Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **6**, Number **4**, November **2021**

Optimization of air quality monitoring network based on LSTM

Li-Wen Chen
Research Institute of ubiquitous sensing and multi-sensor intelligent fusion Fujian University of Technology
Fujian University of Technology
No.33 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, 350118, China
clw@fjut.edu.cn

Ye Zhang*
Research Institute of ubiquitous sensing and multi-sensor intelligent fusion Fujian University of Technology
Fujian University of Technology
No.33 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, 350118, China
Corresponding Author:zhangye@88.com

Wen-Ji Zhang
Research Institute of ubiquitous sensing and multi-sensor intelligent fusion Fujian University of Technology
Fujian University of Technology
No.33 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, 350118, China
zwjfjut@163.com

Jian-Hua Liu
Fujian Provincial Key Laboratory of Automotive Electronics and Electric Drive
Fujian University of Technology
No.33 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, 350118, China
jhliu@fjnu.edu.cn

Jeng-Shyang Pan
Shandong University of Science and Engineering
Department of Information Management
Chaoyang University of Technology
Taichung, 413310, Taiwan
No.579 Qianwangang Road, Huangdao District, Qingdao, Shandong Province, 266590, China
jengshyangpan@gmail.com

Received July 2021; revised September 2021

Abstract. *The air quality monitoring networks are the main tool for measuring, managing and evaluating urban air quality. However, the existing monitoring network has the problem of uneven distribution, some places are dense and some places are sparse. In order to solve this problem, it is necessary to optimize the monitoring network. For this reason, we propose a remote optimization method, which uses the available data of adjacent stations to estimate the pollutant concentration of the target station instead of measurement. For the air quality monitoring networks, we first use a combination of cluster analysis and correlation analysis to identify the six air pollutants PM2.5, PM10, sulfur dioxide, nitrogen dioxide, carbon monoxide, and ozone in the monitoring network whether there is redundant monitoring equipment. Then use long short-term memory (LSTM) predicts the pollutant concentration of the target station based on the data of adjacent stations to replace redundant monitoring equipment. For areas without monitoring stations, we use data from adjacent stations to set up virtual stations to monitor pollutant concentrations. Experimental results show that this method can effectively optimize the layout of the air quality monitoring networks, reduce costs, and improve the integrity of monitoring information.*
Keywords: Air quality monitoring network, Remote optimization strategy, Long short-term memory, Redundant equipment, Virtual monitoring station.

1. **Introduction.** In the past few decades, air pollution has become a major problem in the world [1,2]. Air pollution affects people's daily life and health [3-5]. In response to this, the Chinese government has established air pollution monitoring networks in most cities. But the layout of the monitoring network has problems with unreasonable distribution, which will not only increase the financial burden but also affect the integrity of air pollutant monitoring information.To this end, the air pollution monitoring network needs to be optimized.

To this end, some researchers have used several methods to optimize the air pollution monitoring network. Gómez-Losada et al. [6] used a hierarchical clustering algorithm, to examine the similar behavior in Seville air quality monitoring network. Lu et al. [7] adopted principle component analysis to discover the redundant equipment for optimizing the air monitoring networks. Cotta et al. [8] proposed robust principal component analysis to identify air quality monitoring stations that present similar behavior for any pollutant or meteorological measure. D'Urso et al. [9] adopt a Fuzzy C-Medoids to detect possible information redundancy in the monitoring networks and then, decreasing the number of monitoring stations.

Artificial intelligence algorithms are widely used in predictive models [10-13]. Some statistical and machine learning techniques have been developed to predict the pollutant concentration in the air quality monitoring network [14-18]. Random forest and xgboost are two more commonly used methods. Yang et al. [19] used the random forest to predict PM2.5 concentration. Besides algorithms developed based on traditional statistical methods and machine learning methods, more and more studies recently started to implement deep learning technologies for air pollution prediction. Prakash et al. [20] proposed a wavelet-based recurrent neural network (RNN) model to forecast one step hourly, daily mean, and maximum daily concentrations of ambient CO, NO, PM2.5, and other most prevalent air pollutants. Krishan et al. [21] extended a long short-term memory network for air pollution prediction and achieved better performance than other deep learning methods. Eric KeWang et al. [22] proposed a method of Object segmentation of monitor systems based on the Internet of drones provides a reliable theoretical basis for key property monitoring, environmental monitoring, disaster monitoring, and agricultural monitoring.

With the continuous deepening of artificial intelligence and machine learning research, various types of neural networks and their variants are also emerging in an endless stream, and they all play an important role in their respective fields. The long short-term memory (LSTM) network improved on the basis of recurrent neural networks (RNN) compensates for the disappearance and explosion of gradients in RNN networks, and can make full use of longer-distance time series information. LSTM network has many successful application cases in the fields of pattern recognition, machine translation, traffic pattern detection, traffic flow prediction, stock trading, medicine, etc. [23-25], but there are very few researches on environmental monitoring.

All of these studies indicate that there exists redundant information in the air quality monitoring network. However, the existing method only reduces redundant equipment and does not make an overall optimization of the air quality monitoring network. In this article, we try to explore some feasible methods to optimize the air quality monitoring network. We take the Fuzhou Air Quality Monitoring Network as a research case. For this reason, the situation of redundant equipment in the monitoring network is first revealed. Then based on the similarity, LSTM is used to predict the air concentration of the target station, which ensures that the prediction can replace the lost value caused by the disassembly of the equipment. Then, for areas without monitoring stations, we set up virtual monitoring stations through the air data of adjacent stations to monitor the areas without monitoring stations.We used LSTM, XGBoost and RF models to predict and compare the pollutant concentration results. It shows that the prediction accuracy and effect of LSTM network model are better than the other two models. Experiments show that this method can improve the comprehensiveness and accuracy of urban pollution monitoring.

2. **Materials and methods.**

2.1. **Study area.** Fuzhou is located in the middle east of Fujian Province. The landform of Fuzhou is a typical estuary basin, and its elevation is mostly between 600 and 1,000 meters. Fuzhou lies on the north bank of the estuary of Fujian's largest river and is an important commercial and government center on the southeast coast of China. Fuzhou has a typical subtropical monsoon climate with suitable temperature, warm and humid.

2.2. **Data sources.** There are five national-controlling air quality monitoring sites in the urban area of Fuzhou city (Figure 1). There are five national-controlling air quality monitoring stations in the urban area of Fuzhou city (Figure 1). Shida station(SD), Ziyang station(ZY), Yangqiaoxilu station(YQXL), and Wusibeilu station(WSBL) are located in densely populated urban areas. Gushan station(GS) located in Gu Mountain. At these stations, concentrations of PM2.5, PM10, sulfur dioxide, nitrogen dioxide, ozone and carbon monoxide are monitored. In this study, the real-time hourly mass concentrations of six kinds of pollutants from January 1, 2019, to September 31, 2019, at the five monitoring stations were collected from China National Environmental Monitoring Center.

2.3. **Remote optimization strategy.** We propose a remote optimization strategy. We estimate the air data of the target station through the air data of adjacent stations. For the monitoring network, we first perform cluster analysis and correlation analysis on the monitoring sites of the six air pollutants, so that we can determine which air pollutant monitoring equipment in the monitoring network has redundancy. Then we use the available data of adjacent monitoring stations to predict the air data of the target station and use the prediction results to replace the actual values measured by redundant monitoring equipment. For areas that are not monitored by monitoring stations, we use

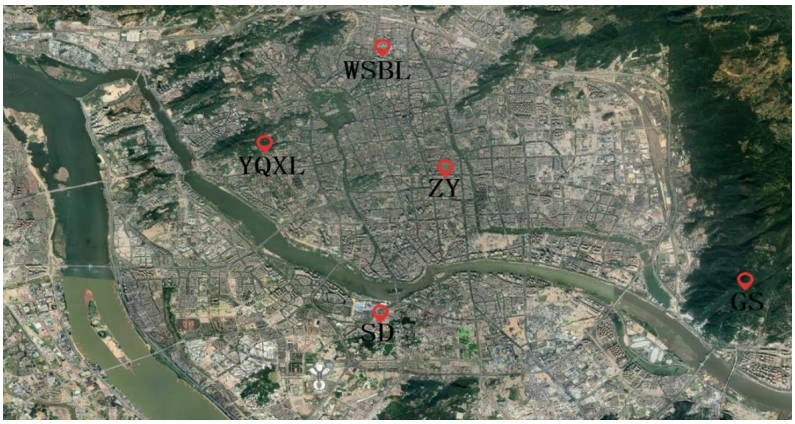

FIGURE 1. Location of the Fuzhou air monitoring station

a remote optimization strategy to set up virtual monitoring stations to monitor the air in locations where no monitoring stations have been deployed.

2.4. **Cluster analysis.** Cluster analysis is a multivariate statistical method, and the most used cluster analysis method is hierarchical clustering. Consider multiple samples of the research object as one type, and consider several samples as the same type, calculate their mutual distance or similarity coefficient, and combine the samples with the smallest distance or the largest similarity into one category. The size of Euclidean distance reflects the similarity relationship between individuals, as shown in Eq.(1).

$$d_{ij} = \sqrt{\sum_{t=1}^{n} \left| x_{it} - x_{jt} \right|^2} \tag{1}$$

2.5. **Correlation analysis.** Cluster analysis is a multivariate statistical method, and the most used cluster analysis method is hierarchical clustering. Consider multiple samples of For a certain pollutant, if the pollution information of the monitoring station has a high degree of correlation, it means that the environment around the two monitoring stations has high similarity. We use the Pearson correlation coefficient r to analyze the correlation, as shown in Eq.(2).

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{2}$$

2.6. **Long Short Term Memory.** LSTM improved from the recurrent neural network model (RNN). The unit of LSTM shown in Figure 2. There are three types of gates in a unit: forget gate, Input gate, Output gate. The gate mechanism used to control the discarding or to retain information, making the memory information in the time series controllable. The formula of LSTM shown in Eq.(3)-(7).

$$f_t = \sigma(W_f [h_{t-1}, x_t] + b_f) \tag{3}$$

$$i_t = \sigma(W_i [h_{t-1}, x_t] + b_i) \tag{4}$$

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \tag{5}$$

$$c_t = f_t c_{t-1} + i_t \tanh (W_c [h_{t-1}, x_t] + b_c) \tag{6}$$

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

W_f、W_i、W_o、W_c are input weights.b_f、b_i、b_o、b_c are bias weights.c_t is the internal memory computed in this unit.

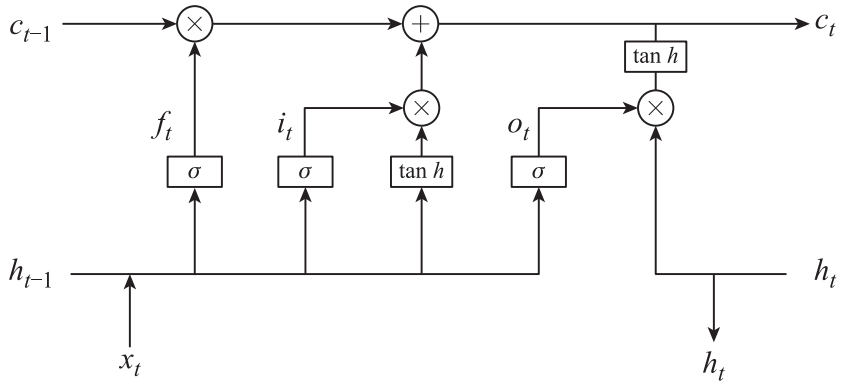

FIGURE 2. LSTM unit

2.7. **Model performance.** In our study, three indexes were employed to measure accuracy. These indexes are the coefficient of determination (R^2), root mean square error ($RMSE$), Mean absolute percentage error($MAPE$), and calculated as follows:

$$R^2 = \frac{\left(\sum_{i=1}^{N}\left(y_{true}^{i}-\overline{y}_{true}\right)\left(y_{pred}^{i}-\overline{y}_{pred}\right)\right)^2}{\sum_{i=1}^{N}\left(y_{true}^{i}-\overline{y}_{true}\right)^2\sum_{i=1}^{N}\left(y_{pred}^{i}-\overline{y}_{pred}\right)^2} \tag{8}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{true}^{i}-y_{pred}^{i}\right)^2} \tag{9}$$

$$MAPE = \sum_{i=1}^{N}\left|\frac{y_{true}^{i}-\overline{y}_{pred}}{y_{true}^{i}}\right|\frac{100}{N} \tag{10}$$

N denotes the number of evaluation samples,y_{true}^{i} is the observed value,y_{pred}^{i} is the predicted value,$\overline{y}_{true}$ is the average of the observed value,$\overline{y}_{pred}$is the average of the predicted value.

3. Results and discussion.

3.1. **Monitoring network redundancy analysis.** The Pearson correlation analysis is employed to have a first look at the linear relationships of six kinds of pollutants among all stations. Then the cluster analysis is utilized to reveal the redundant information in the monitoring network.

Figure 3 displays the results of Spearman correlation analysis among five stations. First, it is found that the correlation coefficient between GS station with other stations is low. GS station located in Gu Mountain, which results in distinctive behavior with others. Thus, GS station is not redundant monitoring stations. Secondly, for the four stations located in the urban area of ZY, WSBL,YQXL, and SD, we found that PM2.5, PM10, nitrogen dioxide, and ozone have relatively high correlations, and the correlation coefficients in most cases greater than 0.8. For carbon monoxide, sulfur dioxide, the correlation coefficient is relatively low. Therefore, sulfur dioxide and carbon dioxide monitoring equipment are not redundant monitoring equipment.

Besides, cluster analysis carried out and the results exhibited in Figure 4. Analyze six kinds of pollutants according to the RDCC value calculated from SPSS. For PM2.5, ZY and WSBL stations are one category, SD and YQXL stations are the other. For PM10, ZY

FIGURE 3. Analysis of the correlation (Pearson's r) between monitoring stations for six kinds of pollutants

and SD stations are one category, WSBL and YQXL stations are the other. For nitrogen dioxide, WSBL, YQXL and SD stations are one category. For ozone, ZY and WSBL stations are one category. For sulfur dioxide, SD and WSBL stations are one category, ZY and YQXL stations are the other. For carbon monoxide, SD and WSBL stations are one category. The lower the RDCC value within the category, the more similar behaviors the category presents. The results of Pearson correlation analysis and cluster analysis show that there are redundant devices in the air quality monitoring networks.

3.2. **Analysis of prediction results of remote optimization strategy.** According to redundancy analysis, we use LSTM, XGBoost, and Random forest (RF) three models to predict PM2.5, PM10, ozone, and nitrogen dioxide concentration at the target station. For a better illustration, the predicted results at ZY station is selected as an example to examine three model performance. The adjacent stations of ZY station are WSBL, SD, YQXL, GS stations, but the correlation coefficient of GS station is low, so the GS station is not considered. The prediction result is shown in Figure 5. For PM2.5 and nitrogen dioxide, the prediction effect of the LSTM model is better than the other two models. For PM10 and ozone, the prediction effects of the three models are similar.

For a better illustration, Table 1 presents the statistical evaluation indices of the three models at all stations except the GS station. From the table, some conclusions can be addressed 1), it can be recognized that the LSTM model has a better prediction effect than the XGBoost model and the RF model; 2), the remote optimization strategy confirmed

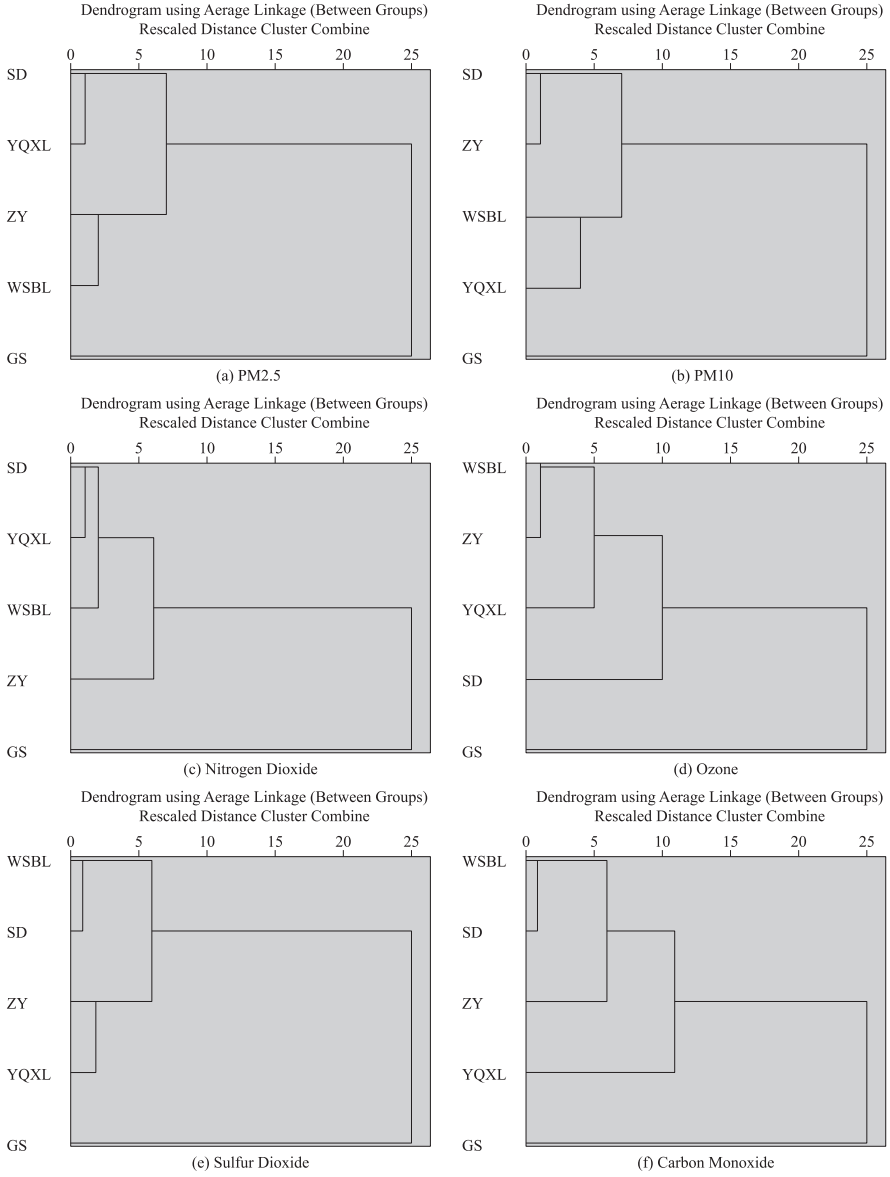

FIGURE 4. The dendrogram for six kinds of pollutants

the possibility of predicting pollutant concentration at target station using available data from adjacent stations.

Overall, this strategy confirmed that the pollution level at some stations could be well predicted instead of measurement directly. This strategy reveals the possibility of removing or relocating some equipment in the air quality monitoring network in the future.

3.3. **Feasibility analysis of virtual monitoring station.** For areas not monitored by monitoring stations, we use available data from adjacent stations to set up a virtual monitoring station. Because adjacent monitoring stations may not have a high correlation with unmonitored areas, we need to test whether the target station with low correlation is suitable for remote optimization strategy. The correlation between the GS station and its adjacent stations is low, so we choose GS station as our target station. The adjacent stations of GS stations are ZY and SD stations.

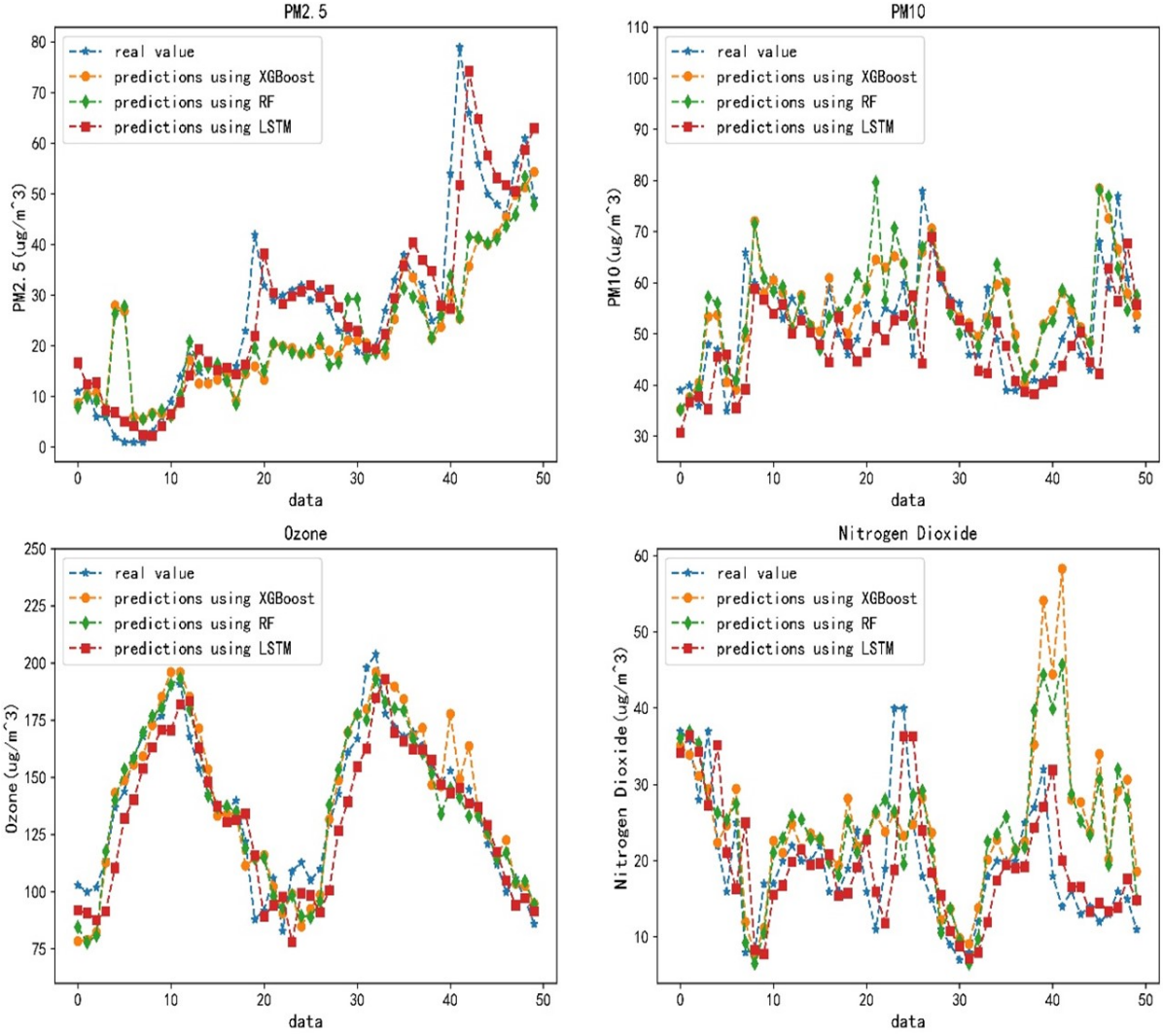

FIGURE 5. Comparison of prediction results of three models at ZY station

The prediction results using the LSTM model are shown in Figure 6. For six kinds of pollutants, $0.68 < R^2 < 0.94$. This means that it is feasible to use a remote optimization strategy to set up virtual monitoring sites in unmonitored areas.

TABLE 1. Forecast results evaluation

		XGBoost			RF			LSTM		
		MAPE	RMSE	R^2	MAPE	RMSE	R^2	MAPE	RMSE	R^2
ZY	PM2.5	21.29%	6.96	0.79	22.49%	7.05	0.79	15.65%	5.68	0.86
	PM10	14.01%	11.54	0.81	14.27%	11.23	0.82	15.10%	13.03	0.76
	Ozone	38.63	12.16	0.91	39.84%	12.05	0.91	30.39%	12.83	0.90
	Nitrogen Dioxide	28.45%	8.83	0.71	29.48%	8.81	0.71	17.25%	6.54	0.84
YQXL	PM2.5	16.70%	5.86	0.78	17.89%	5.72	0.78	13.31%	4.49	0.86
	PM10	15.72%	11.05	0.72	15.63%	10.81	0.73	16.75%	11.31	0.71
	Ozone	30.03%	14.63	0.87	30.72%	13.55	0.89	24.45%	12.10	0.91
	Nitrogen Dioxide	37.75%	8.03	0.67	38.99%	8.05	0.67	17.28%	6.53	0.84
SD	PM2.5	24.26%	7.02	0.74	25.76%	7.32	0.72	13.62%	5.33	0.85
	PM10	13.88%	11.07	0.79	14.92%	11.16	0.79	14.12%	10.76	0.80
	Ozone	18.35%	14.73	0.87	20.20%	15.57	0.85	15.42%	12.26	0.91
	Nitrogen Dioxide	22.92%	8.76	0.64	24.63%	8.95	0.62	16.89%	6.09	0.83
WSBL	PM2.5	19.07%	6.83	0.77	20.42%	6.92	0.76	12.91%	5.07	0.87
	PM10	18.44%	12.31	0.69	20.10%	12.93	0.67	15.62%	9.76	0.80
	Ozone	19.57%	10.52	0.93	21.41%	10.87	0.93	25.51%	11.98	0.92
	Nitrogen Dioxide	20.91%	7.67	0.70	22.00%	7.81	0.69	16.42%	6.02	0.82

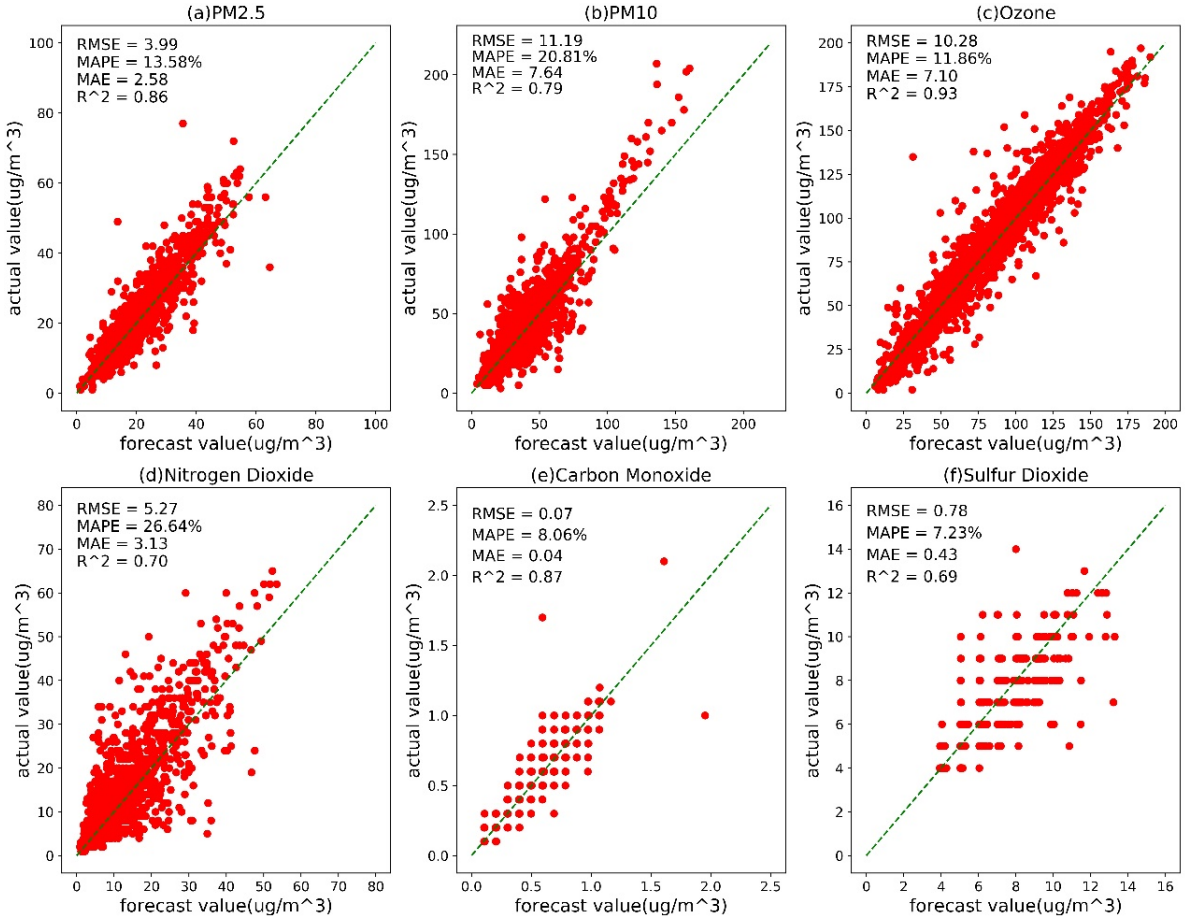

FIGURE 6. Prediction results of six kinds of pollutants at GS station

4. **Conclusions.** We have presented a remote optimization strategy to optimize the air quality monitoring network. Firstly, the correlation analysis and cluster analysis are

employed to reveal the redundant information in the monitoring network. The results verified the existence of redundant equipment in the monitoring network. Then the LSTM, XGBoost and RF are developed to predict Pollutants concentration. The prediction results successfully confirmed that the prediction effect of the LSTM model is better than the other two models and the pollutant concentrations measured by redundant equipment can be well estimated with the available data at adjacent stations. Secondly, we test the feasibility of setting up virtual monitoring stations through remote optimization strategies. The result proves that the air data of the unmonitored area can be monitored through the adjacent stations. This strategy can optimize the air quality monitoring network and expand its monitoring range. The results, therefore, have great practical significance for improving the comprehensiveness and accuracy of urban pollution monitoring.

Acknowledgment. This work was supported in part by the Fujian Science and Technology Department Grant 2018H0003, and FJUT scientific research GJ-YB-20-07 and GY-Z21035.

REFERENCES

[1] S. H. L. Yim, Y. Gu, M. A. Shapiro, and B. Stephens, Air quality and acid deposition impacts of local emissions and transboundary air pollution in Japan and South Korea, *Atmospheric Chemistry and Physics*, vol. 19, no. 20, pp. 13309–13323, 2019.

[2] J. B. Beltman, C. Hendriks, M. Tum, and M. Schaap, The impact of large scale biomass production on ozone air pollution in Europe, *Atmospheric Environment*, vol. 71, pp. 352–363, 2013.

[3] R. Liu, M. T. Young, J. C. Chen, J. D. Kaufman, and H. Chen, Ambient air pollution exposures and risk of parkinson disease, *Environmental Health Perspectives*, vol. 124, no. 11, pp. 1759–1765, 2016.

[4] L. Stockfelt, E. M. Andersson, P. Molnár, L. Gidhagen, D. Segersson, A. Rosengren, and G. Sallsten, Long-term effects of total and source-specific particulate air pollution on incident cardiovascular disease in Gothenburg, Sweden, *Environmental Research*, vol. 158, pp. 61-71, 2017.

[5] M. A. Kioumourtzoglou, J. D. Schwartz, M. G. Weisskopf, S. J. Melly, Y. Wang, F. Dominici, and A. Zanobetti, Long-term PM2.5 exposure and neurological hospital admissions in the northeastern United States, *Environmental Health Perspectives*, vol. 124, no. 1, pp. 23–29, 2016.

[6] Á. Gómez-Losada, A. Lozano-García, R. Pino-Mejías, and J. Contreras-González, Finite mixture models to characterize and refine air quality monitoring networks, *Science of the Total Environment*, vol. 485–486, no. 1, pp. 292-299, 2014.

[7] W. Z. Lu, H. D. He, and L. Y. Dong, Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis, *Building and Environment*, vol. 46, no. 3, pp. 577-583, 2011.

[8] H. H. A. Cotta, V. A. Reisen, P. Bondon, and P. R. P. Filho, Identification of Redundant Air Quality Monitoring Stations using Robust Principal Component Analysis, *Environmental Modeling and Assessment*, vol. 25, no. 4, pp. 521–530, 2020.

[9] P. D'Urso, D. Di Lallo, and E. A. Maharaj, Autoregressive model-based fuzzy clustering and its application for detecting information redundancy in air pollution monitoring networks, *Soft Computing*, vol. 17, no. 1, pp. 83–131, 2013.

[10] E. K. Wang, F. Wang, S. Kumari, J. H. Yeh, and C. M. Chen, Intelligent monitor for typhoon in IoT system of smart city, *The Journal of Supercomputing*, vol. 77, no. 3, pp. 3024-3043, 2021.

[11] K. K. Tseng, R. Zhang, C. M. Chen, and M. M. Hassan, DNetUnet: a semi-supervised CNN of medical image segmentation for super-computing AI service, *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3594–3615, 2021.

[12] F. Zhang, T. Y. Wu, Y. Wang, R. Xiong, G. Ding, P. Mei, and L. Liu, Application of quantum genetic optimization of LVQ neural network in smart city traffic network prediction, *IEEE Access*, vol. 8, pp. 104555-104564, 2020.

[13] M. A. Elshafey, A. S. Amein and K. S. Badran, Universal Image Steganography Detection using Multimodal Deep Learning Framework, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 12, no. 3, pp. 152-161, 2021.

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

[15] E. G. Ortiz-García, S. Salcedo-Sanz, Á. M. Pérez-Bellido, J. A. Portilla-Figueras, and L. Prieto, Prediction of hourly O3 concentrations using support vector regression algorithms, *Atmospheric Environment*, vol. 44, no. 35, pp. 4481-4488, 2010.

[16] L. K. Kwok, Y. F. Lam, and C. Y. Tam, Developing a statistical based approach for predicting local air quality in complex terrain area, *Atmospheric Pollution Research*, vol. 8, no. 1, pp. 114-126, 2017.

[17] G. Grivas, and A. Chaloulakou, Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece, *Atmospheric Environment*, vol. 40, no. 7, pp. 1216-1229, 2006.

[18] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar, Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities, *IEEE Access*, vol. 7, pp. 128325–128338, 2019.

[19] L. Yang, H. Xu, and S. Yu, Estimating PM2.5 concentrations in Yangtze River Delta region of China using random forest model and the Top-of-Atmosphere reflectance, *Journal of Environmental Management*, vol. 272, no. 15, pp. 111061, 2020.

[20] A. Prakash, U. Kumar, K. Kumar, and V. K. Jain, A Wavelet-based Neural Network Model to Predict Ambient Air Pollutants' Concentration, *Environmental Modeling and Assessment*, vol. 16, no. 5, pp. 503–517, 2011.

[21] M. Krishan, S. Jha, J. Das, A. Singh, M. K. Goyal, and C. Sekar, Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India, *Air Quality, Atmosphere and Health*, vol. 12, no. 8, pp. 899–908, 2019.

[22] E. K. Wang, C. M. Chen, F. Wang, M. K. Khan, and S. Kumari, Joint-learning segmentation in Internet of drones (IoD)-based monitor systems, *Computer Communications*, vol. 152, pp. 54-62, 2020.

[23] S. Kumar, A. Damaraju, A. Kumar, S. Kumari, and C. M. Chen, LSTM Network for Transportation Mode Detection, *Journal of Internet Technology*, vol. 22, no. 4 , pp. 891-902, Jul. 2021.

[24] K. Wang, C. M. Chen, M. S. Hossain, G. Muhammad, S. Kumar, and S. Kumari, Transfer reinforcement learning-based road object detection in next generation IoT domain, *Computer Networks*, vol. 193, 2021.

[25] J. M. T. Wu, L. Sun, G. Srivastava, and J. C. W. Lin, A Novel Synergetic LSTM-GA Stock Trading Suggestion System in Internet of Things, *Mobile Information Systems 2021*, 2021, `https://doi.org/10.1155/2021/6706345`.