_{t-r+1}\right\}$ to prepare for prediction. In the prediction stage, ConvLSTM inputs the output state h_{t+j} at each moment to the fully connected layers to generate $PM_{2.5}$ predicted value according to the extracted spatiotemporal correlation feature h_t. In the ConvLSTM training process, a single layer architecture ConvLSTM is used.

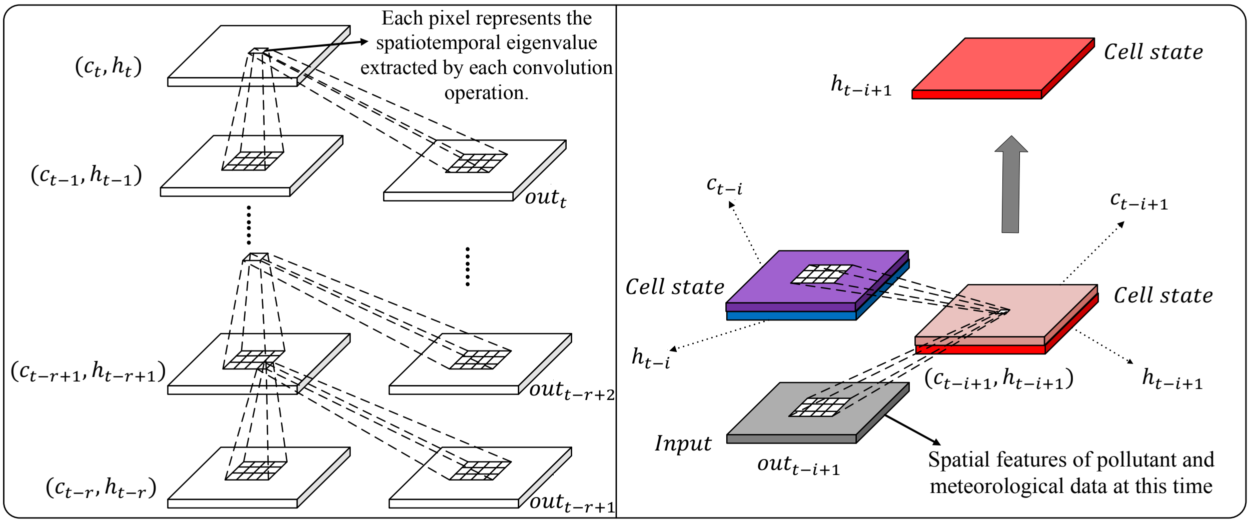


Fig. 7: Implementation of ConvLSTM. Left: Spatiotemporal feature extraction process of ConvLSTM. Right: extraction and generation process of spatiotemporal features at one time.

As shown in Fig. 7, we show the detailed process of ConvLSTM's complete spatiotemporal feature extraction, where (c_{t-i}, h_{t-i}) indicates the cell state. Assume that i, f and o respectively represent the input gate, forge gate and output gate, $W_{\cdot}$ represents the convolution kernel, b represents the bias, and '$\circ$' denotes the Hadamard product.

The spatiotemporal feature extraction process at each time series of ConvLSTM can be expressed by the following formulas:
(1) ConvLSTM selectively forgets the feature information of the cell state at time $t-i+1$,

$$\left\{\begin{array}{c} f_{t-i+1}=\sigma\left(W_f * out_{t-i+1}+W_f * h_{t-i}+W_f * c_{t-i}+b_f\right) \\ c_{t-i+1}^{\prime}=f_i \circ c_{t-i} \end{array}\right. \tag{5}$$

We need to selectively forget the information in the memory unit c_{t-i}. Therefore, we choose the $sigmoid$ function as the activation function of the forget gate. By multiplying the memory unit c_{t-i} and the forget gate f_{t-i+1}, part of the memory information is forgotten.
(2) ConvLSTM selects important information from the input feature out_{t-i+1} that is used to update the memory unit $c_{t-i+1}^{\prime}$,

$$\left\{\begin{array}{c} \tilde{c}_{t-i+1}=tanh\left(W_{\tilde{c}} * out_{t-i+1}+W_{\tilde{c}} * h_{t-i}+W_{\tilde{c}} * c_{t-i}+b_{\tilde{c}}\right) \\ i_{t-i+1}=\sigma\left(W_i * out_{t-i+1}+W_i * h_{t-i}+W_i * c_{t-i}+b_i\right) \\ c_{t-i+1}=c_{t-i+1}^{\prime}+i_{t-i+1} \circ \tilde{c}_{t-i+1} \end{array}\right. \tag{6}$$

In the above formulas, $\widetilde{c}_{t-i+1}$ represents the initial feature used to update the information of the memory unit $c_{t-i+1}^{\prime}$. The function of the input gate i_{t-i+1} is mainly to assign different weight values to the elements in each dimension of $\widetilde{c}_{t-i+1}$ and to select the important feature information for updating the memory unit $c_{t-i+1}^{\prime}$.
(3) Finally, it determines what ConvLSTM must output,

$$\left\{\begin{array}{c} o_{t-i+1}=\sigma\left(W_o * out_{t-i+1}+W_o * h_{t-i}+W_o * c_{t-i}+b_o\right) \\ h_{t-i+1}=o_{t-i+1} \circ \tanh\left(c_{t-i+1}\right) \end{array}\right. \tag{7}$$

The $sigmoid$ function of output gate o_{t-i+1} is mainly to assign different weight values to the elements in each dimension of c_{t-i+1} and to select the important feature information for output $h_{t-i+1} \in R^{d*e*c}$, where d and e represent the size of the final output state, and c represents the number of channels of the output state.

In the prediction stage, the entire working process of ConvLSTM is the same as the above process; that is, repeat the work of formulas (5), (6), and (7). The only difference from the first stage is that the input of ConvLSTM is the output at the last moment $\hat{h}_{t+j-1}$ at time $t+j$, as shown in formula (8). The initial value of $\hat{h}_t$ is the output state h_t at time t in the first stage. The $PM_{2.5}$ concentration prediction process of ConvLSTM at each moment can be expressed by the following formula:

$$\left\{\begin{array}{c} \hat{h}_{t+j}=tanh\left(W_{\hat{h}} * h_{t+j-1}\right) \\ h_{t+j}^{\prime}=resize\left(ave\_pool\left(\hat{h}_{t+j}\right)\right) \\ \hat{y}_{t+j}=W_{\hat{y}} h_{t+j}^{\prime} \end{array}\right. \tag{8}$$

where $W_{\hat{y}}$ represents the weight parameters of the fully connected layers, ave_pool and $resize$ represent average pooling and reshape operations, respectively, which are mainly used for feature dimensionality reduction. The output of the ConvLSTM is $\hat{h}_{t+j}$. After feature dimensionality reduction, we enter the calculation result $h_{t+j}^{\prime}$ into the fully connected layers to generate the predicted value $\hat{y}_{t+j}$.

4.4. Loss function

In the RCL-Learning model, the loss function is used to measure the degree of inconsistency between the predicted values $\hat{y}$ and the observed values y. The loss function is given in (9):

$$loss=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{n}}+\frac{\lambda}{2}\|W\|^2 \tag{9}$$

where n is the length of the predicted sequence, y_i denotes the observed value of the $PM_{2.5}$ concentration, and $\hat{y}_i$ is the predicted value of the $PM_{2.5}$ concentration, where λ is the regularization parameter, and W is the weight parameter of the network.

4.5. Metrics

The RCL-Learning model presented in this study was compared with other prediction models on the same dataset. Root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (Corr) were used as metrics to confirm the effectiveness of the proposed method. Experimental metrics were calculated by the following formulas:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{T}\left(y_i - \hat{y}_i\right)^2}{T}} \tag{10}$$

$$MAE = \frac{1}{T}\sum_{i=1}^{T}\left|y_i - \hat{y}_i\right| \tag{11}$$

$$Corr = \frac{cov\left(y, \hat{y}\right)}{\sqrt{var[y] * var[\hat{y}]}} \tag{12}$$

where y_i is the observed value, $\hat{y}_i$ denotes the predicted value, T is the test set size, $cov(y, \hat{y})$ is the covariance of y and $\hat{y}$, and $var[y]$ and $var[\hat{y}]$ represent the variance of y and $\hat{y}$, respectively.

5. Results

5.1. Parameter setting

The setting of hyperparameters in this study is based on the results of many experiments, leading to the final selection of the optimal set of hyperparameters. The validation set used in this study is closely related to the training stage, and after each epoch, the RMSE and MAE of the prediction model on the validation set are calculated. Therefore, the optimal model is selected based on the model error calculated on the validation set. The specific process is as follows: for each experiment, the number of epochs selected was 100. After training an epoch, we tested the trained model on the validation set. If the RMSE and MAE of the prediction model on the validation set became smaller, we updated and saved the model parameters. After many parameter adjustments and experiments, when the prediction effect of the prediction model on the validation set was optimal, the training ended. In the experiment, dropout was used as a general trick to avoid model overfitting. After the experiments, the parameters used for model testing are shown in Table 3. The implementation codes of our proposed RCL-Learning model and the comparison models are open source; please refer to our personal GitHub homepage [2] or code capsule homepage [3].

5.2. Single-step prediction

In the single-step prediction experiment, the input model data time series length r to 3, and the predicted length n to 1. Our predictor variable is the hourly $PM_{2.5}$ concentration in Shanghai, and the goal of this task is to predict the $PM_{2.5}$ concentration in the next hour. For example, we use the historical three hours multi-city pollutant concentration and meteorological data from 6:00-9:00 to predict the $PM_{2.5}$ concentration of Shanghai at 10:00 in the next hour.

5.2.1. The impact of related factors on $PM_{2.5}$ concentration prediction

In the study of $PM_{2.5}$ concentration prediction, different input variables may have different impacts on the prediction results of the RCL-Learning model. Table 4 lists the performance of $PM_{2.5}$ concentration prediction under different variable pairs, that is, the combination of $PM_{2.5}$ and other variables, but it is worth noting that our input still includes 14 cities. From Table 4, we can see that those different input variables positively impact the $PM_{2.5}$ concentration prediction. Among them, NO_2, O_3, SO_2, humidity, wind speed, and conditions have the most significant impact. This result corresponds to the analysis results in section 3.3.1; that is, the hidden correlation among data variables impacts our $PM_{2.5}$ concentration prediction research.

Similarly, different cities may have different impacts on the prediction results of the RCL-Learning model. Table 5 lists the performance of $PM_{2.5}$ concentration prediction under different city pairs, that is, the combination of Shanghai

[2] https://github.com/zouguojian/RCL-Learning
[3] https://codeocean.com/capsule/6299493/tree

Table 3
Model parameters

Layer name	Output size	Parameters	Values
ResNet	7×8×32	[filter,channel]×number of layers	[2×2, 8] ×1 [3×3, 8] ×1 [2×2, 8] ×1 [2×2, 16]×1 [3×3, 16]×1 [2×2, 16]×1 [2×2, 32]×1 [3×3, 32]×1 [2×2, 32]×1 [2×2, 32]×1 [3×3, 32]×1 [2×2, 32]×1
ConvLSTM	512	[filter,channel]×number of layers	[2×2, 32]×1
Full connected layer	256 128 1	[layer nodes]×number of layers	[256]×1 [128]×1 [1]×1
-	-	Batch size	64
-	-	Dropout	0.5
-	-	Learning rate	0.001
-	-	Epochs	100
-	-	λ	0.0001
-	-	Moving step size S	1
-	-	Training method	SGD

and other neighboring cities, but it is worth noting that our input still includes 16 variables. We can see from Table 5 that different neighboring cities positively impact Shanghai's $PM_{2.5}$ concentration prediction. The significance of the impact is related to distance, among which Nangtong, Wuxi, Hangzhou, Huzhou, and Zhoushan have the most significant impacts. This result corresponds to the analysis results in section 3.3.2, reflecting the importance of spatial features. According to the experimental results in Table 4 and Table 5, we obtain that the $PM_{2.5}$ concentration prediction is affected by pollutants and meteorological factors and by surrounding cities. Therefore, we use 16 pollutants and meteorological factors from 14 cities in the following experiments.

5.2.2. Comparison with state-of-the-art methods

Table 6 lists the [3-1 hour] $PM_{2.5}$ concentration prediction performance of our proposed RCL-Learning model and the baseline models on the whole test set.

Fig. 8 shows the generalization ability of different models on the same test set in the [3-1 hour] task. The length of Fig. 8's x-axis is 4000 hours, which means that 4000 consecutive hours were randomly selected in the test set to test the performance of the prediction model in this time period. This verification method is based on the method given in previous research papers (Huang & Kuo, 2018; Park et al., 2018), and the main purpose is to visualize the prediction effect of the model and highlight the prediction performance and fitting ability of the model. We combine the prediction of pollutant with the change of AQI, and describe the location of mutation points more scientifically through AQI. According to the description of (Yi et al., 2018b), when the AQI value fluctuates sharply, the mutation point appears. Therefore, we combine the test results with the mutation points to further verify the superiority of our RCL-Learning model. The blue curve represents the observed value, the red curve represents the predicted value and the yellow curve represents the AQI value. Owing to space considerations in this study, Fig. 8 only shows the experimental results of the four state-of-the-art prediction models, representing the fitting trends of the ConvLSTM, Att-ConvLSTM, GC-LSTM, and RCL-Learning models were tested on the whole test set.

To demonstrate the predictive performance of the RCL-Learning model we chose, we compared it with the latest research results. We selected four prediction models, including the proposed RCL-Learning. Fig. 9 depicts the predic-

Table 4
The impacts of pollutants and meteorological factors on the performance of $PM_{2.5}$ concentration prediction

Input factor	RMSE	MAE	Corr
PM2.5	5.823	4.416	0.992
(PM2.5 & AQI)	5.763	4.379	0.992
(PM2.5 & PM10)	5.721	4.342	0.992
(PM2.5 & SO2)	**5.626**	**4.224**	**0.993**
(PM2.5 & NO2)	**5.519**	**4.051**	**0.993**
(PM2.5 & O3)	**5.551**	**4.160**	**0.993**
(PM2.5 & CO)	5.801	4.409	0.992
(PM2.5 & temperature)	5.710	4.320	0.992
(PM2.5 & humidity)	**5.640**	**4.240**	**0.993**
(PM2.5 & air pressure)	5.722	4.344	0.993
(PM2.5 & wind direction)	5.679	4.272	0.993
(PM2.5 & wind speed)	**5.641**	**4.243**	**0.993**
(PM2.5 & clouds)	5.800	4.408	0.992
(PM2.5 & maximum temperature)	5.713	4.321	0.993
(PM2.5 & minimum temperature)	5.713	4.323	0.992
(PM2.5 & conditions)	**5.673**	**4.269**	**0.993**

Table 5
The impacts of neighboring cities on the performance of $PM_{2.5}$ concentration prediction

City pair	RMSE	MAE	Corr
Shanghai	9.806	7.043	0.985
(Shanghai & Nanjing)	9.782	6.882	0.983
(Shanghai & Suzhou)	8.292	6.139	0.987
(Shanghai & Nantong)	**7.773**	**5.536**	**0.991**
(Shanghai & Wuxi)	**8.240**	**5.752**	**0.989**
(Shanghai & Changzhou)	9.250	6.673	0.985
(Shanghai & Zhenjiang)	9.574	6.761	0.985
(Shanghai & Hangzhou)	**7.723**	**5.784**	**0.989**
(Shanghai & Ningbo)	8.818	6.147	0.986
(Shanghai & Shaoxing)	9.629	6.896	0.983
(Shanghai & Huzhou)	**7.221**	**5.366**	**0.990**
(Shanghai & Jiaxing)	8.880	6.475	0.964
(Shanghai & Taizhou)	9.805	7.042	0.983
(Shanghai & Zhoushan)	**7.020**	**4.784**	**0.992**

tion performance of different prediction models on the test set in the [3-1 hour] task. The x-axis represents the observed value of $PM_{2.5}$ and the y-axis represents the predicted value of $PM_{2.5}$. The black line indicates the $y = \hat{y}$ function, and the black dots indicate the degree of deviation between the observed and predicted values. In the dispersion comparison, when the concentration of $PM_{2.5}$ is greater than 100 $\mu g/m^3$, the dispersion of ConvLSTM is the largest, and that of RCL-Learning model is the smallest, meaning that the prediction performance is the best. When the values of $PM_{2.5}$ are between 0 $\mu g/m^3$ and 100 $\mu g/m^3$, the dispersion degree of RCL-Learning model is still the smallest. Fig. 9 shows that the RCL-Learning predicted values are generally consistent with the observed values. In the correlation comparison, in the whole test set, the correlation coefficients of ConvLSTM, Att-ConvLSTM, GC-LSTM, and RCL-Learning are 0.970, 0.975, 0.974, and 0.993, respectively, which means that the correlation between predicted values and observed values of RCL-Learning is the largest.

5.3. Multi-step prediction

The existing $PM_{2.5}$ prediction research mainly focuses on single-step prediction for the next time, which may not be sufficient to meet the needs of actual application scenarios. Therefore, the significance of multi-step $PM_{2.5}$

Table 6
Performance comparison of all models for the [3-1 hour] task

Model	RMSE	MAE	Corr
CAMx (Zhu et al., 2019)	34.454	-	0.712
CMAQ (Chen et al., 2014)	34.087	-	0.708
NAQPMS (Wang et al., 2001)	36.649	-	0.690
WRF-Chen (Saide et al., 2011)	37.316	-	0.683
SVM (Suleiman et al., 2019)	25.820	18.415	0.858
SVR (Yang et al., 2018)	23.564	16.001	0.869
MLR (Feng et al., 2020)	19.023	13.284	0.934
HMM (Sun et al., 2013)	22.361	15.034	0.882
XGBoost (Zamani Joharestani et al., 2019)	21.090	14.964	0.887
BP Chen et al. (2019)	18.915	13.043	0.934
RNN (Chang-Hoi et al., 2021)	15.932	11.715	0.962
LSTM (Karim & Rafi, 2020)	15.721	9.895	0.966
ConvLSTM (Le et al., 2020)	13.041	9.234	0.970
CNN-LSTM (Huang & Kuo, 2018; Qin et al., 2019)	11.366	8.221	0.974
GC-LSTM (Qi et al., 2019)	10.478	7.503	0.975
Att-ConvLSTM (Xu & Lv, 2019)	11.476	8.321	0.974
ResNet-LSTM	10.320	7.087	0.975
CNN-ConvLSTM	9.587	6.842	0.983
RCL-Learning	**5.478**	**3.897**	**0.993**

concentration prediction is self-evident. We divide the future 1-15 hours into six multi-step prediction tasks (1-1, 1-2, 1-3, 1-6, 1-8, and 1-15 hours) and trained separate models to predict the $PM_{2.5}$ concentration of each task. In each task, we use the historical pollutant concentration and meteorological data of multiple cities to achieve a multi-step prediction of the future $PM_{2.5}$ concentration of the target city, as shown in Fig. 10. Table 7 lists the performance of the RCL-Learning model in six multi-step prediction tasks. In the experiment involving multi-step prediction, we used the fixed network structure RCL-Learning prediction model in all tasks. The prediction results are shown in Table 7.

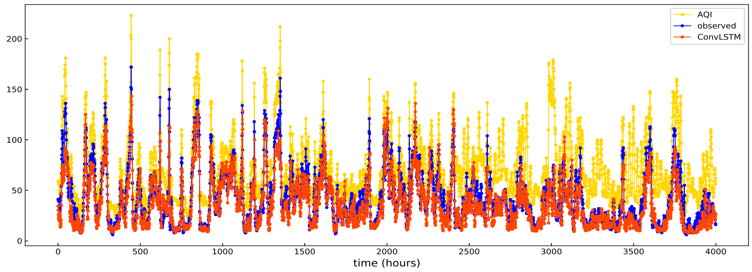
Table 7
$PM_{2.5}$ concentration predictions for multiple durations of time

Task	Historical time period	RMSE	MAE	Corr
1-1 h prediction	3 h	5.478	3.897	0.993
1-1 h prediction	5 h	5.449	3.897	0.993
1-2 h prediction	5 h	9.016	6.075	0.982
1-3 h prediction	10 h	13.622	8.818	0.963
1-6 h prediction	15 h	25.320	17.395	0.907
1-8 h prediction	20 h	31.592	21.803	0.796
1-15 h prediction	20 h	40.376	30.176	0.607

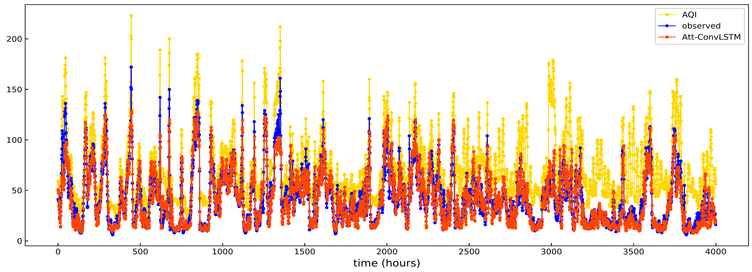
As shown in Table 7, as the prediction time interval increases, the required historical input time series also increases. The prediction performance of the RCL-Learning model gradually decreases with the increase of the prediction step, and RMSE increases from 5.449 to 40.376. In Table 7, for the next one hour's $PM_{2.5}$ concentration prediction, increasing the length r of the historical input time series from three to five hours can indeed improve the accuracy of the prediction, but the accuracy of the prediction RMSE is only improved by 0.029. Considering the improvement of both the prediction accuracy and the calculation cost of the prediction, for the prediction of pollutants at any time, after many experiments, we chose an optimal historical input time series length r.

5.4. Trend prediction

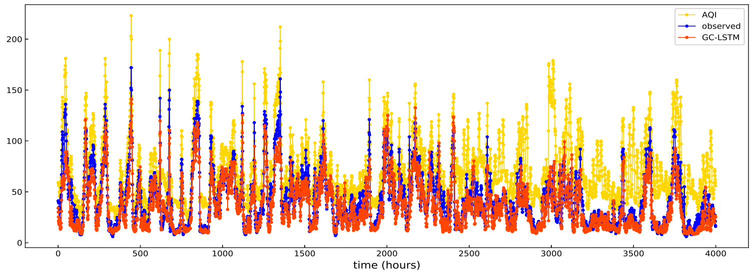
To further confirm the validity of the proposed prediction model, we used the historical 20 hours of multi-city pollutant and meteorological data as input to predict trends in pollutant $PM_{2.5}$ concentrations over the different periods:



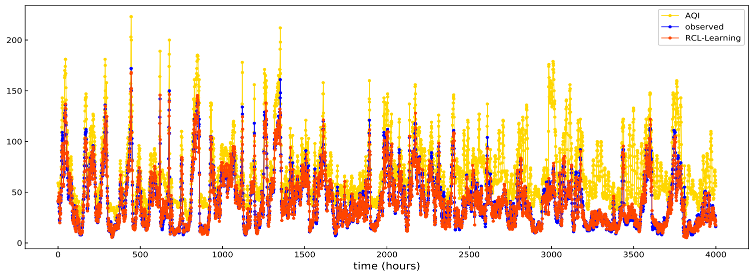
(a)



(b)



(c)



(d)

Fig. 8: Fitting trends of the different models in the [3-1 hour] task. (a)–(d) represent the fitting trends of ConvLSTM, Att-ConvLSTM, GC-LSTM, and RCL-Learning models.

1-10, 11-20, 21-30, and 31-40 hours. We compared the RCL-Learning model with the state-of-the-art $PM_{2.5}$ prediction models: ConvLSTM, Att-ConvLSTM, and GC-LSTM (Qi et al., 2019; Le et al., 2020; Xu & Lv, 2019). Table 8 shows the average error of $PM_{2.5}$ concentration prediction values over the different periods.

To further demonstrate the fitting performance of RCL-Learning on the test set, we predicted the pollutant concentration in the future 1-20 hours and 1-40 hours. Fig. 11 shows the predicted and observed changes in $PM_{2.5}$ over the different periods (with randomly selected samples from different periods on the test set).

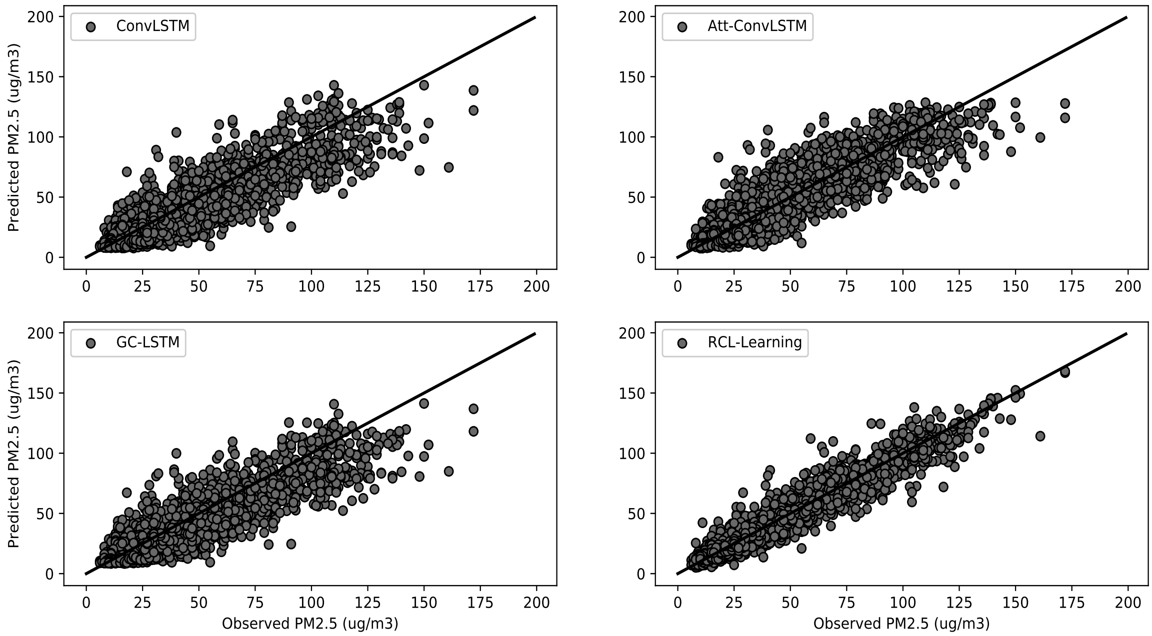


Fig. 9: Degree of fit between the observed and predicted values on the test set in the [3-1 hour] task.

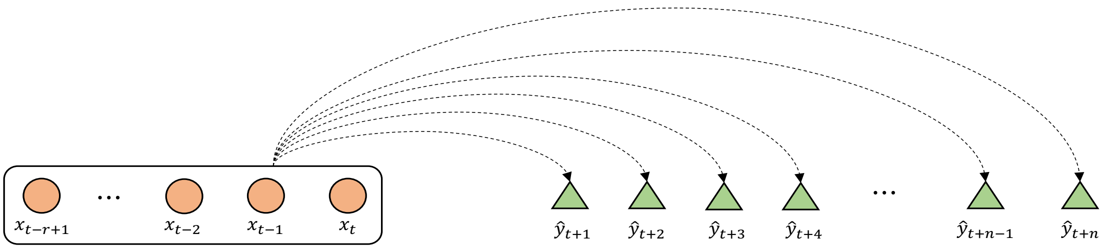


Fig. 10: Illustration of the multi-step prediction.

Table 8
$PM_{2.5}$ concentration prediction over different periods in the future

Model	1-10 h	11-20 h	21-30 h	31-40 h
ConvLSTM (Le et al., 2020)	23.042	36.887	43.874	52.569
Att-ConvLSTM (Xu & Lv, 2019)	22.053	36.010	43.146	52.264
GC-LSTM (Qi et al., 2019)	20.865	34.346	42.532	51.867
RCL-Learning	**17.678**	**31.213**	**40.109**	**48.765**

6. Discussion

6.1. Comparison with previous prediction models

Table 6 shows that, compared with the four single models and eight traditional methods, the CNN-LSTM, GC-LSTM, Att-ConvLSTM, ResNet-LSTM, CNN-ConvLSTM, and RCL-Learning have better predictive results because all six can better handle long-term sequence dependency problems with spatial features. The RMSE for these models ranges from 5.478 to 11.476, MAE ranges from 3.987 to 8.321, and Corr ranges from 0.974 to 0.993. Comparing the prediction results in Table 6 of ResNet-LSTM, GC-LSTM, and CNN-LSTM, the prediction accuracy of ResNet-LSTM is higher than that of GC-LSTM and CNN-LSTM, which proves that deep ResNet has better spatial feature

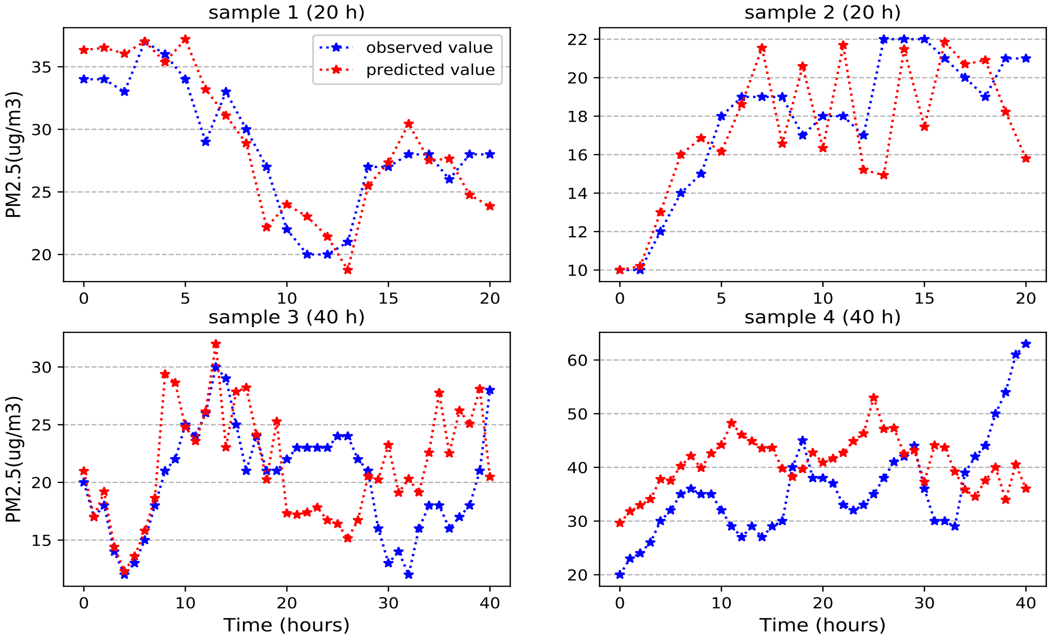


Fig. 11: Prediction of target city pollutant concentration trends over the different periods. The blue curve represents the observed values, the red curve represents the predicted values.

ability to extract pollutant and meteorological data than deep CNN and graph convolutional (GC) neural network. Its RMSE, MAE, and Corr attain the optimal values of 10.320, 7.087, and 0.975, respectively. Next, Comparing the prediction results in Table 6 of Att-ConvLSTM and RCL-Learning, the prediction accuracy of RCL-Learning is higher than that of Att-ConvLSTM, which proves that deep ResNet has better spatial feature ability to extract pollutant and meteorological data than spatiotemporal attention method (Att). Finally, by comparing the results in Table 6 of the CNN-LSTM and CNN-ConvLSTM experiments, and comparing those of ConvLSTM and LSTM, it can be proved that ConvLSTM has better spatiotemporal feature extraction ability for long-term sequences than LSTM. However, using only the ConvLSTM model to extract the temporal and spatial features of the complex pollutant and meteorological data, it is difficult not only to filter the redundant information in these data, but also to deeply extract the spatiotemporal features of time series. Therefore, this study combines the advantages of ResNet and ConvLSTM, and proposes a new type of prediction framework: the RCL-Learning model. The experimental results of the RCL-Learning model in Table 6 also confirm that the combination of ResNet and ConvLSTM is very effective for the prediction of $PM_{2.5}$. The RMSE optimal value is only 5.478, and the MAE optimal value is 3.897.

In this paper, 4000 consecutive test samples were randomly selected and presented in the experiment in the form of graph, as shown in Fig. 8 and Fig. 9. Therefore, our focus was on the fitting ability of the model to verify the supposition that RCL-Learning can better fit the mutation points. As shown in Fig. 8 and Fig. 9, when the $PM_{2.5}$ pollution source concentration is unstable, particularly when the concentration value is greater than 100 $\mu g/m^3$, the prediction results of the comparison models could not follow the actual trend and showed a rather disordered pattern. This also reflects the fact that, in terms of the current $PM_{2.5}$ concentration prediction task, it is still difficult for the model to make accurate predictions. Furthermore, the predictions and observations of the proposed RCL-Learning model are almost coincident and have a good fitting effect on the mutation of $PM_{2.5}$ concentration, such as the 46th hour, 165th hour, 288th hour, 444th hour, etc., as shown in Fig. 8.

Combining the fitting ability of each model in Fig. 8 and Fig. 9, we reach the following conclusions: (1) For the Fig. 8, we can get that the prediction performance of the RCL-Learning model is better than the comparison models, and it is suitable for prediction tasks with sudden changes in pollutant concentration; (2) For the Fig. 9, we can get that compared with the comparison models, RCL-Learning model can accurately predict high concentrations of $PM_{2.5}$, so that the predicted value and the observed value are highly consistent; (3) Combining the experimental results in Fig. 8 and Fig. 9, we can intuitively see that for mutation points, the $PM_{2.5}$ concentration is generally relatively high, and the number of mutation points is relatively small. This mainly reflects that in the general data set, the number of samples at mutation points is small, which leads to the problem of uneven data distribution. This phenomenon has caused the problem of insufficient learning of the predictive model, that is, it is difficult to learn the changing regularity of pollutant concentration under sudden changes. Therefore, this is also the reason why some models are difficult to fit in the case of sudden pollutant concentration.

Based on the above experimental results, our analysis result is that the RCL-Learning model proposed in this paper tightly grasps the spatiotemporal characteristics of pollutants. In terms of data, we consider the impact of pollutants and meteorological factors in multiple cities on the target city in the pollutant concentration prediction task; In terms of the model, we utilize the residual network and ConvLSTM as the spatiotemporal feature extractor, and make full use of the advantages of the two networks in feature extraction. Therefore, the characteristics of our prediction model are as follows: on the one hand, in large samples D_1 (the number of samples D_1 with a pollutant concentration less than 100 $\mu g/m^3$ is 94.3% of the total number of training samples) with small vibration amplitude of pollutant concentration, the changing regularity of pollutant concentration in historical data can be fully learned; on the other hand, in small samples D_2 (the number of samples D_2 with a pollutant concentration greater than 100 $\mu g/m^3$ is 5.7% of the total number of training samples) with large fluctuations of pollutant concentration, we utilize the advantages of the RCL-Learning model to learn the changing regularity of pollutant concentration in the target city and neighboring cities, which can solve the problem that it is difficult to accurately predict the mutation of pollutants in the target city. The ability of the RCL-Learning model to predict $PM_{2.5}$ concentration is verified in this experiment.

6.2. Long-term series prediction and model comparison

Making long-term predictions requires the input of historical pollutant concentration and meteorological data with a very high correlation, and the length of the input sequence is difficult to determine. As a result, ensuring prediction accuracy without the task being overly time-consuming is difficult. However, this study proposes that the RCL-Learning model can achieve both. Therefore, the length of the time series of the input data is mainly based on the amount of time spent in training the model and the improvement of the prediction accuracy. As the prediction time period increases, we gradually increase the length of the input sequence, and the longest threshold we set is 20. Because when the input time series length of the model is greater than 20, the time spent to train the model will rise sharply. Therefore, the data sequence length is an empirical value, which is set according to the experience of each researcher.

The RCL-Learning model can predict the concentration of pollutants in the target city in the near future, as shown in Table 7. When predicting the concentration of pollutants in the target city within the next three hours, the RMSE value can be maintained between 5.449 and 13.622. For longer-term sequence prediction tasks, such as predicting the concentration of pollutants in the target cities in the next 1 to 15 hours, the RCL-Learning prediction model also shows satisfactory performance. The average RMSE value reaches 22.927, and the average Corr value reaches 0.800.

As shown in Table 8, when we compare the average prediction errors of the ConvLSTM, Att-ConvLSTM, GC-LSTM and RCL-Learning models for different prediction time periods, the prediction errors of ConvLSTM, Att-ConvLSTM, and GC-LSTM are larger than the RCL-Learning model, meaning that RCL-Learning has the highest prediction accuracy. Fig. 11 shows the performance of the RCL-Learning model in predicting pollutant trends for the 20- and 40-hour time periods, that is, four test samples were randomly selected from the test set as the reference basis for the analysis of the experimental results. From Fig. 11 we can see that the trends indicated by the blue observation curve and the red prediction curve are consistent. The experiments verified that for the long-term prediction of pollutant concentration, the trend predicted of RCL-Learning model has wide application value. Therefore, we can improve the accuracy of pollutant prediction by considering combining the trend of pollutant concentration predicted by the RCL-Learning model with the state-of-the-art prediction methods.

7. Conclusions

Based on the combination of deep learning and big data correlation principles, we propose in this paper an RCL-Learning prediction model based on ResNet and ConvLSTM. The model is mainly used to predict the concentration of pollutants in target cities. ResNet is primarily employed to extract the spatial features of pollutant and meteorological data in multiple cities. ConvLSTM is used to extract the spatiotemporal features of high-dimensional data output by the ResNet layer. The advantages of the proposed method are summarized as follows.

(1) Compared with the traditional CNN, GC, and Att network, ResNet can better extract the spatial features in the same depth of the network situation.

(2) Because of the temporal and spatial correlation of air pollutants, compared with the traditional LSTM, the prediction model proposed in this paper adds the ConvLSTM layer on the basis of ResNet. ConvLSTM extracts the spatiotemporal correlation features of the data more effectively.

Experiments showed that, compared with the other models, the RCL-Learning model made more accurate predictions by fully extracting the correlation of pollutant and meteorological data, and it solved other problems, such as long-term dependency. Moreover, it fully considered the spatiotemporal correlation of pollutant and meteorological data. According to the correlation between the cumulative matter ($PM_{2.5}$) and the air quality index (AQI), the ability to accurately prediction $PM_{2.5}$ is very important in warning of pollutant hazards. Compared with the traditional machine learning methods and single classical network, the RCL-Learning has become one of the practical auxiliary models in the tasks of monitoring and predicting air pollution at the regional and national levels.

The limitation of this study is that the location information of multiple cities did not make a significant contribution to the prediction of pollutants at the target city were not considered. Accordingly, the location information will be added as the input feature to the prediction model in future work.

Acknowledgement

This work was partially supported by Natural Science Foundation of Shanghai (18ZR1428300), National Natural Science Foundation of China (61572326, 61802258, 61702333), the Shanghai Committee of Science and Technology (17070502800, 16JC1403000).

References

Akimoto, H. (2003). Global air quality and pollution. *Science*, *302*, 1716–1719.

Cairncross, E. K., John, J., & Zunckel, M. (2007). A novel air pollution index based on the relative risk of daily mortality associated with short-term exposure to common air pollutants. *Atmospheric environment*, *41*, 8442–8454.

Chang-Hoi, H., Park, I., Oh, H.-R., Gim, H.-J., Hur, S.-K., Kim, J., & Choi, D.-R. (2021). Development of a pm2. 5 prediction model using a recurrent neural network algorithm for the seoul metropolitan area, republic of korea. *Atmospheric Environment*, *245*, 118021.

Chen, J., Lu, J., Avise, J. C., DaMassa, J. A., Kleeman, M. J., & Kaduwela, A. P. (2014). Seasonal modeling of pm2. 5 in california's san joaquin valley. *Atmospheric environment*, *92*, 182–190.

Chen, Y., An, J. et al. (2019). A novel prediction model of pm2. 5 mass concentration based on back propagation neural network algorithm. *Journal of Intelligent & Fuzzy Systems*, *37*, 3175–3183.

Corani, G., & Scanagatta, M. (2016). Air pollution prediction via multi-label classification. *Environmental modelling & software*, *80*, 259–264.

Cordano, M., & Frieze, I. H. (2000). Pollution reduction preferences of us environmental managers: Applying ajzen's theory of planned behavior. *Academy of Management journal*, *43*, 627–641.

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In *European conference on computer vision* (pp. 184–199). Springer.

Feng, R., Gao, H., Luo, K., & Fan, J.-r. (2020). Analysis and accurate prediction of ambient pm2. 5 in china using multi-layer perceptron. *Atmospheric Environment*, *232*, 117534.

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., & Wang, J. (2015). Artificial neural networks forecasting of pm2. 5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmospheric Environment*, *107*, 118–128.

Fong, I. H., Li, T., Fong, S., Wong, R. K., & Tallon-Ballesteros, A. J. (2020). Predicting concentration levels of air pollutants by transfer learning and recurrent neural network. *Knowledge-Based Systems*, *192*, 105622.

Gu, K., Qiao, J., & Li, X. (2018). Highly efficient picture-based prediction of pm2. 5 concentration. *IEEE Transactions on Industrial Electronics*, *66*, 3176–3184.

Hao, Y., & Liu, Y.-M. (2016). The influential factors of urban pm2. 5 concentrations in china: a spatial econometric analysis. *Journal of Cleaner Production*, *112*, 1443–1453.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*, 1904–1916.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630–645). Springer.

Hossain, M., Rekabdar, B., Louis, S. J., & Dascalu, S. (2015). Forecasting the weather of nevada: A deep learning approach. In *2015 international joint conference on neural networks (IJCNN)* (pp. 1–6). IEEE.

Hu, J., Wang, Y., Ying, Q., & Zhang, H. (2014). Spatial and temporal variability of pm2. 5 and pm10 over the north china plain and the yangtze river delta, china. *Atmospheric Environment*, *95*, 598–609.

Huang, C.-J., & Kuo, P.-H. (2018). A deep cnn-lstm model for particulate matter (pm2. 5) forecasting in smart cities. *Sensors*, *18*, 2220.

Karim, R., & Rafi, T. H. (2020). An automated lstm-based air pollutant concentration estimation of dhaka city, bangladesh. *Int. J. Eng. & Inf. Sys*, *4*, 88–101.

Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating lstm with multiple garch-type models. *Expert Systems with Applications*, *103*, 25–37.

Kolehmainen, M., Martikainen, H., & Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. *Atmospheric Environment*, *35*, 815–825.

Le, V.-D., Bui, T.-C., & Cha, S.-K. (2020). Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 55–62). IEEE.

Lee, A., Szpiro, A., Kim, S., & Sheppard, L. (2015). Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics*, *26*, 255–267.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental pollution*, *231*, 997–1004.

Lin, Z., Li, M., Zheng, Z., Cheng, Y., & Yuan, C. (2020). Self-attention convlstm for spatiotemporal prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 11531–11538). volume 34.

Liu, J., Han, Y., Tang, X., Zhu, J., & Zhu, T. (2016). Estimating adult mortality attributable to pm2. 5 exposure in china with assimilated pm2. 5 concentrations based on a ground monitoring network. *Science of the Total Environment*, *568*, 1253–1262.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, .

Maleki, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Birgani, Y. T., & Rahmati, M. (2019). Air pollution prediction by using an artificial neural network model. *Clean technologies and environmental policy*, *21*, 1341–1352.

Martins, N. R., & Da Graca, G. C. (2018). Impact of pm2. 5 in indoor urban environments: A review. *Sustainable Cities and Society*, *42*, 259–275.

Mayer, H. (1999). Air pollution in cities. *Atmospheric environment*, *33*, 4029–4037.

McKinley, G., Zuk, M., Höjer, M., Avalos, M., González, I., Iniestra, R., Laguna, I., Martínez, M. A., Osnaya, P., Reynales, L. M. et al. (2005). Quantification of local and global benefits from air pollution control in mexico city.

Mokhtari, M., Miri, M., Mohammadi, A., Khorsandi, H., Hajizadeh, Y., & Abdolahnejad, A. (2015). Assessment of air quality index and health impact of pm10, pm2. 5 and so2 in yazd, iran. *Journal of Mazandaran University of Medical Sciences*, *25*, 14–23.

Park, S., Kim, M., Kim, M., Namgung, H.-G., Kim, K.-T., Cho, K. H., & Kwon, S.-B. (2018). Predicting pm10 concentration in seoul metropolitan subway stations using artificial neural network (ann). *Journal of hazardous materials*, *341*, 75–82.

Qi, Y., Li, Q., Karimian, H., & Liu, D. (2019). A hybrid model for spatiotemporal forecasting of pm2. 5 based on graph convolutional neural network and long short-term memory. *Science of the Total Environment*, *664*, 1–10.

Qin, D., Yu, J., Zou, G., Yong, R., Zhao, Q., & Zhang, B. (2019). A novel combined prediction scheme based on cnn and lstm for urban pm 2.5 concentration. *IEEE Access*, *7*, 20050–20059.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*, 91–99.

Russell, A. G., McCue, K. F., & Cass, G. R. (1988). Mathematical modeling of the formation of nitrogen-containing air pollutants. 1. evaluation of an eulerian photochemical model. *Environmental science & technology*, *22*, 263–271.

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. *Atmospheric Environment*, *45*, 2769–2780.

Sainath, T. N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A.-r., Dahl, G., & Ramabhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks. *Neural networks*, *64*, 39–48.

Sønderby, S. K., Sønderby, C. K., Nielsen, H., & Winther, O. (2015). Convolutional lstm networks for subcellular localization of proteins. In *International Conference on Algorithms for Computational Biology* (pp. 68–80). Springer.

Song, C., He, J., Wu, L., Jin, T., Chen, X., Li, R., Ren, P., Zhang, L., & Mao, H. (2017). Health burden attributable to ambient pm2. 5 in china. *Environmental pollution*, *223*, 575–586.

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Highway networks. *arXiv preprint arXiv:1505.00387*, .

Suleiman, A., Tight, M., & Quinn, A. (2019). Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (pm10 and pm2. 5). *Atmospheric Pollution Research*, *10*, 134–144.

Sun, W., Zhang, H., Palazoglu, A., Singh, A., Zhang, W., & Liu, S. (2013). Prediction of 24-hour-average pm2. 5 concentrations using a hidden markov model with different emission distributions in northern california. *Science of the total environment*, *443*, 93–103.

Tian, J., & Chen, D. (2010). A semi-empirical model for predicting hourly ground-level fine particulate matter (pm2. 5) concentration in southern ontario from satellite remote sensing and ground-based meteorological measurements. *Remote Sensing of Environment*, *114*, 221–229.

Wang, J., & Christopher, S. A. (2003). Intercomparison between satellite-derived aerosol optical thickness and pm2. 5 mass: Implications for air quality studies. *Geophysical research letters*, *30*.

Wang, J., Wang, S., Voorhees, A. S., Zhao, B., Jang, C., Jiang, J., Fu, J. S., Ding, D., Zhu, Y., & Hao, J. (2015). Assessment of short-term pm2. 5-related mortality due to different emission sources in the yangtze river delta, china. *Atmospheric Environment*, *123*, 440–448.

Wang, Y., Ying, Q., Hu, J., & Zhang, H. (2014). Spatial and temporal variations of six criteria air pollutants in 31 provincial capital cities in china during 2013–2014. *Environment international*, *73*, 413–422.

Wang, Z., Maeda, T., Hayashi, M., Hsiao, L.-F., & Liu, K.-Y. (2001). A nested air quality prediction modeling system for urban and regional scales: Application for high-ozone episode in taiwan. *Water, Air, and Soil Pollution*, *130*, 391–396.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, .

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (pp. 802–810).

Xu, Z., & Lv, Y. (2019). Att-convlstm: Pm 2.5 prediction model and application. In *The International Conference on Natural Computation, Fuzzy*

Systems and Knowledge Discovery (pp. 30–40). Springer.

Xue, F., Ji, H., Zhang, W., & Cao, Y. (2019). Attention-based spatial–temporal hierarchical convlstm network for action recognition in videos. *IET Computer Vision*, *13*, 708–718.

Yang, J., & Hu, M. (2018). Filling the missing data gaps of daily modis aod using spatiotemporal interpolation. *Science of the Total Environment*, *633*, 677–683.

Yang, W., Deng, M., Xu, F., & Wang, H. (2018). Prediction of hourly pm2. 5 using a space-time support vector regression model. *Atmospheric Environment*, *181*, 12–19.

Yi, J., Wen, Z., Tao, J., Ni, H., & Liu, B. (2018a). Ctc regularized model adaptation for improving lstm rnn based multi-accent mandarin speech recognition. *Journal of Signal Processing Systems*, *90*, 985–997.

Yi, X., Zhang, J., Wang, Z., Li, T., & Zheng, Y. (2018b). Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 965–973).

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). Pm2. 5 prediction based on random forest, xgboost, and deep learning using multisource remote sensing data. *Atmosphere*, *10*, 373.

Zhang, B., Zou, G., Qin, D., Lu, Y., Jin, Y., & Wang, H. (2021). A novel encoder-decoder model based on read-first lstm for air pollutant prediction. *Science of The Total Environment*, *765*, 144507.

Zhang, L., Zhu, G., Shen, P., Song, J., Afaq Shah, S., & Bennamoun, M. (2017). Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 3120–3128).

Zhixiang, M., Cai, C., Xiangwei, M., Wei, L., & Chuanzhen, Z. (2021). Short-term effects of different pm2. 5 thresholds on daily all-cause mortality in jinan, china, .

Zhu, G., Zhang, L., Shen, P., & Song, J. (2017a). Multimodal gesture recognition using 3-d convolution and convolutional lstm. *Ieee Access*, *5*, 4517–4524.

Zhu, J. Y., Sun, C., & Li, V. O. (2017b). An extended spatio-temporal granger causality model for air quality estimation with heterogeneous urban big data. *IEEE Transactions on Big Data*, *3*, 307–319.

Zhu, J. Y., Zhang, C., Zhang, H., Zhi, S., Li, V. O., Han, J., & Zheng, Y. (2017c). pg-causality: Identifying spatiotemporal causal pathways for air pollutants with urban big data. *IEEE Transactions on Big Data*, *4*, 571–585.

Zhu, Y.-Y., Gao, Y.-X., Liu, B., Wang, X.-Y., Zhu, L.-L., Xu, R., Wang, W., Ding, J.-N., Li, J.-J., & Duan, X.-L. (2019). Concentration characteristics and assessment of model-predicted results of pm2. 5 in the beijing-tianjin-hebei region in autumn and winter. *Huan jing ke xue= Huanjing kexue*, *40*, 5191–5201.