Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2858471, 13 pages
https://doi.org/10.1155/2020/2858471

Research Article

Probabilistic Forecasting Method of Metro Station Environment Based on Autoregressive LSTM Network

Qing Tian ,[1] Bo Li ,[1] Hongquan Qu ,[1] Liping Pang ,[2] Weihang Zhao,[1] and Yue Han[1]

[1]*School of Information Science and Technology, North China University of Technology, Beijing 100144, China*
[2]*School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China*

Correspondence should be addressed to Hongquan Qu; qhqphd@ncut.edu.cn and Liping Pang; pangliping@buaa.edu.cn

Received 12 March 2020; Revised 16 May 2020; Accepted 25 May 2020; Published 12 June 2020

Academic Editor: Javier Martinez Torres

With the increasing number of metros, the comfort and safety of crew and passengers in metro stations have been paid great attention. The environment forecasting has become very important for decision-making. The outputs of the traditional point prediction methods are some exact values in the future. However, it might be closer to the real conditions that the predicted variables are given a probability range with a different confidence rather than exact values. This paper proposes a probabilistic forecasting method of metro station environment based on autoregressive Long Short Term Memory (LSTM) network. It has a good performance to quantify the uncertainty of environment trend in a metro station. Seven-day field tests were carried out to obtain the measured data of 7 internal environmental parameters in a metro station and 8 external environment parameters. In order to ensure the prediction performance, the random forest algorithm is used to select the input variables for the proposed probabilistic forecasting method. The selected input variables and the previous predicted values are as the input variables to build the probabilistic forecasting model. The proposed method can realize to predict the probabilistic distribution of internal environmental parameters in a metro station. This work may contribute to prevent emergency events and regulate environment control system reasonably.

1. Introduction

The metro is one of the most efficient public transport modes to solve the problem of traffic congestion in urban areas [1]. However, the continuous increase of passengers brings some negative environmental problems [2, 3]. Therefore, it is necessary to analyze the environment trend in a subway station and develop a relative accurate model to predict the internal environmental parameters of the subway station [4].

In recent years, a data-based empirical modeling is a widely used alternative to mechanistic modeling since it requires less specific knowledge of the studied process [5–8]. In previous studies, the goal of environmental prediction is to obtain exact future values. Xiao-Ping et al. made the research progress of air pollution prediction based on artificial neural network [9]. Chen and Shao improved the traditional Back Propagation (BP) [10] neural network algorithm by adding momentum factor and changing learning rate. The established new model was applied to the urban air quality prediction [11]. Wang et al. used genetic algorithm to optimize the initial weights and threshold of the BP neural network in simulation [12]. Lu and Viljanen developed a network by nonlinear autoregressive with external input (NNARX) model and genetic algorithm, and it showed the suitability of neural networks to perform predictions [13]. However, the actual future results are affected by many uncertain factors, and it is very difficult to give accurate prediction values. Kamal et al. investigated the effectiveness of Artificial Neural Network (ANN) model for predicting the ambient air quality. This study illustrates that ANN can simplify and speed up the computation of the ambient air quality and provides an interesting alternative to air quality monitoring [14]. Bodri and Čermák developed an artificial time-delay feed-forward neural networks to predict Surface Air Temperatures (SAT) for six hours up to one day, and the model provided a good fit with the measured data [15].

Ramedani et al. proposed a new methodology based on ANN for generating daily GSR data [16]. Huibing put forward BP neural network prediction method to solve the problem that the environment temperature measurement accuracy is not high and it has large time delay. Simulation results show that the accuracy of temperature measurement has been significantly improved, especially on measurement delay [17]. Qu et al. developed a modeling method based on sliding time window Random Vector Functional Link Neural Network (RVFLNN) and solved the problem of slow computing speed with big data [18]. Their study improved the prediction speed while ensuring the prediction accuracy. All these research studies have good ability to model the nonlinear and dynamic system and can realize the accurately prediction.

The goals of these methods are to get exact values in every time steps. However, the real results in practice can be affected by many factors. It may be more reasonable to predict their probability distributions with different confidences rather than exact values. A good forecasting is to make predictions for an uncertain future, and its forecasting results should be shown in a form of probability distributions [19–21]. Probabilistic forecasts serve to quantify the uncertainty in the future, and they are an essential method to make an optimal decision [22]. Compared with exact prediction, probabilistic forecasts give more information. It can reveal the possible variation range of predicted parameters and determine whether the parameters exceed the maximum allowable values and its probability of occurrence. Thus, it can help an environment control system to adjust its operation for extreme and rare events [23, 24]. This will prevent some emergency accidents [25]. There are many studies of probabilistic forecasts in some fields. Aznarte investigated the convenience of quantile regression to predict extreme concentrations of NO_2, and they improved the probabilistic forecasting and allowed for the prediction of the full probability distribution, which in turn allowed to build models for the tails of this distribution [26]. Wan et al. proposed an Extreme Learning Machine (ELM) based on probabilistic forecasting method for wind power generation using the historical wind power time series as the inputs alone [27].

2. Field Test and Influence Analysis

2.1. Testing Instrument. An environmental monitor, named CPR-KA, as shown in Figure 1, is used to investigate the environmental conditions. Its pump suction rate is 300 mL/min and data sampling period is 2 minutes. This equipment uses highly sensitive electrochemical sensors to monitor the concentrations of environmental pollutants, SO_2 and NO_2, uses Photoionization Detector (PID) and infrared sensors to monitor concentrations of VOC and CO_2, respectively, uses light scattering sensors to monitor concentration of PM_{10}, and uses integrated temperature and humidity sensor to monitor temperature and RH. It can measure a variety of internal environmental parameters and pollutant concentrations. Its measurement range and accuracy are listed in Table 1.

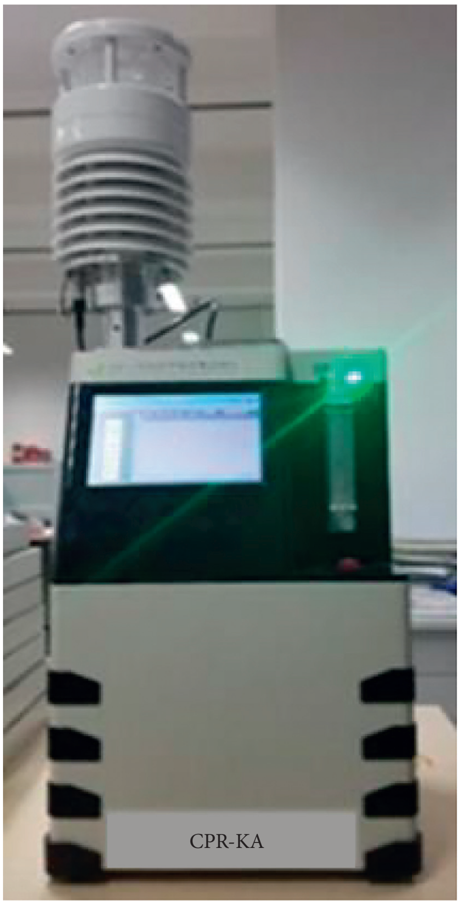


Figure 1: Real-time air quality monitor.

2.2. Measured Metro Station. The measured metro station is a transfer station with full-height platform screen doors. It is an underground station, which adopts a separated island platform design pattern. The design parameters of Heating Ventilation and Air Conditioning (HVAC) system are as follows:

(1) Rated conditions of HVAC system: the dry-bulb temperature is 28°C and the range of relative humidity is 40%–70% in the station platform for summer rated conditions

(2) Ventilation rate: the ventilation air volume in the platform is $5.78 \times 10^4\,m^3/h$ and the fresh air is $1.08 \times 10^4\,m^3/h$

The environmental monitor is located in the middle of the platform and 1.2 m above the platform ground, as shown in Figure 2. The 8 external parameters that may affect the internal environmental parameters in the metro station are also collected at the same time. The passenger flow and the arrival frequency of metro vehicle are automatically recorded. Typical external atmospheric parameters, including outdoor temperature and RH, are collected from http://data.cma.cn/. Typical outdoor air quality data, including PM_{10}, CO, NO_2, and SO_2, are obtained from http://beijingair.sinaapp.com/. During the 7-day investigations, a total number of 2800 observed environmental data are collected from the metro station.

Table 1: Measurement range and accuracy of CPR-KA.

Parameters	Measurement range	Accuracy
SO_2	0~2000 ppb	1 ppb
NO_2	0~2000 ppb	1 ppb
VOC	0~10 ppm	1 ppb
CO_2	0~5% vol	0.01% vol
PM_{10}	0~0.5 mg/m^3	0.001 mg/m^3
Temperature	−50~80°C	0.1°C
Relative humidity (RH)	0~100% RH	0.8% RH

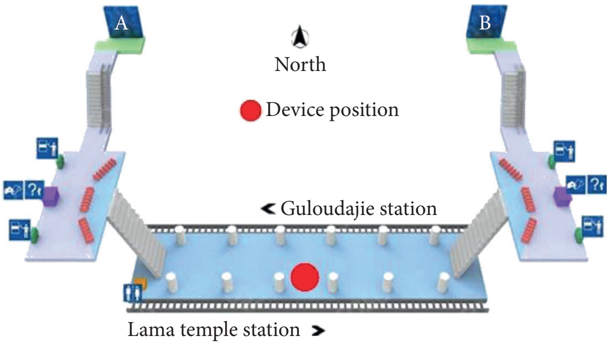


Figure 2: Measured position in the platform.

In this paper, we define some terminologies. Passenger flow, arrival frequency of metro vehicle, outdoor temperature, outdoor RH, outdoor PM10, outdoor CO, and outdoor NO_2 and SO_2 are defined as the external environmental parameters. Eight external environmental parameters are the input variables of the forecasting model. Seven parameters collected in the metro station, including CO_2, VOC, SO_2, NO_2, PM_{10}, temperature, and RH, are defined as the internal environmental parameters. They are the output variables of forecasting models.

2.3. Influence Analysis for Input Variable Selection. In order to eliminate the influence of irrelevant variables on the model performance, this paper uses the random forest algorithm to obtain the influence of external environmental variables on the predicted internal variables in training and prediction [28–30]. According to the results of influence analysis, the key variables can be selected as the input variables of the network.

Figure 3 shows the procedure of external variables' influence analysis. The random forest algorithm is used to analysis the degree of influence, W, of external environmental variables, V, on the prediction parameter, Y. The threshold value, g, is set. When $w_i > g$, the corresponding external variable will be retained into the input variables, X, of autoregressive LSTM network. Note that Y and X are time series; therefore, they can also be denoted as Y_t and X_t, respectively.

Random forest algorithm is based on the ensemble learning method [31] and uses the decision tree as a basic learner [32]. First, the influences of V on X is calculated in one decision tree; then, the average of all decision trees is calculated to get final influence W. For a specific prediction variable, Y, the above process is as follows [33]:

(1) Denote original dataset as $D = [V \circ Y]$, where $[\cdot\circ\cdot]$ represents concatenation. Extract K bootstrap datasets [34], $\{D^k\}_{k=1}^{K}$, from D, and in the meanwhile remains K Out-Of-Bag (OOB) datasets, $\left\{\widetilde{D}^k\right\}_{k=1}^{K}$.

(2) Initialization, $k = 1$.

(3) Train the kth decision tree regression model C^k with dataset D^k, and calculate its prediction accuracy, E^k, using the corresponding OOB dataset, $\widetilde{D}^k$.

(4) Add noise to external parameter, v_i, in the OOB dataset, and calculate the prediction accuracy of model C^k again, and the changed accuracy after adding noise is denoted as $\widetilde{E}_i^k$.

(5) Repeat step (3)~step (4) until $k = K$.

When all the decision trees are processed using the above steps, the degree of influences can be calculated with equation (1) for a given external variable, v_i:

$$w_i = \frac{1}{K}\sum_{k=1}^{K}\left(E^k - \widehat{E}_i^k\right), \tag{1}$$

where E^k is the prediction accuracy without any disturbance to the external parameters when training the kth decision tree; $\widehat{E}_i^k$ is the prediction accuracy after adding noise to external variable; and K is the number of decision trees.

The degree of influences, W, of V on Y can be obtained using the above steps. The larger the w_i, the greater the variable's contribution to the predictor. Therefore, we extract input variables by setting a threshold value, g. The external variable will be retained as an input variable when $w_i > g$. Denote these input variables as $X = \{x_i\}_{i=1}^{Z}$, as shown in Figure 3. Because g is a learnable parameter, we determine the optimal g after comparing different values.

3. Principle of Probabilistic Forecasting Method

3.1. Procedure of Probabilistic Forecasting Method. The probabilistic forecasting method proposed in this paper can obtain the Gaussian distributions of the predicted environmental parameters in the future time points based on past observations. The overall procedure is summarized in Figure 4 and it contains the following four steps:

Step 1: environmental data preprocessing.

The internal environmental parameters in station platform were measured every 2 minutes for 7 days starting from 21 October 2019. The Butterworth low-pass filter algorithm is used to deal with the raw data [35, 36]. The useful signal and noise are separated and the high-frequency interference signals are filtered out [37]. The transfer function of Butterworth low-pass filter is given by

$$|H(\omega)|^2 = \frac{1}{1 + \left(\omega/\omega_c\right)^{2N}}, \tag{2}$$

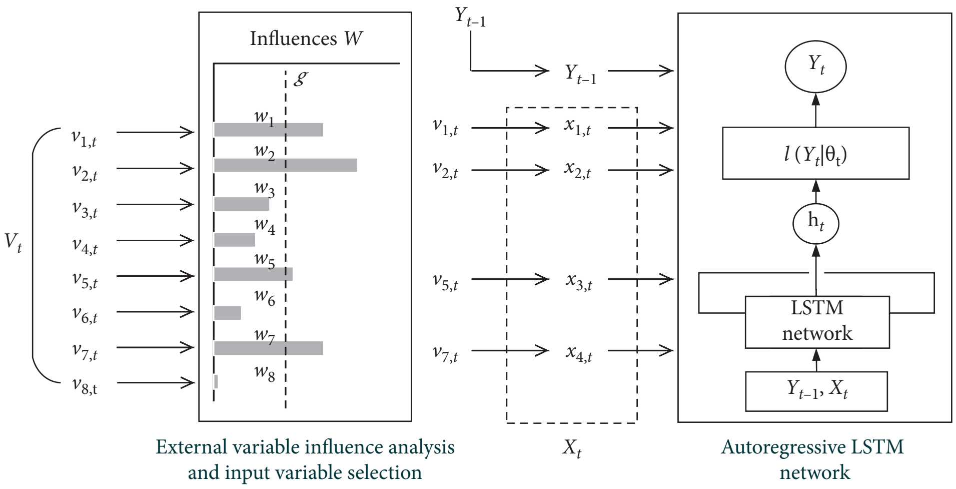


Figure 3: External variable influence analysis and input variable selection.

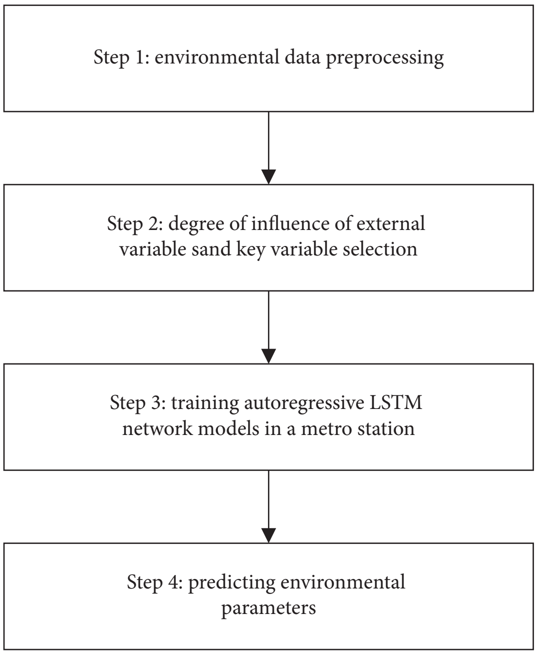

Step 1: environmental data preprocessing
Step 2: degree of influence of external variable sand key variable selection
Step 3: training autoregressive LSTM network models in a metro station
Step 4: predicting environmental parameters

Figure 4: Procedure of probabilistic forecasting method of environmental parameters.

where N is the order of filter; w is the frequency, rad/s; and w_c is cut-off frequency.

The frequency-amplitude curve of a filter includes passband and stopband and transition band. For passband and stopband, its two rules are shown in equations (3) and (4), respectively:

$$1 - \delta_p \leq |H(\omega)| \leq 1 + \delta_p, \quad |\omega| \leq \omega_p, \tag{3}$$

$$|H(\omega)| \leq \delta_s, \quad \omega_s \leq |\omega| \leq \infty, \tag{4}$$

where w_p and w_s are edge frequencies of passband and stopband, respectively; δ_s is the deviation of amplitude between filter and ideal filter in stopband; and δ_p is the deviation of amplitude between the filter and ideal filter in passband.

Step 2: degree of influence of external variable and input variable selection.

Although eight types of external variables are measured, they have different influences on the internal environmental parameters in the metro platform.

If all the external variables are taken as the input variables of the forecasting model, it might lead to worse prediction results. The presented probabilistic forecasting method based on the autoregressive LSTM neural network is a machine learning algorithm. Its input variables have a great impact on the performance of machine learning algorithm [38–40]. In general, the collected data are not entirely suitable as input variables of neural network. It is significant to reduce the number of input parameters in order to avoid overfitting and accelerate the training speed of the model. It has been used in some research studies and has obtained good results, and these research results showed the important role of input variable selection [41, 42].

Seven internal environmental variables are collected in the metro station, including CO_2, CO, CH_2O, VOC, SO_2, NH_3, NO_2, PM_{10}, temperature, and RH. In this paper, the probabilistic forecasting method will predict these parameters. The predicted parameters are denoted as Y. Correspondingly, external environmental variables are denoted as $V = \{v_i\}_{i=1}^{M}$, where $M = 8$. The influence of external variables V on prediction parameter Y will be analyzed using the random forest algorithm. Denoting the degree of influence of external environmental parameters by $W = \{w_i\}_{i=1}^{M}$.

The greater the w_i of the external environmental parameter, the greater its influence on the internal environmental parameter, and vice versa. We denote a manually threshold value, g, to select external environmental parameters. Therefore, variables whose weights are less than g are eliminated in order to exclude their negative influences on model performance.

Step 3: training autoregressive LSTM network models.

After Step 2, the external environmental parameters with high degree of influence are selected as the inputs of the autoregressive LSTM network together with historical data at the previous time step [43]. Denote the input variables for each prediction parameter by $X = \{x_i\}_{i=1}^{Z}$, where Z is the number of input variables. The prediction parameters are the internal environmental parameters in the metro station. Because different prediction parameters have different input variables, it is necessary to adjust the structure of LSTM network in order to better predict the changes of different environment parameters. The number of input layer nodes in the network structure is Z, and it is equal to the number of input variables. The training dataset and the test dataset are divided by 7 : 3; the first 70 percent of observations are used to develop the prediction models and the remaining 30 percent of observations are used as a test dataset.

Step 4: predicting the environmental parameters in a metro station.

In this prediction process, we use the same network structures and parameters in the training process. However, in this process, there is a slight difference from the training process. The prediction variables are known in the training process, but they are unknown in the prediction process. In order to continue the prediction process, we use the rolling window prediction that can feed the last outputs back as the input until the end of the prediction range.

3.2. Model of Autoregressive LSTM Network. For internal environmental parameter prediction, it is important to build a conditional distribution. Thus, the proposed model can be denoted as

$$P\left(Y_{t_0+1:t_0+\tau}\,\middle|\,Y_{1:t_0}, X_{1:t_0+\tau}; \Phi\right), \tag{5}$$

where t_0 is the time point which splits the past and the future; τ is the length of prediction range; $Y_{t_0+1:t_0+\tau}$ and $Y_{1:t_0}$ are the target values in time range $[t_0 + 1: t_0 + \tau]$ and $[1: t_0]$, respectively; $X_{1:t_0+\tau}$ is the value of external variable in time range $[1: t_0 + \tau]$; and Φ denotes the parameters of the model.

In equation (5), the whole time series, $[1: t_0 + \tau]$, is split into two parts, $[1: t_0]$ and $[t_0 + 1: t_0 + \tau]$, by time point t_0. The first half part named the condition range contains the past information, and the remaining part is called the prediction range. The model utilizes the past values of prediction variable, $Y_{1:\,t_0}$, and the external variables, $X_{1:\,t_0+\tau}$, to predict the future values, $Y_{t_0+1:t_0+\tau}$. $Y_{t_0+1:t_0+\tau}$ is assumed to be unknown at prediction time, and $X_{1:t_0+\tau}$ is known external variables.

For each time point, the problem can be parametrized by the output h_t of an autoregressive LSTM network:

$$\begin{aligned}P\left(Y_{t_0+1:t_0+\tau}\,\middle|\,Y_{1:t_0}, X_{1:t_0+\tau}; \Phi\right) &= \prod_{t=t_0+1}^{t_0+\tau} P\left(Y_t\,\middle|\,Y_{t-1}, X_t; \Phi\right)\\ &= \prod_{t=t_0+1}^{t_0+\tau} \ell\left(Y_t\,\middle|\,\theta\left(h_t, \Phi\right)\right),\\ h_t &= h\left(h_{t-1}, Y_{t-1}, X_t, \Phi\right),\end{aligned} \tag{6}$$

where h is a function implemented by LSTM cells; Y_t is internal environmental parameter Y at time t; $\ell(\cdot)$ is the likelihood to fit the distribution of predictive variables; and $\theta(\cdot)$ is a function that computes the parameters of the likelihood.

The autoregressive model means that the observation at last time step, Y_{t-1}, and the previous output of the network, h_{t-1}, are fed back as inputs for the next time step. The likelihood, $\ell(Y_t \mid \theta(h_t, \Phi))$, is a fixed distribution whose parameters are given by a function $\theta(h_t, \Phi)$ of the network output h_t.

It is significant to choose a good distribution for the proposed model. The environmental parameters are assumed following the Gaussian distribution according to the research results by some researchers [44, 45]. It is greatly convenient to construct the LSTM network because the Gaussian distribution has mean and variance. Thus, for the study in this paper, the distribution of Y is determined as Gaussian distribution, so the likelihood can be denoted by equation (7), and its parameters, the mean μ and standard deviation σ, are given by equations (8) and (9). The mean is given by an activation function of the network output, and the standard deviation is obtained by applying an activation function followed by a softplus activation function:

$$\ell_G(Y \mid \theta(h, \Phi)) = \ell_G(Y \mid \mu, \sigma) = \left(2\pi\sigma^2\right)^{-(1/2)} \exp\left(-\frac{(Y-\mu)^2}{(2\sigma^2)}\right), \tag{7}$$

$$\mu\left(h_t\right) = w_\mu^T h_t + b_\mu, \tag{8}$$

$$\sigma\left(h_t\right) = \log\left(1 + \exp\left(w_\sigma^T h_t + b_\sigma\right)\right), \tag{9}$$

where μ and σ are the mean and standard deviation, respectively; h_t is the network output; and w and b are weights and bias of nonlinear transformation, respectively.

For the training and forecasting processes, their network structures are the same, but there is a slightly difference to calculate Y, as shown in Figure 5. For the training process, the values of Y are assumed to be known, but they are unknown in prediction process. The value of Y at the last time step can be the input of model. In order to continue the prediction, a sampled value should be obtained from the distribution of the last time step. They will be described and discussed in Sections 3.3 and 3.4, separately.

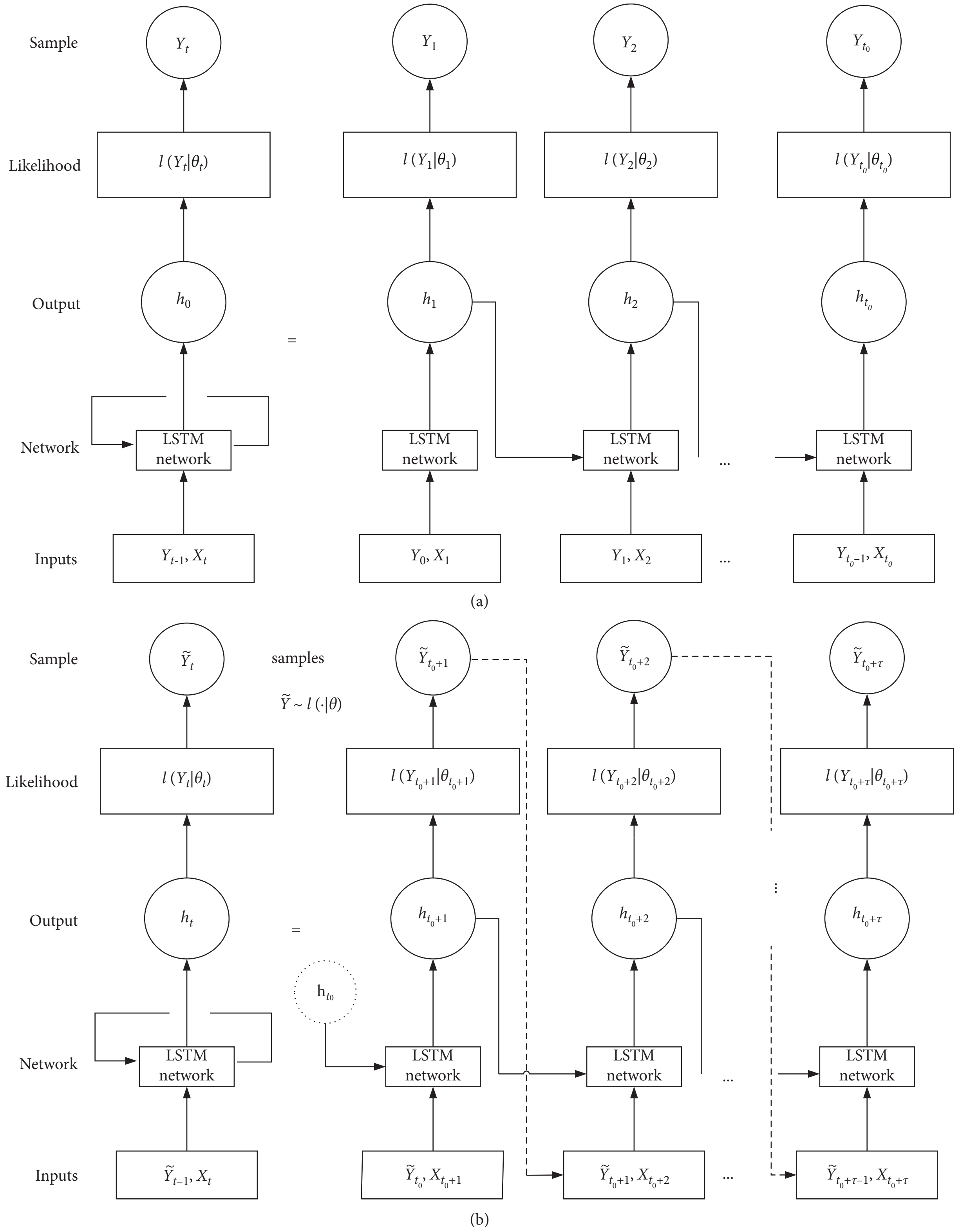


FIGURE 5: Models of autoregressive LSTM network. (a) Training and (b) prediction.

3.3. Training Process. Figure 5(a) illustrates the training process. The inputs of network are X_t and Y_{t-1}. All time steps are in a conditional range, $[1{:}\ t_0]$. The autoregressive LSTM network in Figure 5(a) can be expanded some continuous training process according to the time steps. Their inputs are (Y_{t-1}, X_t) and the previous network output h_{t-1} at each time step t. Here, $t \in [1{:}\ t_0]$. The network output, $h_t = h(h_{t-1}, Y_{t-1}, X_t, \Phi)$, is then used to compute the parameters of the likelihood, $\theta_t = \theta(h_t, \Theta)$, using equations (8) and (9). Finally, the model parameters are optimized using

$$L = \sum_{t=t_0+1}^{t_0+\tau} \log \ell\left(Y_t \mid \theta(h_t\right), \tag{10}$$

where h_t is the output of the network and Y_t is the actual value of prediction variable.

In equation (10), our goal is to maximize the probability of Y_t in the predicted Gaussian distribution. μ and σ can be optimized directly via stochastic gradient descent by computing gradients.

3.4. Prediction Process. Figure 5(b) illustrates the prediction process. The network structure and the parameters in the training process are same as the prediction process. However, the inputs of prediction network are different from training network, and the actual values of prediction variable are unknown in the time range of $[t_0 + 1{:}\ t_0 + \tau]$. Therefore, $\widetilde{Y}_{t_0+1:\ t_0+\tau} \sim P_\Phi(Y_{t_0+1:\ t_0+\tau} \mid Y_{1:\ t_0}, X_{1:\ t_0+\tau})$ can be obtained from the prediction distribution and used as the one of the input values for the next time steps, and the inputs of the prediction network are $(\widetilde{Y}_{t-1}, X_t)$, h_t and $\widetilde{Y}_t$ at each time step t. Here, $t \in [t_0 + 1{:}\ t_0 + \tau]$.

By rolling window prediction, the distributions at all prediction time steps could be given. The whole prediction is as follows. First, h_{t_0} is obtained from the end of training process. Then, h_{t_0+1} is calculated with the inputs of X_{t_0+1}, Y_{t_0}, and h_{t_0}. After getting the network output, h_{t_0+1}, the Gaussian likelihood, $\ell(Y_{t_0+1} \mid \theta_{t_0+1})$, can be built. Finally, $\widetilde{Y}_{t_0+1} \sim \ell(Y_{t_0+1} \mid \theta_{t_0+1})$ is drawn and fed back for the next point $t_0 + 2$. This prediction process is repeated until $[t_0 + 1{:}\ t_0 + \tau]$.

4. Results and Discussion

4.1. Results of Data Preprocessing. The collected environmental data are processed by these steps: removal and replacement of outliers, missing data imputation, and noise smoothing. Figure 6 shows a part of data after preprocessing.

4.2. Time Series Processing for LSTM Network. Probability forecasting method is based on autoregressive LSTM network, so the input variables need to be processed, as shown in Figure 7. Set $t_0 = 8$ as the boundary of $[1{:}\ t_0]$ and $[t_0 + 1{:}\ t_0 + \tau]$. Whole dataset can be built through selecting different start points from the entire time series.

4.3. Determination of Parameter g. For the prediction of Y, 8 external environmental parameters are input to get their degrees of influence, $W = [w_1, w_2, \ldots, w_M]$. Figure 8 shows the results of the influence of external environmental parameters on the internal environmental parameters.

From Figure 8, the results can be observed:

(1) Passenger volume is the main influence factor for carbon dioxide concentration and temperature in the metro station

(2) RH, PM_{10}, and NO_2 in the outside atmosphere have obvious contributions for RH, PM_{10}, and NO_2 in the metro station, respectively

For the proposed method, an external environmental parameter will be retained as an input variable when its influence degree, w_i, is larger than g. So, g is regarded as a hyperparameter to determine the input variables of the network. In general, a grid search and a manual search are the most widely used strategies for hyperparameter optimization [46–48]. Because most of the influence degrees are less than 0.3 according to Figure 8, all available values of g are set as 0.1, 0.2, and 0.3, respectively. The grid search method is used to select the optimal value of g. Each value of g is applied to the model to calculate its RMSE. The value of g with smallest RMSE is selected as the final g. The RMSE is calculated with

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_t \left(Y_t - \widetilde{Y}_t\right)^2}, \tag{11}$$

where T is the length of prediction values and Y_t and $\widetilde{Y}_t$ are predicted and actual data, respectively.

Table 2 shows the results of different values of g. There are 4 kinds of g values. We use every value of them to select the input variables and train different models. The RMSE of these models are shown in Table 2. The value of g corresponding to the minimum RMSE value is considered to be an optimal one, which is a bold font in Table 2.

Finally, every internal environmental parameter has optimal g value according to Figure 8 and Table 2. For example, the minimum RMSE of CO_2 is 0.227, and its corresponding g value is 0.1. The influences exceeding 0.1 are PF and NO_2. Thus, the two external parameters, PF and NO_2, are used to predict CO_2 concentration in a metro station.

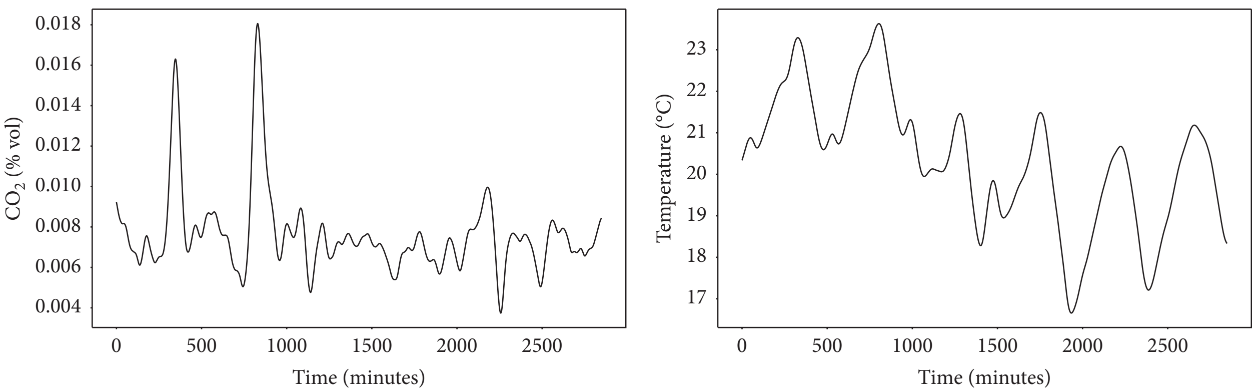
From Table 2 and Figure 8, we can draw the following conclusions:

(1) Different internal environmental parameters have different optimum g values.

(2) The smaller the RMSE value, the better the performance of the model, so the models of CO_2, VOC, PM_{10}, and RH have the best performance with g = 0.1 For these parameters, g is determined as 0.1.

(3) For SO_2, NO_2, and TEM, g is determined as 0.2.

4.4. Results of Environmental Prediction. The collected 2800 data from the metro station are adopted for predictions. We built different prediction models to predict the corresponding internal environmental parameters. The proposed model based in LSTM network has input layers, hidden layers, and output layers. Because the input variables determine its input layer nodes, different models have different input layer nodes. The number of input layers is decided by its degree of influence. Each model has 2 hidden layers with



(a)

(b)

Figure 6: Data preprocessing of (a) CO_2 and (b) Temperature.

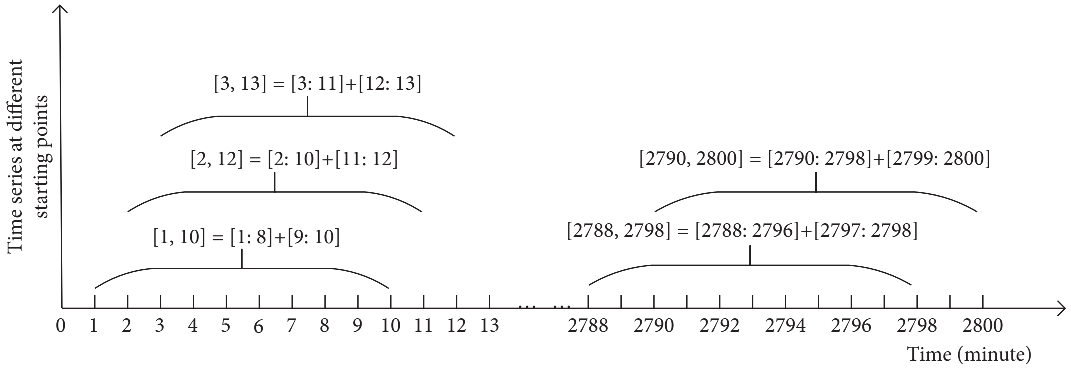
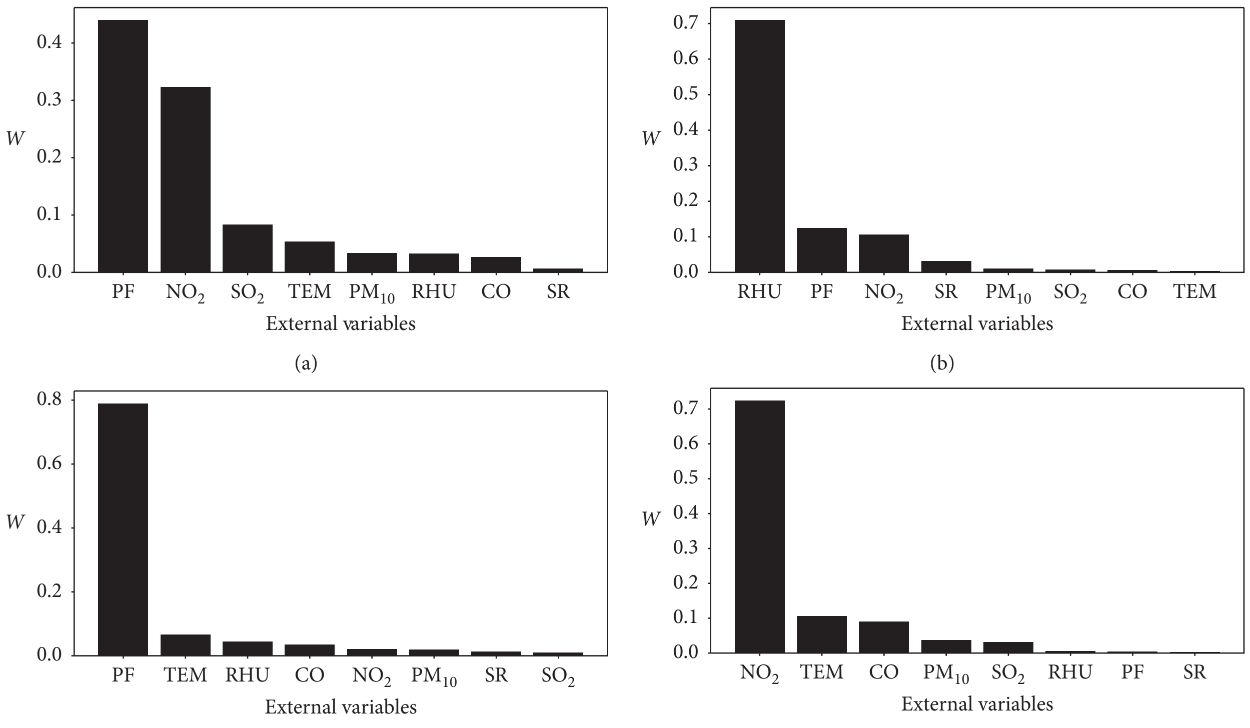


Figure 7: Time series slicing.



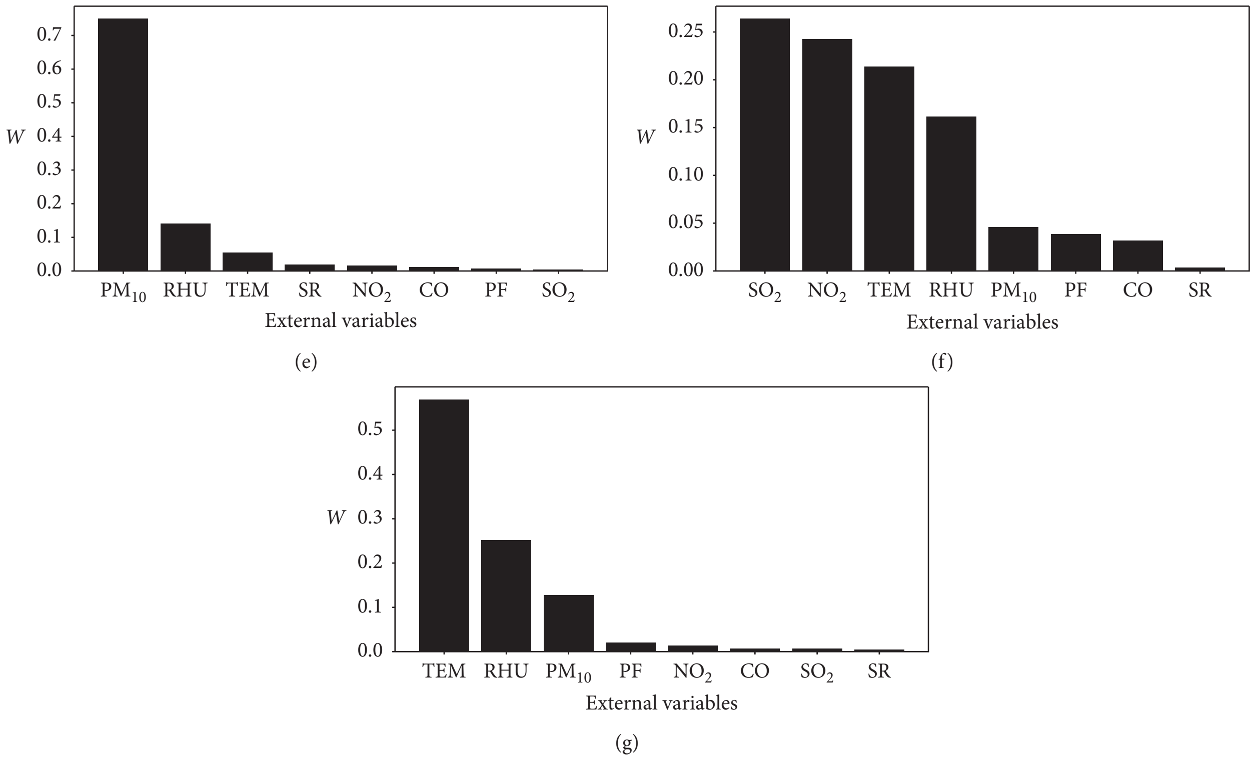
(a)

(b)

(c)

(d)

Figure 8: Continued.



(e)

(f)

(g)

Figure 8: Influence analysis of external environmental parameters. (a) CO_2, (b) RH, (c) temperature, (d) NO_2, (e) PM_{10}, (f) SO_2, and (g) VOC.

Table 2: RMSE of the prediction model with different g.

	$g = 0$	$g = 0.1$	$g = 0.2$	$g = 0.3$
CO_2	0.405	**0.227**	0.481	0.263
VOC	0.137	**0.064**	0.073	0.069
SO_2	0.236	0.303	**0.164**	0.203
NO_2	0.179	0.165	**0.088**	0.196
PM_{10}	0.419	**0.154**	0.209	0.237
TEM	0.401	0.168	**0.115**	0.212
RH	0.215	**0.150**	0.198	0.156

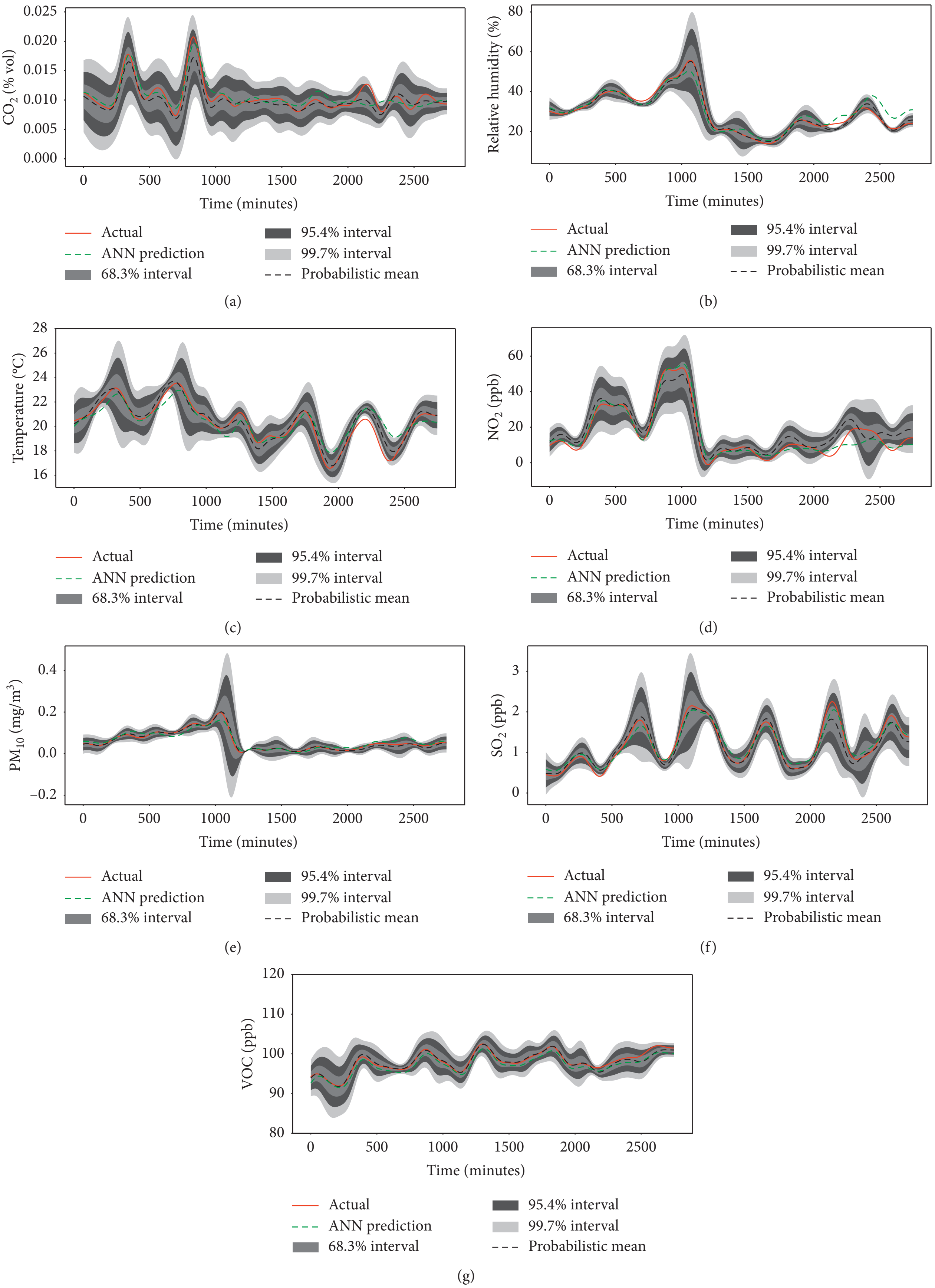
nodes number of 64 and 16, respectively. It has 1 output layer with node number of 2, which represent the mean and variance of normal distribution. The training and test datasets are divided according to a ratio of 7 : 3. The time step is set to 120 s, and the number of training iterations is set to 1000.

4.4.1. Results Based on the Three-Sigma Rule of Distribution. The evaluation of the proposed model is based on the three-sigma rule of distribution. Three-sigma rule is an empirical rule stating that, for many reasonably symmetric unimodal distributions, almost all of the population lies within three standard deviations of the mean [49]. Therefore, we define three ranges of predicted normal distribution in different intervals, $[\mu - \sigma, \mu + \sigma]$, $[\mu - 2\sigma, \mu + 2\sigma]$, and $[\mu - 3\sigma, \mu + 3\sigma]$. For a normal distribution, 68.3% of the observations are within $[\mu - \sigma, \mu + \sigma]$, 95.4% are within $[\mu - 2\sigma, \mu + 2\sigma]$, and 99.7% are within $[\mu - 3\sigma, \mu + 3\sigma]$. Figure 9 shows that the prediction results of probabilistic forecasting method in three ranges. Table 3 shows that the propagation of the actual values falling in different ranges. The traditional ANN results are also calculated and shown in Figure 9, in which "Actual," "ANN prediction," and "Probabilistic mean" represent the measured data, the results of ANN prediction, and the results of the proposed method, respectively.

As shown in Table 3, we calculate the proportion of the actual values falling in three-sigma limits for seven internal environmental parameters in metro station. The higher the proportion, the more accurate the probabilistic distribution and the higher the prediction accuracy. The model accuracy can be verified by analyzing and comparing the proportion. In the range of $[\mu + \sigma, \mu - \sigma]$, the proportion will be 68.3% if it obeys the standard normal distribution. So, our goal is to make the propagation of predicted results in this range exceed 68.3%. Similarly, in the ranges of $[\mu + 2\sigma, \mu - 2\sigma]$ and $[\mu + 3\sigma, \mu - 3\sigma]$, our goals are 95.4% and 99.7%, respectively. The mean proportion of these ranges is calculated and listed in the last row in Table 3. Some conclusions can be obtained from Table 3:

(1) The results show that the mean proportion in three intervals is 76.75%, 93.00%, and 98.12%, respectively. This means that the normal distribution predicted by our model can effectively cover the change range of predicted variables and give different probability intervals according to three-sigma limits.

(2) The proportions of CO_2, VOC, SO_2, PM_{10}, and TEM in the range of $[\mu + \sigma, \mu - \sigma]$ are greater than 68.3%.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

FIGURE 9: Predicting result comparisons. (a) CO_2, (b) RH, (c) temperature, (d) NO_2, (e) PM_{10}, (f) SO_2, and (g) VOC.

Table 3: Proportion of actual values falling in three-sigma limits.

	CO_2 (%)	VOC (%)	SO_2 (%)	NO_2 (%)	PM_{10} (%)	TEM (%)	RH (%)	Mean (%)
$\mu \pm \sigma$	86.07	91.20	71.67	59.75	96.55	70.87	61.13	76.75
$\mu \pm 2\sigma$	96.44	100.00	96.80	85.53	100.00	88.44	83.82	93.00
$\mu \pm 3\sigma$	98.55	100.00	99.38	92.29	100.00	96.62	100.00	98.12

Table 4: Propagation of ANN results located in three-sigma limits.

	CO_2 (%)	VOC (%)	SO_2 (%)	NO_2 (%)	PM_{10} (%)	TEM (%)	RH (%)	Mean
$\mu \pm \sigma$	71.93	82.00	85.53	63.96	65.93	51.42	63.16	69.13
$\mu \pm 2\sigma$	93.42	98.07	96.69	85.38	89.09	77.75	93.86	90.61
$\mu \pm 3\sigma$	98.29	100.00	100.00	93.16	96.66	91.24	96.15	96.50

Table 5: RMSEs and PIs of ANN and probabilistic forecasting methods.

	CO_2	VOC	SO_2	NO_2	PM_{10}	T	RH
RMSE of ANN method	0.57	0.08	0.35	0.33	0.43	0.48	0.30
RMSE of proposed method	0.22	0.06	0.16	0.08	0.15	0.12	0.15
PI	0.60	0.28	0.53	0.73	0.64	0.76	0.50

On the contrary, the proportions of NO_2 and RH are lower than 68.3%. In the range of $[\mu + 2\sigma, \mu - 2\sigma]$, the proportions of CO_2, VOC, SO_2, and PM_{10} are greater than 95.4%, and they are higher than ones of NO_2, TEM, and RH. Finally, in the range of $[\mu + 3\sigma, \mu - 3\sigma]$, only the proportions of VOC and PM_{10} are greater than 99.7% and the proportions of other variables are lower than 99.7%. In particular, the proportions of NO_2 and RH are smaller than other variables in the range of $[\mu + 2\sigma, \mu - 2\sigma]$ and $[\mu + 3\sigma, \mu - 3\sigma]$. Therefore, we draw a conclusion that the prediction accuracy values of NO_2 and RH are lower than other variables, but those of VOC and PM_{10} are higher than other variables.

(3) Most of the ANN prediction results fall into these three intervals, as shown in Figure 9. It means that most of the ANN prediction results are included in the range of normal distribution predicted by the probabilistic forecasting model.

4.4.2. Result Comparison with the ANN Model. In order to further reveal the prediction performance of the presented probability forecasting method, it is compared with the traditional ANN. As shown in Table 4, we calculated the proportions of the ANN results falling in three-sigma limits for seven internal environmental parameters in metro station.

Some conclusions can be obtained from Table 4:

(1) There are 96.50% of ANN prediction results falling in the range $[\mu + 3\sigma, \mu - 3\sigma]$ on average, which means that the results of probabilistic prediction include the most of ANN prediction results. The results of probabilistic prediction can replace ANN prediction results to some extent.

(2) There are 69.13%, 90.61%, and 96.50% of ANN results falling in three-sigma limits on average, respectively. The average proportions of actual values falling in three-sigma limits are 76.75%, 93.00%, and 98.12%, respectively. They are higher than the ones of ANN prediction.

The outputs of the probabilistic forecasting model are a series of normal distributions, and the outputs of the ANN model are a series of exact values. We use the mean, μ, of the probabilistic forecasting model to calculate RMSE and compare it with the ANN model. In addition, the improvement performance of the probabilistic forecasting method compared with the ANN method is calculated with equation (12) and listed in Table 5:

$$\mathrm{PI} = \frac{Y_{\mathrm{ANN}} - Y_{\mathrm{prob}}}{Y_{\mathrm{ANN}}}, \quad (12)$$

where PI is the relative percentage of performance improvement and Y_{ANN} and Y_{prob} are the prediction result RMSE of traditional ANN and the proposed methods, respectively.

The results show that the RMSE values of the presented model are lower than the ones of ANN. The mean of all PIs is 0.58, which means 58% improvement over ANN on average. These data illustrate that if only using the mean value to evaluate, the probabilistic forecasting model is closer to the actual values, and its accuracy is better than that of the ANN.

5. Conclusion

A probabilistic forecasting method for the internal environment in metro station is proposed on the basis of the autoregressive LSTM network. This method can predict the probabilistic distribution of internal environmental parameters in a metro station. The 2800 observations from the measured metro station are used to illustrate the proposed model and its performance is well compared with ANN. Some results can be obtained from our study:

(1) The random forest algorithm is used to analyze the degree of influence of external environmental parameters on the predicted variables. For CO_2 and temperature, the results show that the passenger flow is the most important influence parameter. Other

internal environmental parameters, such as RH, PM_{10}, SO_2, and NO_2, are mainly influenced by their corresponding parameters in the outside atmosphere, respectively.

(2) The proposed model can build a conditional distribution between past data and future data, and its results are a series of distributions of mean and standard deviation. The proportions of the actual values falling in three-sigma limits are 76.75%, 93.00%, and 98.12%, respectively, which shows the reliability of the proposed model.

(3) Compared with the ANN model, the proposed model can predict the internal environmental parameters in the metro station platform. The results show that there are 69.13%, 90.61%, and 96.50% of ANN prediction results falling in three-sigma limits on average, respectively, which is lower than the proportions of probabilistic forecasting. In addition, the proposed model has 58% improvement over the ANN on average if using its mean of predicted distribution to compare.

The above results show that probabilistic forecasting model is more suitable for predicting internal environmental parameter in a metro station than the ANN. The most important contribution of the proposed method is that it can provide extreme prediction values in future time, such as the upper and lower boundaries with corresponding probability. Therefore, it can support the decision-making and give more information to adjust the operation of HVAC system in a metro station.

There are many factors affecting the internal environmental parameters in the metro station. In this paper, only 8 external variables are selected as the input variables of the prediction, and this may miss some useful variables, such as outdoor weather. This may reduce the accuracy of the prediction to a certain extent. In the future research, more factors may be considered to increase the reliability of the model. Multiple metro stations will be considered to collect their experimental data in the future in order to improve the performance of the model.

Data Availability

The author have no right to share data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors gratefully acknowledge the financial support of National Key R&D Program of China (no. 2017YFB1201100) and Liao Ning Revitalization Talents Program (XLYC1802092). Thanks to researchers who have given me selfless help on field testing in metro station.

References

[1] V. Martins, T. Moreno, M. C. Minguillón et al., “Exposure to airborne particulate matter in the subway system,” *Science of the Total Environment*, vol. 511, pp. 711–722, 2015.

[2] M. Kim, B. Sankararao, O. Kang, J. Kim, and C. Yoo, “Monitoring and prediction of indoor air quality (IAQ) in subway or metro systems using season dependent models,” *Energy and Buildings*, vol. 46, pp. 48–55, 2012.

[3] H. Liu, S. Lee, M. Kim et al., “Multi-objective optimization of indoor air quality control and energy consumption minimization in a subway ventilation system,” *Energy and Buildings*, vol. 66, pp. 553–561, 2013.

[4] M. Kim, R. D. Braatz, J. T. Kim, and C. Yoo, “Indoor air quality control for improving passenger health in subway platforms using an outdoor air quality dependent ventilation system,” *Building and Environment*, vol. 92, pp. 407–417, 2015.

[5] H. Liu and J. Wang, “Integrating independent component analysis and principal component analysis with neural network to predict Chinese stock market,” *Mathematical Problems in Engineering*, vol. 2011, Article ID 382659, 15 pages, 2011.

[6] F. Leon and M. H. Zaharia, “Stacked heterogeneous neural networks for time series forecasting,” *Mathematical Problems in Engineering*, vol. 2010, Article ID 373648, 20 pages, 2010.

[7] T. Tunç, “A new hybrid method logistic regression and feedforward neural network for lung cancer data,” *Mathematical Problems in Engineering*, vol. 2012, Article ID 241690, 10 pages, 2012.

[8] Y. Bo, Y. Cui, L. Zhang et al., “Beam structure damage identification based on BP neural network and support vector machine,” *Mathematical Problems in Engineering*, vol. 2014, Article ID 850141, 8 pages, 2014.

[9] B. Xiao-Ping, L. I. Hong, Z. Qi-Ming et al., “Progress of research on artificial neural network in air pollution prediction,” *Science & Technology Review*, vol. 24, no. 12, pp. 77–81, 2006.

[10] Q. Chen and Y. Shao, “The application of improved BP neural network algorithm in urban air quality prediction: evidence from China,” in *Proceedings of the 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, December 2008.

[11] M. Buscema, “Back propagation neural networks,” *Substance Use & Misuse*, vol. 33, no. 2, pp. 233–270, 1998.

[12] F. Wang, S. Y. Cheng, M. J. Li et al., “Optimizing BP networks by means of genetic algorithms in air pollution prediction,” *Journal of Beijing University of Technology*, vol. 35, no. 9, pp. 1230–1234, 2009.

[13] T. Lu and M. Viljanen, “Prediction of indoor temperature and relative humidity using neural network models: model comparison,” *Neural Computing and Applications*, vol. 18, no. 4, pp. 345–357, 2009.

[14] M. M. Kamal, R. Jailani, and R. L. A. Shauri, “Prediction of ambient air quality based on neural network technique,” in *Proceedings of the 2006 4th Student Conference on Research and Development*, Selangor, Malaysia, June 2006.

[15] L. Bodri and V. Čermák, “Prediction of surface air temperatures by neural network, example based on three-year temperature monitoring at spořilov station,” *Studia Geophysica et Geodaetica*, vol. 47, no. 1, pp. 173–184, 2003.

[16] Z. Ramedani, M. Omid, and A. Keyhani, “A method based on neural networks for generating solar radiation map,”

International Journal of Energy, Environment and Economics, vol. 3, no. 5, pp. 775–786, 2012.

[17] L. Huibing, "BP neural network prediction method based on temperature," *Electronic Test*, vol. 2013, no. 19, pp. 62–64, 2013.

[18] H. Qu, S. Fu, L. Pang, C. Ding, and H. Zhang, "Rapid temperature prediction method for electronic equipment cabin," *Applied Thermal Engineering*, vol. 138, pp. 83–93, 2018.

[19] V. Flunkert, D. Salinas, and J. Gasthaus, "Deepar: probabilistic forecasting with autoregressive recurrent networks," 2017, https://arxiv.org/abs/1704.04110.

[20] H. V. Roberts, "Probabilistic prediction," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 50–62, 1965.

[21] P. Friederichs and A. Hense, "A probabilistic forecast approach for daily precipitation totals," *Weather and Forecasting*, vol. 23, no. 4, pp. 659–673, 2008.

[22] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 243–268, 2007.

[23] R. Buizza, "The value of probabilistic prediction," *Atmospheric Science Letters*, vol. 9, no. 2, pp. 36–42, 2008.

[24] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 125–151, 2014.

[25] P. C. Miclea, "Metro ventilation," *Tunnels & Tunnelling International*, vol. 16, no. 12, pp. 51–54, 2011.

[26] J. L. Aznarte, "Probabilistic forecasting for extreme NO_2 pollution episodes," *Environmental Pollution*, vol. 229, pp. 321–328, 2017.

[27] C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong, "Probabilistic forecasting of wind power generation using extreme learning machine," *IEEE Transactions on Power Systems*, vol. 29, no. 3, pp. 1033–1044, 2014.

[28] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[29] H. Pang, A. Lin, M. Holford et al., "Pathway analysis using random forests classification and regression," *Bioinformatics*, vol. 22, no. 16, pp. 2028–2036, 2006.

[30] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[31] T. Hastie, R. Tibshirani, and J. Friedman, "Ensemble learning," in *The Elements of Statistical Learning*, Springer, New York, NY, USA, 1970.

[32] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems Man & Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[33] L. I. Xian, W. Yan, L. Yong et al., "Segmentation of nasopharyngeal neoplasms based on random forest feature selection algorithm," *Journal of Computer Applications*, vol. 2019, no. 5, pp. 1485–1489, 2019.

[34] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[35] L. Kai-Si, Z. Lu, and L. Kai-Wu, "Research on optimization design and simulation of butterworth low-pass filter," *Journal of Chongqing Technology and Business University (Natural Science Edition)*, vol. 2014, no. 6, pp. 58–62, 2014.

[36] W. Yu, Z. Ming, L. I. Zong et al., "Normalized design and application of Butterworth low-pass filter," *Ship Electronic Engineering*, vol. 2018, no. 1, pp. 61–64, 2018.

[37] J. Roberts and T. D. Roberts, "Use of the butterworth low-pass filter for oceanographic data," *Journal of Geophysical Research*, vol. 83, no. C11, p. 5510, 1978.

[38] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, NY, USA, 2003.

[39] Md Raihan-Al-Masud and M. M. R. Hossain, "Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms," *PLoS One*, vol. 15, no. 2, 2020.

[40] S.-Y. Jiang and L.-H. Wang, "Enhanced machine learning feature selection algorithm for cardiac arrhythmia in a personal healthcare application," in *Proceedings of the 2018 IEEE Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia)*, October 2018.

[41] M. H. Kim, Y. S. Kim, J. Lim, J. T. Kim, S. W. Sung, and C. Yoo, "Data-driven prediction model of indoor air quality in an underground space," *Korean Journal of Chemical Engineering*, vol. 27, no. 6, pp. 1675–1680, 2010.

[42] J. Lim, Y. Kim, T. Oh et al., "Analysis and prediction of indoor air pollutants in a subway station using a new key variable selection method," *Korean Journal of Chemical Engineering*, vol. 29, no. 8, pp. 994–1003, 2012.

[43] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*, Oklahoma State University, Stillwater, OK, USA, 2002.

[44] X. Fernandez and D. Caroselli, "Environmental distribution [DB/OL]," 2014, http://citeseerx.ist.psu.edu/viewdoc/summary?.

[45] A. G. Lynch, "Normal distribution," in *Encyclopaedic Companion to Medical Statistics*, Wiley, Hoboken, NJ, USA, 2011.

[46] F. J. Pontes, G. F. Amorim, P. P. Balestrassi, A. P. Paiva, and J. R. Ferreira, "Design of experiments and focused grid search for neural network parameter optimization," *Neurocomputing*, vol. 186, pp. 22–34, 2016.

[47] K. Omata and M. Yamada, "Artificial neural network and grid search aided optimization of temperature profile of temperature gradient reactor for dimethyl ether synthesis from syngas," *Industrial & Engineering Chemistry Research*, vol. 48, no. 2, pp. 844–849, 2009.

[48] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm," *A Big Comparison for NAS*, https://arxiv.org/abs/1912.06059, 2019.

[49] P. Friedrich, "The three sigma rule," *American Statistician*, vol. 48, no. 2, pp. 88–91, 1994.