Predicting of air pollutant concentrations based on spatio-temporal attention convolutional LSTM networks

Peng Jiang[1], **Igor Bychkov**[4], **Jun Liu**[2,3], **Alexei Hmelnov**[4]

[1]Department of Science and Technology Cooperation, Westlake University, No.18, Shilong Mountain Street, Xihu District, Hangzhou, China

[2]"the Belt and Road" Institute for Information Technology, Hangzhou Dianzi University, No.115, Wenyi Road, Xihu District, Hangzhou, China

[3]School of Automation (Artificial Intelligence), Hangzhou Dianzi University, No.1158, Number Two Street, Jianggan District, Hangzhou, China

[4]Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences, 134 Lermontov st. Irkutsk, Russia

E-mail: `jiangpenghz@163.com`

Abstract. Forecasting of air pollutant concentration, which is influenced by air pollution accumulation, traffic flow and industrial emissions, has attracted extensive attention for decades. In this paper, we propose a spatio-temporal attention convolutional long short term memory neural networks (Attention-CNN-LSTM) for air pollutant concentration forecasting. Firstly, we analyze the Granger causalities between different stations and establish a hyperparametric Gaussian vector weight function to determine spatial autocorrelation variables, which is used as part of the input feature. Secondly, convolutional neural networks (CNN) is employed to extract the temporal dependence and spatial correlation of the input, while feature maps and channels are weighted by attention mechanism, so as to improve the effectiveness of the features. Finally, a depth long short term memory (LSTM) based time series predictor is established for learning the long-term and short-term dependence of pollutant concentration. In order to reduce the effect of diverse complex factors on LSTM, inherent features are extracted from historical air pollutant concentration data meteorological data and timestamp information are incorporated into the proposed model. Extensive experiments were performed using the Attention-CNN-LSTM, autoregressive integrated moving average (ARIMA), support vector regression (SVR), traditional LSTM and CNN, respectively. The results demonstrated that the feasibility and practicability of Attention-CNN-LSTM on estimating CO and NO concentration.

1. Introduction

In recent years, air pollution, especially the large-scale haze caused by ultrafine particles and volatile organic compounds (VOCs) of mobile pollution sources, has attracted worldwide attention [1],[2]. Ultrafine particles and VOCs not only harm to human health directly, but also is the important precursor of fine particulate matter (PM2.5) and major component of photochemical smog. Therefore, monitoring the emission of ultrafine particles and VOCs from mobile pollution sources is one of the effective means to reduce smog weather and can improve the quality of regional urban atmospheric environment. Knowing the source and concentration of these pollutants is essential to reduce the adverse effects of air pollution on health [3]. Thus, in the spatial dimension, the characteristics of pollutant changes in different regions are considered, and the differences and causes of pollutant concentration in time and space are analyzed, so as to improve the efficiency and reliability of pollutant concentration prediction, and provide decision-making basis for the government to control air pollution, traffic control and life travel.

The approaches for forecasting air pollutant concentrations mainly include deterministic and statistical models. Deterministic models simulate the atmospheric physic and chemistry in the processes of emission, diffusion and transformation of air pollutions, cannot explain the non-linearity and heterogeneity of some factors on the formation of pollutants [4],[5],[6],[7]. Statistical models are based on a data-driven manner ranther than sophisticated theoretical models to estimate air quality, has shown a virtue of obvious advantages [8],[9].

Recently, deep neural networks (DNNs) can automatically learn salient feature mappings from high-dimensional input data, and avoid the complicated process of artificial design and extraction of features to solve a wide variety prediction of complex problems, such as further perfect the prediction performance of air pollutant concentrations. Inherently considering spatio-temporal correlations of historical air pollutant data, meteorological data and timestamp data Li et al. proposed a novel LSTM model to forecast air pollutant concentration [10]. Qi embedded feature selection and spatial-temporal semi-supervised learning (ST-SSL) in the deep network to infer the PM2.5 concentration for the next few hours at all locations [11]. However, since the input of the model is a fixed-length sequence, the model's representation of the context is also a sequence of the same length, which limits the performance of the model. So it is difficult to get a suitable vector representation as the output.

In general, the air polluting process usually involves a variety of interacting pollutants, which are affected by local reactions, spatio-temporal evolution properties of air pollutant concentration and confounding factors, such as the direction of wind and humidity. Therefore, the research on the prediction of air pollutant concentration still faces the following two challenges:

(i) LSTM is hard to deal with the time series with long-term dependency and complex task,
(ii) air pollution causal pathways are complex among different locations in nature, since they may be influenced by geography, atmospheric phenomena and other complex factors.

To handle both challenges outlined above, we propose a deep spatio-temporal hybrid model to estimate air pollutant concentrations. The main contributions of the paper are given as follows:

(i) Granger causality is used to model the spatial correlation between different stations in adjacent regions, and consider the spatial dependence of air pollutant concentrations between the propagation of air pollution under different wind directions in each sub-region by constructing a hyperparametric Gauss vector weight function.
(ii) By constructing the Attention-CNN hybrid model, we can effectively extract the intrinsic features from historical air pollutant concentrations, meteorological and timestamp data by learning over a long time span, and then use LSTM layer to extract temporal information from these feature mappings.
(iii) We use air pollutant concentration data and meteorological monitoring data from northern Taiwan in 2015 for research and analysis to evaluate our methods. Abundant experiments prove that the model is superior to traditional machine learning methods.

The rest of the paper is organized as follows: Section 2 mainly introduce the data description, Attention-CNN-LSTM model, spatio-temporal correlation using Granger causality analysis, extraction of spatial and temporal features, attention mechanism in feature map and channel and prediction for air pollution concentration of multiple monitoring stations. Section 3 shows

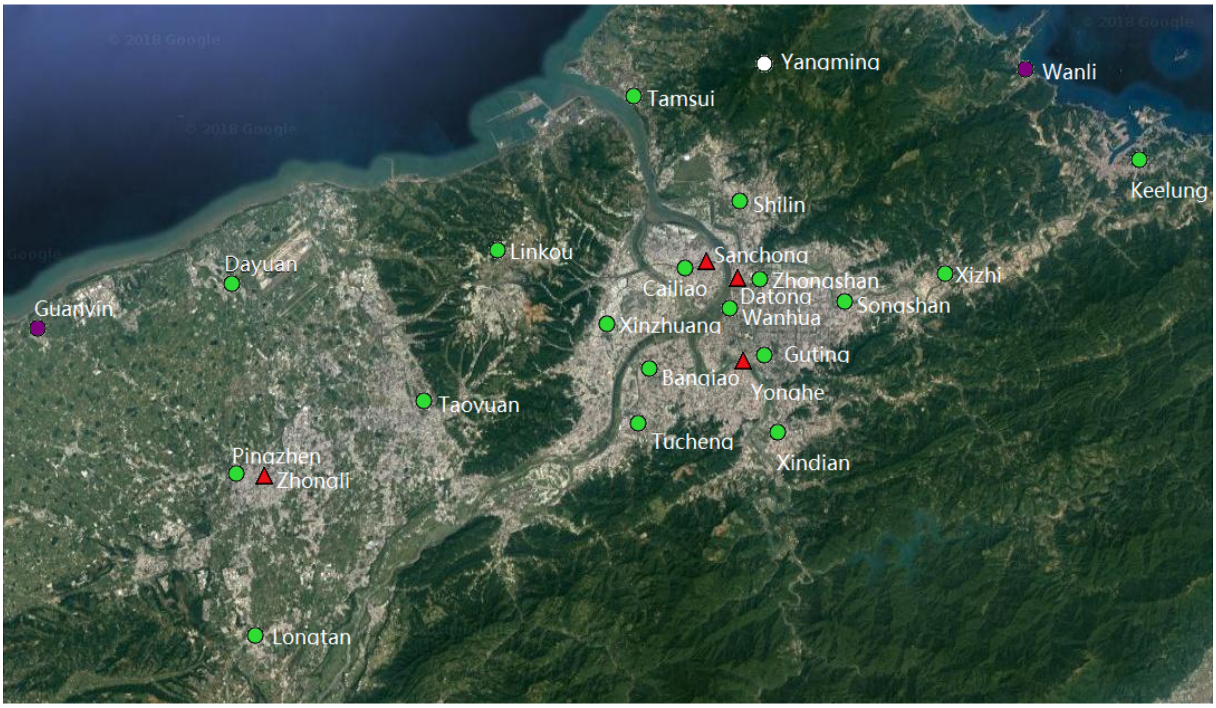

Figure 1. Topographic map and the location of monitoring sites in Northern Taiwan.

the experimental results. Finally, some concluding remarks and suggestions for future work are in Section 4.

2. Materials and methods

2.1. Data Preprocessing

The experimental data in this paper is from the Environmental Protection Administration, Executive Yuan, R.O.C. In the current experiments, air pollutant concentration data and meteorological monitoring data is collected every hour for 25 monitoring sites (shown in Figure 1) from Jan/01/2015 to Dec/31/2016. The meteorological data and air quality data used for CO and NO concentration prediction are shown in Table 1.To mitigate the negative impact of missing values on data analysis performance, we delete the timestamp to eliminate missing values, because filling in missing values requires accurate prediction of spatial and temporal correlation between different time series. If filling in missing time series data is not representative, temporal autocorrelation and spatial correlation may not be strong. In all the experiments, the data is divided into the training set (80%) and testing set (20%).

2.2. Attention-CNN-LSTM

The framework of the proposed spatial-temporal prediction model for multi-scale pollutant concentrations is shown in Figure 2. The main inputs (historical CO/NO concentration data) are included in brown box, and auxiliary inputs (meteorological data, related pollutant concentration and the time of day) are included in light blue box; r represents the number of time steps used, and the numbers in the parentheses represent the dimensions of each type of feature.

Considering the spatio-temporal correlation between 25 stations and their historical information using GC (Granger causality) analysis, as an index to measure the interaction

Table 1. Atmospheric pollutants and meteorological data applied in model studies.

Input parameters	Unit	Input parameters	Unit
NO	ppb	O_3	ppb
CO	ppm	Average temperature	$°C$
PM_{10}	$\mu g/m^3$	Relative humidity	%
$PM_{2.5}$	$\mu g/m^3$	Wind direction	degree
THC	ppm	Wind speed	m/sec
$NMHC$	ppm	Rainfall	mm
SO_2	ppb		

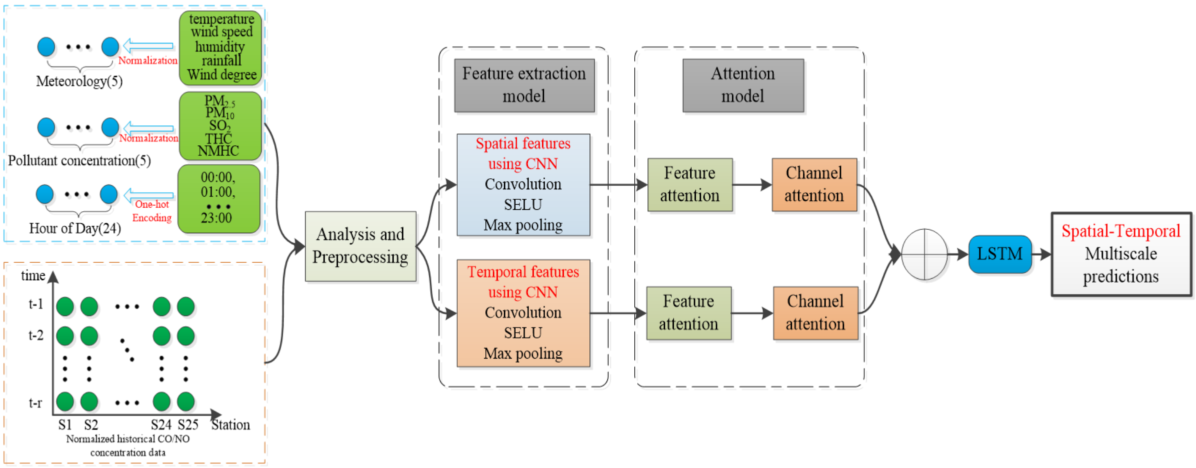

Figure 2. Attention-CNN-LSTM Architecture.

between time series, has been favored in recent decades. For complex spatial factors, we use GC to analyze the correlation between the air concentration time series. We define the time series of air pollutants at two monitoring sites as Y_i and X_i respectively. The formula of GC and the null hypothesis are given as follows:

$$Y_i(t) = \sum_{j=1}^{n} \Phi_i(j)Y_i(t-j) + \sum_{j=1}^{n} \mu_i(j)X_i(t-j) + \epsilon_t, \text{if } i \in N_d \quad (1)$$

$$Y_i(t) = \sum_{j=1}^{n} \Phi_i(j)Y_i(t-j) + \epsilon_t, \text{if } i \notin N_d \quad (2)$$

where N_d is the neighborhood set of the spatial clustering (The K-Means algorithm is adopted to gather them); ϵ_t is a white noise Gaussian random vector; n is the number of time stamps; vector Φ_i is the correspondent weights for Y_i; vector μ_i represents the spatial weight between spatial locations Y_i and X_i.

In the section of Spatial-temporal feature extraction, two or more CNN layers are selected to extract the intrinsic features from historical air pollutant data for long-term span learning, and then the one-hot encoding is used. The method encodes the hourly data and combines the extracted features with current meteorological data and related pollutant data to improve the predictive performance. In addition, we add batch normalization (BN) after the second and third convolutional layers of the model. Considering that the scaled exponential linear units (SELU) function has better convergence performance and can effectively avoid the gradient disappearance problem, it is taken as activation function in this paper [12].

In the section of Feature map attention, We adopted an attention mechanism to weigh the hidden features to enhance their validity. In the attention mechanism, $F = \{f(1), f(2), ..., f(j)\}$ is the output hidden feature maps of the convolutional layer, where $j \in R$ is the number of the convolution kernel. The weighted feature maps F' are computed by *softmax*:

$$F' = W \cdot F \tag{3}$$

$$f'(i) = \omega_i * f(i)$$
$$\omega_i = \mathit{softmax}(F) = \frac{\exp(f(i))}{\sum_{t=1}^{i} \exp(f(t))} \tag{4}$$

here $W = \{\omega_1, \omega_2, ..., \omega_j\}$ is a weight matrix and its size is the same as that of the feature maps. To generate W, the attention mechanism consists of 3 convolution layers with the stride 1. The first convolution layer has $k \times s$ filters with convolution kernel size 5×5, the second and third layers have k filters with convolution kernel size 3×3, and the number of filters is 100 for each convolutional layer.

By stacking several layers of LSTM, the features in spatially correlated contaminant data with long-term dependence can be automatically extracted layer by layer, and the fused features can be used to compute multiscale time series prediction of air pollutants. For the LSTM layer, one input is the temporal information of $X = (x_1, x_2, ..., x_t)$, and another input is the hidden unit h_{t-1} from the last time step. The forward training process of Attention-CNN-LSTM can be expressed by the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

$$C_t = f_t * C_{t-1} + i_t * \mathit{tanh}(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{7}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t * \mathit{tanh}(C_t) \tag{9}$$

where i_t, o_t, and f_t denote the activation of the input gate, output gate and forget gate, respectively; C_t and h_t denote the activation vector for each cell and memory block, respectively; and W and b denote the weight matrix and bias vector. The output from the last step of the LSTM is then fed to the fully connected layer for spatio-temporal air pollution prediction.

3. Results and discussion

3.1. Performance metric and settings

Each monitoring station collects air quality data once an hour, and the dataset contains more than 400000 instances, each with concentrations of CO and NO. To prove the effectiveness of the proposed Attention-CNN-LSTM, several models are used for comparison, including ARIMA [13], SVR [14], CNN [15], LSTM [16], and CNN-LSTM [17]. In order to prevent inconsistency of data magnitude differences and gradient explosion, we need to convert all the input data by scaling the attributes between $[0, 1]$ using the min-max normalization. The performance evaluation indicators, including the root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and correlation coefficient (R), were used to evaluate the effectiveness of our model in our experiments.

3.2. *Performance comparisons*

Obviously, the current state has different effects on different time intervals in the future [18]. Therefore, we analyze the input data with the air pollution concentration at multiple time intervals on the next 1st hour to 24th hour to develop different training sets. Over the next 3h, we train a model for each hour, respectively. With respect to the next $4-24$h, which are divided into the three groups ($4-6$h, $7-12$h, $13-24$h) and the models are trained for each time interval.

As it is shown in the Tables 2,3,4,5, the accuracy of all the models decreases as the prediction time extends. However, for CO concentration, the RMSE standard deviation of Attention-CNN-LSTM, which varies from 0.4602 to 0.5803, is much lower than that for the other models, indicating that Attention-CNN-LSTM achieves higher accuracy and stability in long-term prediction. This is due to the combination of 3D-CNN, attention mechanism and LSTM, which extract advanced spatio-temporal features while maintaining the transmission of state information, rather than using LSTM alone or CNN. Prediction results show that Attention-CNN-LSTM is valid for air pollutant concentration forecasting in total data set.

Table 2. Comparisons of MAE using different models for the next 1st to the 24th hour predictions

		1h	2h	3h	4-6h	7-12h	13-24h
NO	ARIMA	13.42	13.28	13.64	14.49	14.89	15.28
	SVR	11.95	11.94	12.26	12.70	13.02	13.13
	CNN	10.67	11.86	11.64	11.76	12.41	12.68
	LSTM	9.906	10.08	10.90	11.26	11.87	12.14
	CNN-LSTM	9.193	9.727	10.09	10.52	11.19	11.55
	Attention-CNN-LSTM	8.568	8.693	9.270	9.830	10.68	10.77
CO	ARIMA	0.5131	0.5478	0.5910	0.6253	0.6479	0.6876
	SVR	0.4516	0.4632	0.5124	0.5706	0.5880	0.6176
	CNN	0.4365	0.4570	0.4565	0.4745	0.5031	0.5466
	LSTM	0.4149	0.4323	0.4240	0.4400	0.4586	0.4919
	CNN-LSTM	0.4079	0.4078	0.4115	0.4395	0.4439	0.4458
	Attention-CNN-LSTM	0.3481	0.3627	0.3770	0.3954	0.4027	0.4360

4. Conclusions

In this paper, an attention-based CNN-LSTM model has been proposed to forecast air pollutant concentration. Granger causality analysis is utilized to explore the spatial correlation among different monitoring sites and spatial features are added into the prediction model. The model combines ordinary convolution units and attention mechanism to extract the spatial and temporal feature maps. Finally, the air pollutant concentration of multiple monitoring stations is predicted by LSTM, which can learn temporal dependencies on time series of pollutant concentrations. Experimental results have demonstrated that the proposed Attention-CNN-LSTM outperforms the other state-of-the-art algorithms in terms of RMSE, MAE, MAPE and R values. In order to further improving the performance of the proposed method, several aspects remain to be investigated in the future work: (1) Exploring the spatio-temporal clustering method based on weather patterns because the air polluting process may be affected by multiple weather patterns; (2) Exploring multi-faceted causality analysis and environmental factors.

Table 3. Comparisons of RMSE using different models for the next 1 to 24 hour prediction

		1h	2h	3h	4-6h	7-12h	13-24h
NO	ARIMA	17.80	18.27	18.75	19.42	19.62	20.04
	SVR	16.07	16.07	16.61	17.04	17.66	17.71
	CNN	14.05	15.20	15.83	16.05	16.52	16.95
	LSTM	13.19	14.49	14.58	15.18	15.59	15.72
	CNN-LSTM	13.08	13.82	14.09	14.75	15.04	15.59
	Attention-CNN-LSTM	12.27	12.87	13.29	13.69	14.18	14.79
CO	ARIMA	0.6125	0.6759	0.6874	0.7372	0.7775	0.7854
	SVR	0.5536	0.6191	0.6365	0.6490	0.6703	0.6803
	CNN	0.5269	0.5710	0.5865	0.6125	0.6365	0.6423
	LSTM	0.5043	0.5597	0.5699	0.5961	0.6027	0.6106
	CNN-LSTM	0.4988	0.5784	0.5820	0.6037	0.6121	0.6145
	Attention-CNN-LSTM	0.4602	0.4634	0.4969	0.5654	0.5710	0.5803

Table 4. Comparisons of R using different models for the next 1 to 24 hour prediction

		1h	2h	3h	4-6h	7-12h	13-24h
NO	ARIMA	0.7910	0.7770	0.7672	0.7443	0.7346	0.7290
	SVR	0.8307	0.8286	0.8158	0.8056	0.7934	0.7883
	CNN	0.8519	0.8464	0.8348	0.8289	0.8183	0.8078
	LSTM	0.8867	0.8755	0.8631	0.8534	0.8385	0.8248
	CNN-LSTM	0.8980	0.8834	0.8729	0.8616	0.8594	0.8363
	Attention-CNN-LSTM	0.9057	0.8933	0.8832	0.8787	0.8648	0.8386
CO	ARIMA	0.8146	0.7892	0.7796	0.7379	0.7163	0.7093
	SVR	0.8374	0.8274	0.8138	0.8044	0.7908	0.7785
	CNN	0.8674	0.8518	0.8434	0.8271	0.8138	0.8090
	LSTM	0.8858	0.8742	0.8639	0.8490	0.8319	0.8249
	CNN-LSTM	0.8919	0.8805	0.8766	0.8633	0.8304	0.8360
	Attention-CNN-LSTM	0.9219	0.9091	0.8970	0.8736	0.8518	0.8508

Acknowledgments

This work was supported in part by the Leading Talents of Science and Technology Innovation in Zhejiang Province 10 Thousands Plan under Grant 2018R52040, in part by the National Key Research and Development Program of China under Grant 2016YFC0201400, in part by the Provincial Key Research and Development Program of Zhejiang Province under Grant 2017C03019, and in part by the International Science and Technology Cooperation Program of Zhejiang Province for Joint Research in High-tech Industry under Grant 2016C54007.

References

[1] Kurt A and Oktay A B 2010 Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks *Expert Systems with Applications* **37(12)** 7986-7992

[2] Arhami M, Kamali N and Rajabi M 2013 Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations *Environmental Science and Pollution Research* **20(7)** 4777-4789

[3] Yang W, Deng M, Xu F and Wang H 2018 Prediction of hourly PM2.5 using a space-time support vector regression model *Atmospheric Environment* **181** 12-19

Table 5. Comparisons of MAPE using different models for the next 1 to 24 hour prediction

		1h	2h	3h	4-6h	7-12h	13-24h
NO	ARIMA	53.01	52.37	53.18	59.45	61.45	65.23
	SVR	45.05	44.18	46.08	51.42	54.10	56.45
	CNN	38.76	39.15	41.47	44.79	45.79	47.69
	LSTM	31.61	36.96	39.45	41.03	43.73	44.80
	CNN-LSTM	30.78	34.00	35.14	39.43	40.16	42.09
	Attention-CNN-LSTM	28.06	28.43	29.04	29.75	31.27	34.02
CO	ARIMA	43.72	45.29	48.50	53.18	57.94	59.44
	SVR	35.75	37.57	38.54	41.24	45.05	51.79
	CNN	25.96	26.40	28.22	29.55	30.54	32.96
	LSTM	22.27	25.23	27.12	28.62	29.35	30.29
	CNN-LSTM	22.81	23.98	24.29	26.96	27.84	29.13
	Attention-CNN-LSTM	21.26	22.48	23.99	25.23	27.98	28.17

[4] Kim Y, Fu J S and Miller T L 2010 Improving ozone modeling in complex terrain at a fine grid resolution: Part I–examination of analysis nudging and all PBL schemes associated with LSMs in meteorological model *Atmospheric Environment* **44(4)** 523-532

[5] Jeong J I, Park R J, Woo J H, Han Y J and Yi S M 2011 Source contributions to carbonaceous aerosol concentrations in Korea *Atmospheric environment* **45(5)** 1116-1125

[6] Crippa M, Canonaco F, Slowik J G, El Haddad I, De-Carlo P F, Mohr C, ... and Abidi E 2013 Primary and secondary organic aerosol origin by combined gas-particle phase source apportionment. *Atmos. Chem. Phys* **13(16)** 8411-8426

[7] Zhou G, Xu J, Xie Y, Chang L, Gao W, Gu Y and Zhou J 2017 Numerical air quality forecasting over eastern China: An operational application of WRF-Chem *Atmospheric environment* **153** 94-108

[8] Wang D, Wei S, Luo H, Yue C and Grunder O 2017 A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine *Science of The Total Environment* **580** 719-733

[9] Prasad K, Gorai A K and Goyal P 2016 Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time. *Atmospheric Environment* **128** 246-262

[10] Li X, Peng L, Yao X, Cui S, Hu Y, You C and Chi T 2017 Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation *Environmental Pollution* **231** 997-1004

[11] Qi Z, Wang T, Song G, Hu W, Li X and Zhang Z M 2018 Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. *IEEE Transactions on Knowledge and Data Engineering*

[12] Klambauer G, Unterthiner T, Mayr A and Hochreiter S 2017 Self-normalizing neural networks *Advances in Neural Information Processing Systems* 971-980

[13] Dickey D 1989 Time series theory and methods. *Technometrics* **31(1)** 121-121

[14] Nieto P J, Combarro E F, Diaz J J and Montanes E 2013 A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study *Applied Mathematics and Computation* **219(17)** 8923-8937

[15] Chen Y N, Han C C, Wang C T, Jeng B S and Fan K C 2009 A CNN-based face detector with a simple feature map and a coarse-to-fine classifier-Withdrawn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*

[16] Ghaderi A, Sanandaji B M and Ghaderi F 2017 Deep forecast: deep learning-based spatio-temporal forecasting *arXiv preprint* arXiv:1707.08110

[17] Wu Y and Tan H 2016 Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework. *arXiv preprint* arXiv:1612.01022

[18] Zheng Y, Yi X, Li M, Li R, Shan Z, Chang E, Li T 2015 Forecasting finegrained air quality based on big data. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2267-2276