IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 18, 2019, accepted July 29, 2019, date of publication August 1, 2019, date of current version August 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2932445

Spatiotemporal Prediction of PM2.5 Concentrations at Different Time Granularities Using IDW-BLSTM

JUN MA[1,2], YUEXIONG DING[2], VINCENT J. L. GAN[1], CHANGQING LIN[1], AND ZHIWEI WAN[3]

[1]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong
[2]Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong
[3]Shenzhen Inovance Technology Company Ltd., Shenzhen, China

Corresponding author: Zhiwei Wan (wanzhiwei0213@outlook.com)

ABSTRACT As air pollution becomes an increasing concern globally, governments, and research institutions have attached great importance to air quality prediction to help give early warnings and prevent the impacts of air pollution. The existing prediction methods for air quality forecasting include deterministic methods, statistical methods, machine learning, and deep learning methods. Deep learning-based prediction methods have attracted much attention these years due to its high performance and powerful modeling capability. However, the majority of the deep learning methods only focus on the prediction of the places where there have monitoring stations, and limited studies have integrated deep learning to predict places without monitoring stations. To address the limitations, this paper proposes a new methodology framework combining a deep learning network, namely, bi-directional long short-term memory (BLSTM) network and the inverse distance weighting (IDW) technique for the spatiotemporal predictions of air pollutants at different time granularities. The BLSTM can effectively capture the long-term temporal mechanism of air pollution. The IDW layer, on the other hand, can consider the spatial correlation of air pollution and interpolate the spatial distribution. A case study is conducted to validate the effectiveness of the proposed methodology. The PM2.5 concentration at Guangdong, China is forecasted. Prediction performances of the LSTM network at hourly, daily, and weekly granularities and over different time spans are presented. Spatial distribution of the predicted PM2.5 concentrations and the prediction errors are analyzed. The experimental results demonstrate that the proposed method can achieve better prediction performance for the PM2.5 concentration compared with other models.

INDEX TERMS Air pollution, machine learning, neural networks, spatiotemporal phenomena, deep learning, long short-term memory, inverse distance weighting.

I. INTRODUCTION

Air pollution has now become one of the most significant environmental problems in the world. It will not only accelerate climate change but also pose a severe threat to the health systems of human beings. According to World Health Organization (WHO), around 4.2 million people die every year from exposure to ambient air pollution, and 9 out of 10 people in the world breathe air that contains high levels of pollutants [1]. In developing countries like India and China, air pollution is even worse due to the rapid urbanization and industrialization in the last decades [2]–[4].

Against this background, governments and research institutions have attached great importance to the prevention and control of air pollution. Although a large number of air quality monitoring stations have been built to observe the air pollutant concentration, there still are many places that require scientific predictions to report the air quality. It can help give early warnings, suggest outdoor activities and prepare control strategies. Many researchers have been working on proposing a better method to improve prediction accuracy [5]–[9]. Existing methods for air quality prediction can be divided into three major categories: deterministic methods, statistical methods, and machine learning methods. Deterministic methods deploy fundamental principles to simulate the dispersion and transport mechanisms of air pollution.

The associate editor coordinating the review of this manuscript and approving it for publication was Yungang Zhu.

Commonly seen deterministic methods include Operational Street Pollution Model (OSPM) [10], Community Multiscale Air Quality (CMAQ) model [11], Nested Air Quality Prediction Modeling System (NAQPMS) [12]. However, traditional deterministic methods were usually designed under specific assumptions and cannot fit other real-world conditions. The use of default parameters and the lack of real observations limit their performance [13].

Statistical methods, on the other hand, apply data-based models to predict air quality [5]. For example, Sharma *et al.* [14] forecasted the air pollution load in Delhi using time series regression models. Gupta and Christopher [15] used a multiple regression approach to monitor particulate matter air quality. Slini *et al.* [16] developed a stochastic Autoregressive Integrated Moving Average (ARIMA) model for maximum ozone concentration forecasts. Deng *et al.* [17] proposed a cellular automata (CA) model based on a multivariate regression model and several physical models to analyze the generation and diffusion of PM2.5. Compared with deterministic methods, statistical methods can be applied under a broader range of conditions as long as enough data are given. However, the linear assumption behind the traditional statistical methods is opposite to the non-linear characteristics of the real world, and this limits their performance.

To address this problem, many researchers started to adopt non-linear machine learning methods, such as Support Vector Machine (SVM), Artificial Neural Networks (ANN), Random Forest (RF) to predict the air quality. Among these machine learning methods, ANN has been one of the most widely used methods. For example, Li *et al.* [5] proposed the self-adaptive neuro-fuzzy weighted extreme learning machine (ANFIS-WELM) to predict the air pollutant concentration. Alimissis *et al.* [18] conducted a spatial estimation of urban air pollution with the use of ANN and Multiple Linear Regression (MLR). To improve the forecasting accuracy of air pollutant concentration, Yang and Wang [19] proposed a new hybrid model that combined Complementary Ensemble Empirical Mode Decomposition (CEEMD), Modified Cuckoo Search and Differential Evolution algorithm (MCSDE) and Elman Neural Network (ENN). Gardner and Dorling [20] trained multilayer perceptron (MLP) neural networks to model hourly pollutant concentrations in Central London. Niska *et al.* [21] adopted feed-forward neural networks for forecasting hourly concentrations of nitrogen dioxide at a traffic station.

Air pollution at one moment could have a short or long-term impact on future status. The influence might last for hours, days or even weeks. Therefore, when forecasting air quality, it is necessary to consider the time traverse. However, most of the ANN methods above fail to strengthen the time lag of air pollution or learn the long-term dependencies. This limits their performance to some extent. To overcome this problem, some studies adopted advanced deep learning techniques to model the time series data. Commonly seen networks include Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM). These methods were reported to have high performance in different domains. For example, Donahue *et al.* [22] used the LSTM networks for visual recognition and description. Fu *et al.* [23] applied LSTM to predict short-term traffic flow and achieved high performance. Shi *et al.* [24] predicted the future rainfall intensity in a local region over a relatively short period using the LSTM network. However, not so many studies in air quality prediction have applied LSTM as the regression method until recently. Zhao *et al.* [25] implemented a long short-term memory-fully connected (LSTM-FC) neural network to predict PM2.5 contamination of a specific air quality monitoring station over 48 h. Tong *et al.* [26] developed a bi-directional LSTM RNN network for spatiotemporal interpolation of air pollutants. Qi *et al.* [27] proposed Deep Air Learning (DAL) for the prediction, interpolation, and feature analysis of fine-grained air quality. Fan *et al.* [28] proposed a spatiotemporal prediction framework based on missing value processing algorithms and deep RNN.

Still, their results were not perfect. Most of them focus on the predictions where there have monitoring stations. Limited studies have extended their methods on places without stations. This is, in fact, a typical spatial interpolation/extrapolation problem of air quality predictions [29], [30]. However, traditional interpolation methods either overlooked the temporal effect from historical data, or just used subjectively defined liner/non-linear equations to define the complicated real-world temporospatial relationships [26]. Such a problem can be improved by deep learning techniques to some extent. However, limited studies have tried to address the traditional temporospatial interpolation problems with advanced deep learning technologies. Existing ones have different kinds of limitations that have not been well addressed. For example, Qi *et al.* [27] proposed a well-structured interpolation model but their results are not based on the latest LSTM models, and this may limit their performance. Tong *et al.* [26] proposed a deep learning based interpolation method but unfortunately, they did not provide any spatial interpolation results on places with no data records.

To overcome the limitations and fill the research gap, this paper combines the bi-directional Long Short-Term Memory (BLSTM) network and the Inverse Distance Weighting (IDW) for spatiotemporal prediction of the concentration of PM2.5 at different time granularities. A new IDW-BLSTM model is proposed and PM2.5 concentrations at different time granularities and over different time spans are predicted. Figure 1 shows the proposed research framework. Firstly, the raw data set is collected and the time series samples are built. Then the IDW-BLSTM model is constructed, and its parameters are optimized. Comparison of IDW-BLSTM network and five other algorithms is conducted. Results demonstrate that the proposed method outperforms other models in air pollution estimation. Next, PM2.5 concentrations at different time granularities and over different time spans are estimated. Results show that the predicted distribution is

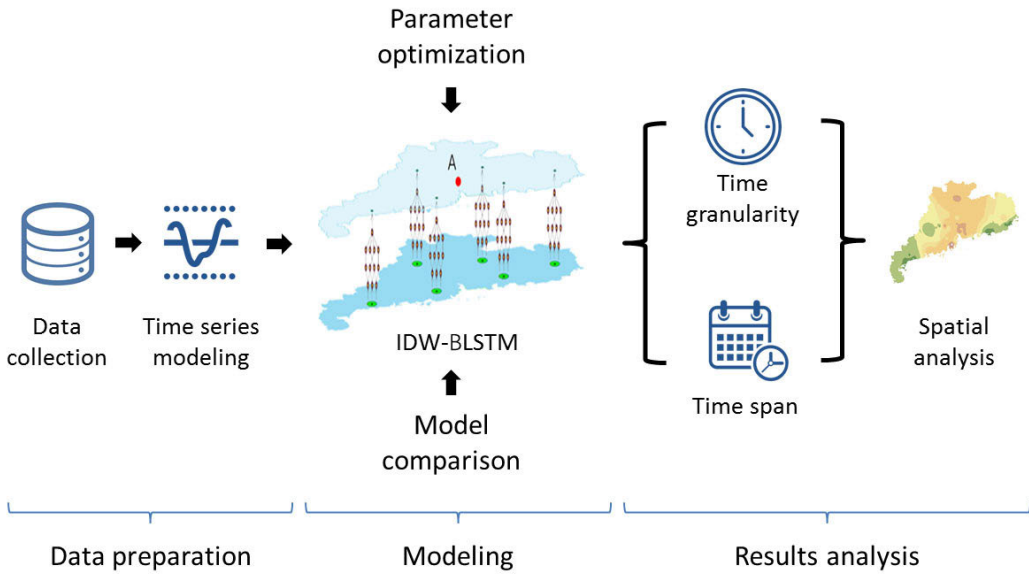

FIGURE 1. Research framework.

generally consistent with the real one. Spatial analysis of the PM2.5 concentrations in the study area is also conducted to provide practical suggestions to the local government.

The remained paper is organized as follows: Section 2 presents the collected data and sample modeling. Section 3 introduces the methods and model optimization. Results are analyzed and discussed in Section 4. Section 5 concludes the study.

II. DATA COLLECTION AND PREPROCESSING

A. DATA COLLECTION

To validate the effectiveness of the proposed research framework, we collected 2017/1/1-2017/12/31 one-year air quality data of the monitoring stations in Guangdong province, China. The province has 100 stations in total. Each station records 8,760 hourly PM2.5 concentration data. The distribution of the 100 stations is shown in Figure 2. Each green circle represents a station.

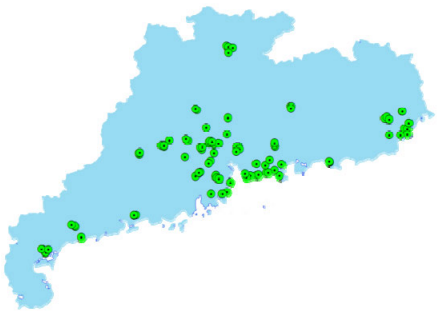

FIGURE 2. Distribution of the monitoring stations in Guangdong, China.

B. TIME SERIES SAMPLE MODELING

After data collection, preprocessing should be conducted. Since there is no missing value in the collected data, the remaining yet the most essential step is sample modeling. The data should be transformed in a way that can be implemented in machine learning/deep learning models. For time series problems, a typical way of modeling is called rolling-window [31].

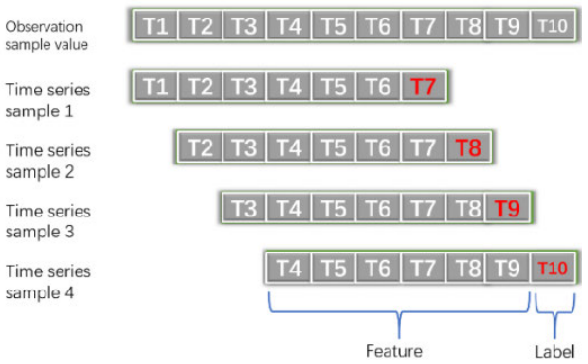

FIGURE 3. An example of time series samples modeling.

The rolling window method will construct one sample for each time record t. The sample for t_0 is constructed using the values within $[t_0 - \Delta t, t_0)$ as the features, and the value at t_0 as the label or target. Δt is called the window size. Figure 3 shows an example of how to build time-series samples. Assume there were 10 time-series records in the dataset, including T1, T2, …, T10. If $\Delta t = 6$, then for Sample 1, it has T1, T2, T3, T4, T5 and T6 as its features and T7 as its label. For Sample 2, it has T2, T3, T4, T5, T6 and T7 as its features and T8 as its label. Sample 3 and Sample 4 were given in a similar way. As a result, four time-series samples can be built when the time records are 10 and the window size is 6.

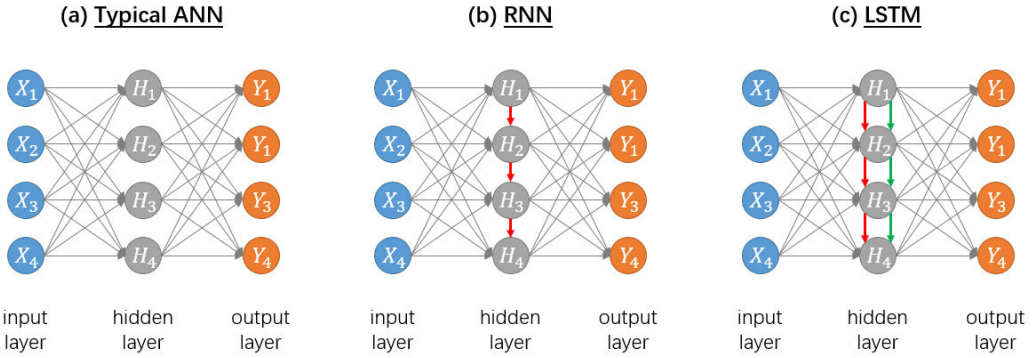

FIGURE 4. Typical examples of ANN, RNN, and LSTM.

In our dataset, each station has 8,760 hourly records. For convenience, we pre-define the window size Δt as 24, and there will be 8,736 time-series samples in total for each station. More optimal window sizes for different time granularities will be identified later.

III. MODELING AND OPTIMIZATION

A. FROM ANN TO RNN

As mentioned above, an IDW-BLSTM model is proposed as the main regression algorithm in this study. It is developed based on LSTM and IDW. Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) specially designed to prevent the neural network output for a given input from either decaying or exploding as it cycles through the feedback loops [32]. It effectively accounts for the long-term dependencies and has been applied to a lot of fields, such as meteorological forecasting [33], weather forecasting [34], network traffic prediction [35], air pollution forecasting [36], etc. LSTM is developed based on Artificial Neural Network (ANN) and Recurrent Neural Network (RNN). Therefore, to better understand LSTM, it is necessary to start from ANN and RNN.

ANN was initially introduced in the 1970s, but its importance wasn't fully appreciated until a famous paper by David Rumelhart, Geoffrey Hinton and Ronald Williams [37]. It is a multilayer feedforward neural network using the technique of backward propagation of errors. One typical structure is presented in Figure 4 (a). It can be seen that the network consists of an input layer, hidden layers, and an output layer. Usually, it has only one input layer and one output layer but one or more hidden layers [37]. Each neuron is a weight calculation of the inputs from the last layer and followed by an activation function.

However, when it comes to modeling time-series data, traditional ANN models fail to relate the information of the previous moment to the next moment [38]. To resolve the issue, scholars proposed Recurrent Neural Network (RNN). Compared to traditional ANN, RNN is equipped with a novel connection method. As shown in Figure 4 (b), it takes the output of the previous moment as the input of the next moment to affect the weights at the next moment. In this way, it exhibits temporal dynamic behavior for a time sequence and improves the prediction accuracy.

B. LSTM AND BI-DIRECTIONAL LSTM

However, when handling time series problems, RNN is not capable of capturing the long-term dependencies in the input sequences. It may also cause vanishing gradient and exploding gradient problems. To address these, Hochreiter and Schmidhuber proposed the Long Short-Term Memory (LSTM) network [38]. The specialty in LSTM is that it adds self-connected units which allow a value (forward pass) or gradient (backward pass) that flows into the unit to be preserved and subsequently retrieved at the required time step [32]. As shown in Figure 4 (c), the unique connections between neurons are shown as the green arrows. They are used to control the cell states, and help maintain information in memory for long periods of time.

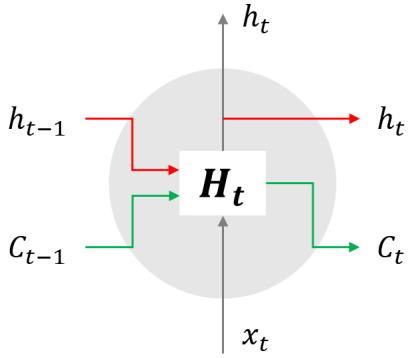

FIGURE 5. Inputs and outputs of an LSTM cell H_t.

Figure 5 shows the inputs and outputs of an LSTM cell. h_t and C_t are the outputs. h_{t-1} and C_{t-1} are the inputs from the last cell, and x_t is the input from the last layer. Equation (1) presents its mathematical format.

$$(h_t, C_t) = H_t(x_t, h_{t-1}, C_{t-1}) \tag{1}$$

where H_t is the whole function within the cell. It consists of three parts, namely the forget gate, the input gate, and the

output gate. Forget gate determines the extent to which the information of the last unit remains in the present cell. This gate is the main difference between typical ANN and LSTM since it records the influence from long and short memories. As shown in Equation (2), it checks h_{t-1} and x_t, and outputs a number between 0 and 1 for each number in the cell state C_{t-1} to represent which information to "remember" when calculating this cell.

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

where f_t is the output of the forget gate, W and b are the relevant parameters, σ is the sigmoid function.

The next part is the input gate. It determines which new information should be stored by updating the cell state from C_{t-1} to C_t. It is calculated based on Equation (3) and (4), where i_t is the output of the input gate.

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_C \cdot [x_t, h_{t-1}] + b_C) \tag{3}$$
$$i_t = \sigma(W_i \cdot [x_t, h_{t-1}] + b_i) \tag{4}$$

After having the updated status, the output gate can help to calculate the outputs of the cell, h_t, by Equation (5) and (6).

$$O_t = \sigma(W_O \cdot [x_t, h_{t-1}] + b_O) \tag{5}$$
$$h_t = O_t \odot \tanh(C_t) \tag{6}$$

However, the LSTM network is still not perfect. The current version only supports forward learning and optimization. In practices, the values in later times can also help adjust the original predictions at earlier times. For example, in speech recognition, a later sentence may help adjust the translation of some earlier words. As a result, a backward LSTM prediction can be very helpful.

Under such a problem, Graves and Schmidhuber [39] designed the bi-directional LSTM/RNN (BLSTM/BRNN) models. Several studies have pioneered its applicability in air quality forecasting. For example, Tong *et al.* [26] implemented BLSTM in predicting the PM2.5 concentrations in Florida, the U.S. However, the way they designed the geo coordinates into time series may reduce the efficiency of modeling and optimization. This study proposes a newly designed network based on BLSTM and traditional spatial interpolation techniques like IDW to predict the air quality in the whole region including the places where there are no monitoring stations.

Different from ordinary LSTM, BLSTM will have a set of LSTM units to backwardly learn and optimize the time series. In this case, it will double the original units, and the additional LSTM units will train the inputs in a reverse sequence. These two groups of LSTM units will output two different values, which are donated as $\vec{h}$, and $\overleftarrow{h}$. As shown in Equation (7), the eventual outputs of a BLSTM unit are the sum of these two predictions.

$$h_{bi}^{(t)} = \vec{h}(t) + \overleftarrow{h}(t) \tag{7}$$

C. IDW-BLSTM

As discussed in the introduction, although deep learning-based techniques have been more and more popular in air quality research. Limited studies have explored the possibility of combining traditional spatial interpolation methods like IDW to predict the air quality in the whole region including the places where there are no monitoring stations. To address this gap, this study proposes the Inverse Distance Weighting BLSTM (IDW-BLSTM) network for spatial-temporal air quality prediction.

As shown in Figure 7, the main idea of IDW-BLSTM is integrating spatially allocated BLSTM predictors by introducing an IDW layer on top of them. Different from the spatial first mechanism in CNN-LSTM [40], IDW-BLSTM is temporal modeling first and then spatial. This is in line with the discoveries in related studies that, the temporal correlation is stronger than the spatial correlation for air quality prediction [26]. The weights in the IDW layer are combined using inversed distance and additional neural network weight vectors. In this way, the method can utilize the back-propagation techniques in neural networks to update and optimize the weights.

The proposed structure of IDW-BLSTM is shown in Figure 6. It consists of five kinds of layers, including the input layer, the BLSTM layers, the fully connected layer, the IDW layer, and the output layer. The number of input layer to FC layer chains equals to the number of monitoring stations. This could largely reduce the calculation in training compared with similar complexity CNN-LSTM/ConvLSTM and fine-grid modeling.

For example, for point A, IDW-BLSTM predicts its air quality values by de-normalizing the output value, as shown in Equation (8).

$$\mathrm{Pred_A} = \mathrm{f_D}\,(\mathrm{O_{IDW}}) \tag{8}$$

where O_{IDW} is the output of the IDW layer, which is calculated by Equation (9).

$$\mathrm{O_{IDW}} = \sigma(\mathrm{W_{IDW}} \cdot \mathrm{I_{IDW}} + \mathrm{b_{IDW}}) \tag{9}$$

where σ is the sigmoid function, W_{IDW} and b_{IDW} are the neural network parameters, and can be optimized by back-propagation. $I_{IDW} = [i_1^{IDW}, i_2^{IDW}, \ldots, i_n^{IDW}]$ is the input of the IDW layer, and it is given by Equation (10). n is the number of stations.

$$\mathrm{i_i^{IDW}} = \frac{\mathrm{Z_i} \cdot \mathrm{w_i}}{\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{w_j}} \tag{10}$$

where Z_i is the output value of the i^{th}-station ordinary BLSTM network. w_i is the weight assigned to x_i. The weight is calculated based on the inverse distance value, the calculation of which is shown in Equation (11).

$$\mathrm{w_i} = \frac{1}{\mathrm{d(x, x_i)^p}} \tag{11}$$

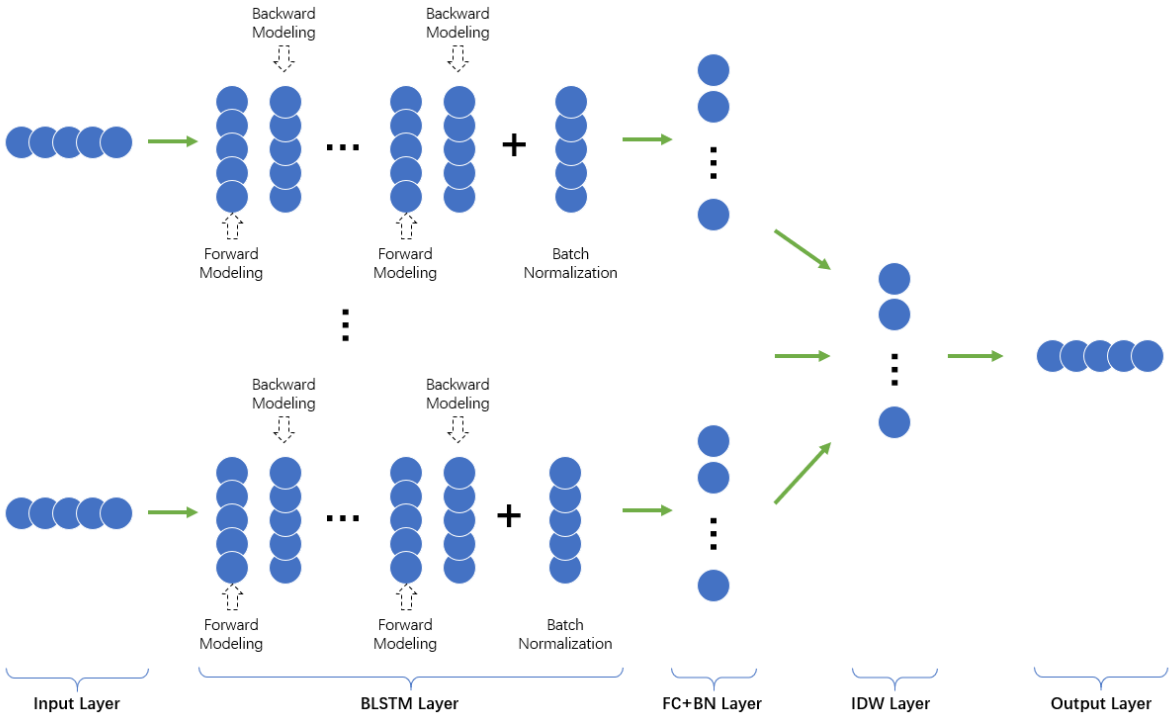

FIGURE 6. Structure of the proposed IDW-BLSTM.

where $d(x, x_i)$ is the distance between point x and x_i, p is the power factor used to adjust the sensitivity from the known value and is set as 1 in this study.

IV. RESULTS AND DISCUSSION

A. PARAMETER OPTIMIZATION

The parameters of the IDW-BLSTM model need to be tuned for better results. Important parameters in this study include N_L^{BLSTM}-the number of layers in the BLSTM part, N_N^{BLSTM}-the number of neurons in each BLSTM layer, N_N^{FC}-the number of neurons in the FC layer, N_N^{IDW}-the number of neurons in the IDW layer, batch size, dropout rate, etc.

After referring to the literature [41]–[43] and some initial tests, we set the candidates for the parameters as $N_L^{BLSTM} = \{1, 2, 3, 4\}$, $\{N_N^{BLSTM}, N_N^{FC}, N_N^{IDW}\} = \{16, 32, 64, 128, 256\}$, batch size $= \{32, 64, 128, 256, 512\}$, dropout rate $= \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. When optimizing the parameters, we preset the window size as 24, epoch as 200, and learning rate as 0.015. 70% of the shuffled samples are used as the training set, while the remained 30% are for testing. The parameters are optimized by the integration of the Mini-Batch Gradient Descent algorithm, dropout neuron algorithm and L2 regularization algorithm. RMSE was used as the optimization criteria. Its calculation is shown in Equation (12).

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y_i} - \mathrm{y_i^*}\right)^2} \tag{12}$$

where y_i is the observed value of the i^{th} case and y_i^* is the predicted value. A smaller RMSE means better performance.

After two rounds of one-sided grid search test, it is discovered that when $N_L^{BLSTM} = 2$, $N_N^{BLSTM} = 64$, $N_N^{FC} = 256$, $N_N^{IDW} = 128$, batch size $= 128$, dropout rate $= 0.1$, the model was able to get an optimal result with RMSE $= 8.24$.

B. ALGORITHM COMPARISON

To further evaluate the performance of the proposed method, we compared it with other algorithms, including Autoregressive Integrated Moving Average (ARIMA), ElasticNet, Support Vector Regression (SVR), Gradient Boosting Decision Tree (GBDT), Artificial Neural Network (ANN) [44], [45], Recurrent Neural Network (RNN), ordinary LSTM, BLSTM, Convolutional Neural Network-LSTM (CNN-LSTM). To evaluate the performance of these algorithms, this study implemented three different indicators, including root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). Calculations of these three indicators are presented in Equation (12), (13) and (14), respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i^*\right| \tag{13}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - y_i^*\right|}{y_i} \tag{14}$$

where n is the number of cases, y_i is the observed value of the i^{th} case and y_i^* is the predicted value of the i^{th} case.

A smaller value of these indicators means better performance in prediction.

These three indicators evaluate the prediction performance of the models from different angles. RMSE reflects the gap between the real value and the estimated value, and shows the sensitivity of the models to huge errors. MAE also evaluates the gap between the two values, but it reflects the robustness of models. MAPE represents the percentage of the error over the real value.

TABLE 1. Comparison of different algorithms.

Algorithm	RMSE	MAE	MAPE(%)
ARIMA	14.59	13.95	26.58
ElasticNet	13.95	13.32	24.65
SVR	13.87	12.83	25.27
GBDT	12.10	9.59	21.32
ANN	11.20	8.31	19.42
RNN	10.34	5.24	13.73
LSTM	8.98	5.01	10.45
BLSTM	8.73	5.03	9.56
CNN-LSTM	8.40	4.88	8.59
IDW-BLSTM	8.24	4.80	9.01

The calculation results are shown in Table 1. It can be seen that IDW-BLSTM, CNN-LSTM, BLSTM, ordinary LSTM, and RNN have better performance than other algorithms. This reflects that (1) the networks designed for time series problems outperform typical machine learning algorithms in air quality predictions. Also, LSTM based algorithms have lower error measures than RNN means (2) the longer-term dependency does have an important impact on PM2.5 concentrations. (3) BLSTM was again proved to perform better than ordinary LSTM [26] due to its bi-directional modeling concept. (4) The newly designed IDW layer additional improved the performance of BLSTM by 5.6% (comparing BLSTM and IDW-BLSTM), and this proved the effectiveness of the proposed model. (5) CNN-LSTM and IDW-BLSTM perform the best because these two algorithms not only considered the temporal correlation on the historic values but also included the spatial correlation from nearby stations. (6) The proposed IDW-BLSTM obtained the lowest RMSE and MAE, and this proved the effectiveness of the proposed method.

C. OPTIMAL WINDOW SIZE FOR DIFFERENT TEMPORAL GRANULARITIES

In this study, another objective is to explore the air quality prediction at different time granularities. The optimal window size Δt can be different for different time granularities. This parameter is important to time series problems and needs to be optimized. It is because many time-series problems follow periodic changes to some extent. This means the values at

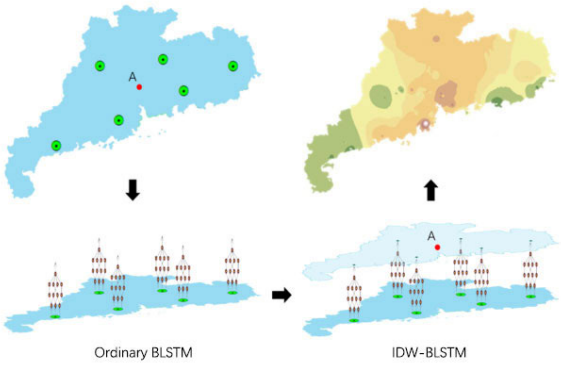

FIGURE 7. Graphical illustration of the IDW-BLSTM network.

previous moments have a lagged effect on the value at the next moment. The lagged effect might be strong in the short term or weak in the long term. Therefore, when predicting the air pollutant concentration, the influence of the window size in different lengths needs to be considered. A smaller window size Δt cannot guarantee enough long-term memory inputs for IDW-BLSTM model while a larger window size Δt will increase unrelated inputs and the computation complexity [46]. Hence, it is necessary to identify the most appropriate window size.

To tackle this problem, this study adopted the auto-correlation function. This function helps determine the time correlation among the time series data itself. Larger auto-correlation coefficients mean stronger time correlations and stronger lagged effects. Calculation of the auto-correlation coefficient at time lag Δt is shown in Equation (15) [47].

$$\rho_{\Delta t} = \frac{\sum_{t=\Delta t+1}^{T} (Y_t - \bar{Y})(Y_{t-\Delta t} - \bar{Y})}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2} \tag{15}$$

where Y_t is the data set sorted by ascending date, and $Y_{t-\Delta t}$ represents the same data set with Δt time lag. $\bar{Y}$ is the average value.

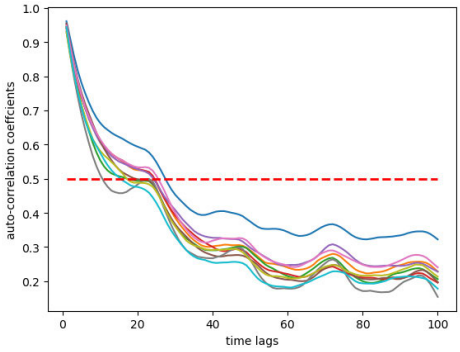

FIGURE 8. The autocorrelation coefficients of the stations with respects to different time lag values.

Figure 8 shows the autocorrelation coefficients of the stations in Guangdong. The horizontal axis represents the time lag, and the vertical axis represents the autocorrelation coefficients. It can be observed from Figure 8 that when

TABLE 2. Performance of different window sizes at hour granularity.

Window size	RMSE	MAE	MAPE(%)	Window size	RMSE	MAE	MAPE(%)
1	9.03	5.96	10.32	13	9.05	5.12	9.26
2	8.52	4.61	8.37	14	9.63	5.67	9.81
3	8.93	4.96	8.95	15	7.73	4.29	9.01
4	8.53	4.88	9.20	16	8.17	4.59	9.28
5	7.41	3.91	8.12	17	9.15	5.21	9.31
6	7.65	4.25	8.23	18	9.11	5.13	9.03
7	10.03	5.96	10.32	19	9.52	5.56	9.65
8	8.22	4.41	8.87	20	8.19	4.69	9.39
9	8.93	4.96	8.95	21	9.83	5.78	9.88
10	8.34	4.66	8.65	22	7.99	4.57	9.32
11	9.10	5.2	9.82	23	9.28	5.47	9.76
12	8.97	5.12	9.13	24	8.24	4.80	9.01

time lag increases, the coefficient shows a descending trend. This confirms that earlier events have a weaker influence on the current status. Furthermore, when the time lag is smaller than 24, auto-correlation coefficients of most stations are higher than 0.5, which indicates the time correlations are stronger. Therefore, the range of time lag Δt for hourly granularity is pre-set as [1], [24].

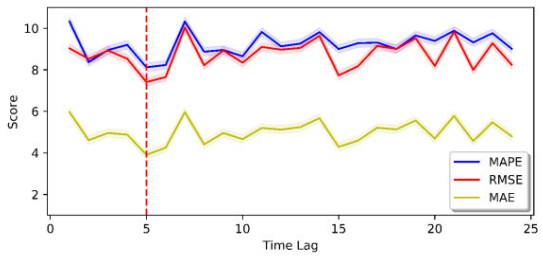

FIGURE 9. Scores of three indicators with different window sizes.

To further identify the optimal window sizes at hourly granularity, we tested the prediction performance using different Δt within [1], [24]. 70% of the data are taken as the training set and the remaining 30% as the testing set. Results of the estimation performance of the IDW-BLSTM network are then shown in Table 2. When the window size varies from 1 to 24, RMSE varies around 8 to 9, MAE around 5 and MAPE around 8 to 10. To better observe the change of the three indicators, we presented their scores in a line chart, which is shown in Figure 9. The horizontal axis shows the window size and the vertical axis represents the score. The blue line represents the scores of MAPE, the red line RMSE and the yellow line MAE. It can be observed that the trends of these three lines are similar. When the window size is smaller than 5, the scores decrease as the window size increases. This is because when the window size is too small, the input features would have too limited information to have a lower error score. On the other hand, when the window size is larger than 5, the error scores increase a bit and then gradually become stable. This may because when the window size is large enough, increasing its value will add more noise and therefore interfere with the performance. As a result, it can be concluded that the optimal window size for the hourly granularity is 5.

TABLE 3. Performance of different window sizes at daily granularity.

Window size	RMSE	MAE	MAPE(%)
1	15.81	11.46	26.60
2	14.41	10.40	25.51
3	13.28	9.17	26.69
4	14.50	9.78	27.42
5	12.03	8.49	24.32
6	13.38	8.85	24.92
7	13.58	7.04	23.30

D. DAILY GRANULARITY AND WEEKLY GRANULARITY

When it comes to daily granularity and weekly granularity, the number of cases becomes smaller as there are only 365 days and 52 weeks in the collected data. Therefore, this study set the range of window size as (1, 7). Results of the prediction performance at the daily granularity and weekly granularity are presented in Table 3 and Table 4 respectively. It can be seen that when the window size is 5, all the error indicators at the daily granularity and weekly granularity reach the smallest scores. Therefore, 5 is the optimal window size at these two granularities.

TABLE 4. Performance of different window sizes at the weekly granularity.

Window size	RMSE	MAE	MAPE(%)
1	20.33	19.58	34.39
2	19.17	16.67	33.43
3	17.03	16.35	32.27
4	16.73	15.60	31.86
5	15.70	12.84	30.25
6	16.83	16.35	31.10
7	18.35	17.15	31.73

On the other hand, compared with Table 2, values of RMSE, MAE, and MAPE are much higher at daily/weekly granularities than those at hourly granularity. This means the prediction accuracy of LSTM decreases as the time dimensions expand from hours to weeks. The reason behind is that the size of data used for training at daily and weekly granularities is far less than that used at hourly granularity. Consider that the time span of the collected data is one year, and the minimum unit is an hour. In this case, we have 8760 data records for hourly granularity, while only 365 records for daily granularity and 52 records for weekly granularity. Also, PM2.5 values at larger granularity are more difficult to predict. Because other factors such as weather, climate, and urban activities, may have more substantial impacts on values at larger granularity. Those data would be collected to further test the performance of the proposed method in future studies.

E. SPATIAL ANALYSIS

The experiments in previous sections discussed the feasibility of implementing the proposed methodology and its measurable performance. Another important feature of the proposed method is that it inherited the interpolation method from the traditional IDW method. Therefore, it can be further utilized to calculate and predict the air quality at places where there is no monitoring station. This is very practical since researchers and managers would be able to have a more integrated idea on the overall performance of the air quality in the surrounding regions.

Figure 10 presents the calculated prediction diagrams of the PM2.5 concentration in Guangdong province using the proposed method. It is drawn by the predicted value of each pixel at 12 am, December 30th, 2017. The PM2.5 concentrations in the central part of the Guangdong province are relatively higher than other parts of the province, especially Dongguan city and Zhuhai city. These two cities are therefore suggested to put more effort into their air quality problems.

Furthermore, this study also examined the RMSE distribution of the predicted PM2.5 concentration. Three IDW diagrams of the average RMSE calculated at the record stations using the proposed method are shown in Figure 11. Different diagrams represent the RMSE distribution at

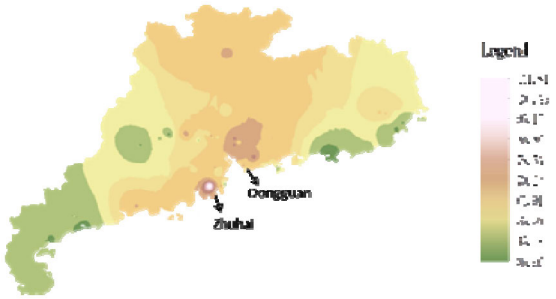

FIGURE 10. Prediction diagram of the PM2.5 concentration in Guangdong province using IDW-BLSTM.

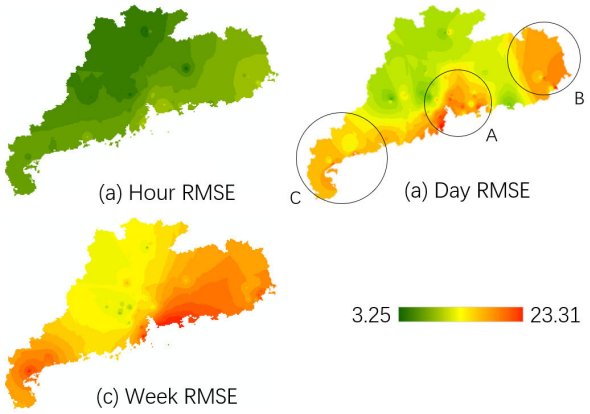

FIGURE 11. The RMSE distribution of IDW-BLSTM at hourly, daily and weekly granularities.

different temporal granularities. Green means lower RMSE and red means higher RMSE. Figure 11 (a) presents the RMSE distribution at hourly granularity, (b) at daily granularity, and (c) at weekly granularity. It can be concluded from Figure 11 that the predictions of PM2.5 concentrations at smaller granularities have lower errors than larger granularities. This is in line with the results in previous sections.

In addition, the calculated diagram of the RMSE distributions can help the government to plan new stations. For example, it can be seen from Figure 11 (b) that the prediction error in area A, B and C are relatively higher than other places at daily granularity. This reflects that the PM2.5 concentrations at these three places are more complicated to predict and require more additional information to help. Therefore, if the government wants to improve the prediction accuracy at the daily granularity of the whole province, more stations could be added to enhance the analysis. Especially for area A, by referring to Figure 10, it can be seen that it is not only a high error district but also a high PM2.5 area. So it is strongly suggested that the government put more effort into studying the PM2.5 concentrations in that area.

On the other hand, it is also understandable that area A has such a different numerical pattern. This district is the Greater Bay area in Guangdong Province, and it is one of the fastest developing areas in China. The central part is Dongguan city, which has a high-ranking manufacturing industry nationwide. The mid-east part is Shenzhen city. It is one of the top 5 fast-growing cities in China, and is famous for its electronic

device market and industry. The complicated PM2.5 concentration there may result from the fast-developing environment. Although economic growth is an important concern in urban management, the environment problem, such as air quality or PM2.5, is also a critical issue that needs enough attention.

V. CONCLUSION

To conclude, this paper proposed a new methodology framework that combines LSTM network and IDW techniques to predict the air pollutant concentration based on historical records. To validate the effectiveness of the proposed framework, a case study in Guangdong, China, was conducted. SVR, GBDT, ANN, RNN, and ordinary LSTM were compared with the proposed method. The prediction performance of IDW-BLSTM at different time granularities and time spans were evaluated. Contributions of this study can be summarized as follows:

1. To the best of our knowledge, this is the first study that tried to integrate the traditional spatial interpolation methods like IDW and advanced deep learning techniques like BLSTM to explore the temporospatial prediction problem. Results show that the proposed network has good performance compared with other latest methods.
2. The implementation of the IDW layer into the neural networks helped the performance of BLSTM improve by 5.6%. This may because the IDW layer helped consider the spatial correlation and impact from other stations into the modeling. This is, on the other hand, supported the advantage of the proposed model.
3. The prediction results of PM2.5 concentration at smaller temporal granularities have smaller errors than larger temporal granularities.
4. According to the case study, two cities, Zhuhai and Dongguan, are suggested to put more effort into their PM2.5 management since the PM2.5 concentrations there are very high.
5. In terms of station planning, areas such as A, B, and C in Figure 11 are suggested to install more monitoring stations since the air quality there is more complicated to predict.

Overall, the proposed methodology framework was able to provide valid results in PM2.5 prediction problems, and further experiments can be conducted to verify its feasibility in studying other pollution sources in other places. On the other hand, this paper has some limitations. Due to the data availability, we only used the historic air pollution data and did not include the meteorological and urban information. In the future, more relevant features should be collected to improve prediction and analysis.

REFERENCES

[1] World Health Organization. *9 Out of 10 People Worldwide Breathe Polluted Air, But More Countries Are Taking Action*. Accessed: Sep. 6, 2018. [Online]. Available: http://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action
[2] L. Morawska, J. Chen, T. Wang, and T. Zhu, "Preface on air quality in China," *Sci. Total Environ.*, vols. 603–604, p. 26, Dec. 2017.
[3] J. Wang, W. Qu, C. Li, C. Zhao, and X. Zhong, "Spatial distribution of wintertime air pollution in major cities over eastern China: Relationship with the evolution of trough, ridge and synoptic system over East Asia," *Atmos. Res.*, vol. 212, pp. 186–201, Nov. 2018.
[4] W. Yanjia and H. Kebin, "The air pollution picture in China," *IEEE Spectr.*, vol. 36, no. 12, pp. 55–58, Dec. 1999.
[5] Y. Li, P. Jiang, Q. She, and G. Lin, "Research on air pollutant concentration prediction method based on self-adaptive neuro-fuzzy weighted extreme learning machine," *Environ. Pollut.*, vol. 241, pp. 1115–1127, Oct. 2018.
[6] Y. Zhou, S. De, G. Ewa, C. Perera, and K. Moessner, "Data-driven air quality characterization for urban environments: A case study," *IEEE Access*, vol. 6, pp. 77996–78006, 2018.
[7] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang, and L. Huang, "A predictive data feature exploration-based air quality prediction approach," *IEEE Access*, vol. 7, pp. 30732–30743, 2019.
[8] J. Ma and J. C. P. Cheng, "Identification of the numerical patterns behind the leading counties in the U.S. local green building markets using data mining," *J. Cleaner Prod.*, vol. 151, pp. 406–418, May 2017.
[9] J. Ma and J. C. P. Cheng, "Data-driven study on the achievement of LEED credits using percentage of average score and association rule analysis," *Building Environ.*, vol. 98, pp. 121–132, Mar. 2016.
[10] R. Berkowicz, "OSPM—A parameterised street pollution model," *Environ. Monit. Assess*, vol. 65, no. 1, pp. 323–331, 2000.
[11] L. Li, C. H. Chen, C. Huang, H. Y. Huang, G. F. Zhang, Y. J. Wang, H. L. Wang, S. R. Lou, L. P. Qiao, M. Zhou, M. H. Chen, Y. R. Chen, D. G. Streets, J. S. Fu, and C. J. Jang, "Process analysis of regional ozone formation over the Yangtze River Delta, China using the Community Multi-scale Air Quality modeling system," *Atmos. Chem. Phys.*, vol. 12, no. 22, pp. 10971–10987, Nov. 2012.
[12] Z. Wang, T. Maeda, M. Hayashi, L.-F. Hsiao, and K.-Y. Liu, "A nested air quality prediction modeling system for urban and regional scales: Application for high-ozone episode in Taiwan," *Water, Air, Soil Pollut.*, vol. 130, nos. 1–4, pp. 391–396, Aug. 2001.
[13] A. Suleiman, M. R. Tight, and A. D. Quinn, "Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM_{10} and $PM_{2.5}$)," *Atmos. Pollut. Res.*, vol. 10, no. 1, pp. 134–144, Jan. 2018.
[14] N. Sharma, S. Taneja, V. Sagar, and A. Bhatt, "Forecasting air pollution load in Delhi using data analysis tools," *Procedia Comput. Sci.*, vol. 132, pp. 1077–1085, Jan. 2018.
[15] P. Gupta and S. A. Christopher, "Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach," *J. Geophys. Res. Atmos.*, vol. 114, no. D14, pp. 1–13, Jul. 2009.
[16] T. Slini, K. Karatzas, and N. Moussiopoulos, "Statistical analysis of environmental data as the basis of forecasting: An air quality application," *Sci. Total Environ.*, vol. 288, pp. 227–237, Apr. 2002.
[17] F. Deng, L. Ma, X. Gao, and J. Chen, "The MR-CA models for analysis of pollution sources and prediction of $PM_{2.5}$," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 4, pp. 814–820, Apr. 2019.
[18] A. Alimissis, K. Philippopoulos, C. G. Tzanis, and D. Deligiorgi, "Spatial estimation of urban air pollution with the use of artificial neural network models," *Atmos. Environ.*, vol. 191, pp. 205–213, Oct. 2018.
[19] Z. Yang and J. Wang, "A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction," *Environ. Res.*, vol. 158, pp. 105–117, Oct. 2017.
[20] M. W. Gardner and S. R. Dorling, "Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London," *Atmos. Environ.*, vol. 33, no. 5, pp. 709–719, Feb. 1999.
[21] H. Niska, T. Hiltunen, A. Karppinen, J. Ruuskanen, and M. Kolehmainen, "Evolving the neural network model for forecasting air pollution time series," *Eng. Appl. Artif. Intell.*, vol. 17, no. 2, pp. 159–167, Mar. 2004.
[22] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2625–2634.
[23] R. Fu, Z. Zhang, and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," in *Proc. 31st Youth Academic Annu. Conf. Chin. Assoc. Automat. (YAC)*, Nov. 2016, pp. 324–328.
[24] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 802–810.

[25] J. Zhao, F. Deng, Y. Cai, and J. Chen, "Long short-term memory—Fully connected (LSTM-FC) neural network for $PM_{2.5}$ concentration prediction," *Chemosphere*, vol. 220, pp. 486–492, Apr. 2019.

[26] W. Tong, L. Li, X. Zhou, A. Hamilton, and K. Zhang, "Deep learning $PM_{2.5}$ concentrations with bidirectional LSTM RNN," *Air Qual., Atmos. Health*, vol. 12, no. 4, pp. 411–423, 2019.

[27] Z. Qi, T. Wang, G. Song, W. Hu, X. Li, and Z. Zhang, "Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 12, pp. 2285–2297, Dec. 2018.

[28] J. Fan, Q. Li, J. Hou, X. Feng, H. Karimian, and S. Lin, "A spatiotemporal prediction framework for air pollution based on deep RNN," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 15–22, Aug. 2017.

[29] J. Li and A. D. Heap, "A review of spatial interpolation methods for environmental scientists," Geosci. Aust., Rec. 2008/23, 2008, pp. 137.

[30] J. Li and A. D. Heap, "A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors," *Ecological Informat.*, vol. 6, nos. 3–4, pp. 228–241, 2011.

[31] E. Zivot and J. Wang, Eds., "Rolling analysis of time series," in *Modeling Financial Time Series with S-Plus*. New York, NY, USA: Springer, 2006, pp. 313–360.

[32] NVIDIA Developer. (Apr. 23, 2018). *Long Short-Term Memory (LSTM)*. Accessed: Sep. 10, 2018. [Online]. Available: https://developer.nvidia.com/discover/lstm

[33] Y.-L. Hu and L. Chen, "A nonlinear hybrid wind speed forecasting model using LSTM network, hysteretic ELM and differential evolution algorithm," *Energy Convers. Manage.*, vol. 173, pp. 123–142, Oct. 2018.

[34] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," *Energy*, vol. 148, pp. 461–468, Apr. 2018.

[35] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intell. Transp. Syst.*, vol. 11, no. 2, pp. 68–75, 2017.

[36] T.-C. Bui, V.-D. Le, and S.-K. Cha, "A deep learning approach for forecasting air pollution in South Korea Using LSTM," Apr. 2018, *arXiv:1804.07891*. [Online]. Available: https://arxiv.org/abs/1804.07891

[37] M. A. Nielsen, *Neural Networks and Deep Learning*, vol. 25. San Francisco, CA, USA: Determination Press, 2015.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, 2005.

[40] Y. Qi, Q. Li, H. Karimian, and D. Liu, "A hybrid model for spatiotemporal forecasting of $PM_{2.5}$ based on graph convolutional neural network and long short-term memory," *Sci. Total Environ.*, vol. 664, pp. 1–10, May 2019.

[41] J. C. P. Cheng and L. J. Ma, "A non-linear case-based reasoning approach for retrieval of similar cases and selection of target credits in LEED projects," *Building Environ.*, vol. 93, pp. 349–361, Nov. 2015.

[42] J. C. P. Cheng and L. J. Ma, "A data-driven study of important climate factors on the achievement of LEED-EB credits," *Building Environ.*, vol. 90, pp. 232–244, Aug. 2015.

[43] M. A. Jun and J. C. P. Cheng, "Selection of target LEED credits based on project information and climatic factors using data mining techniques," *Adv. Eng. Inform.*, vol. 32, pp. 224–236, Apr. 2017.

[44] J. Ma and J. C. P. Cheng, "Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology," *Appl. Energy*, vol. 183, pp. 182–192, Dec. 2016.

[45] J. Ma and J. C. P. Cheng, "Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests," *Appl. Energy*, vol. 183, pp. 193–201, Dec. 2016.

[46] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi, "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation," *Environ. Pollut.*, vol. 231, pp. 997–1004, Dec. 2017.

[47] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.

JUN MA received the Ph.D. degree from the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2016. He is currently the Chief Research Officer with the Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong. His research interests include smart city, urban computing, data mining, and artificial intelligence.

YUEXIONG DING received the B.S. degree from the School of Pharmaceutical Information Engineering, Guangdong Pharmaceutical University, Guangdong, China, in 2017. He is currently pursuing the M.S. degree with the School of Technology, Shantou University, Guangdong, China. He is also a Researcher with the Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong. His research interests include data mining, artificial intelligence, and smart city.

VINCENT J. L. GAN received the M.Sc. and Ph.D. degrees from the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology (HKUST), in 2016, where he was a Research Assistant Professor of civil engineering. His research interests include tall building, optimum design, performance optimization, energy efficiency, indoor human comfort, and construction informatics.

CHANGQING LIN received the Ph.D. degree from the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology (HKUST), in 2016, where he is currently a Research Associate with the Institute for the Environment (IENV). He specializes in atmospheric environment, lidar remote sensing, satellite remote sensing, and air pollution exposure and health impacts.

ZHIWEI WAN received the M.S. degree from the School of Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2014. He is currently a Researcher and an Algorithm/Software Engineer with Shenzhen Inovance Technology Company Ltd., Shenzhen, China. His research interests include the Internet of Things, smart city, and urban computing.