Atmosphere **2015**, *6*, 891-907; doi:10.3390/atmos6070891

OPEN ACCESS

atmosphere

ISSN 2073-4433

www.mdpi.com/journal/atmosphere

Article

Forecasting Urban Air Quality via a Back-Propagation Neural Network and a Selection Sample Rule

Yonghong Liu [1,*] Qianru Zhu [2], Dawen Yao [1] and Weijia Xu [3]

[1] School of Engineering, Sun Yat-Sen University, Guangzhou 510275, China;
E-Mail: yaodawen@mail2.sysu.edu.cn

[2] Guangdong Provincial Academy of Environmental Science, Guangzhou 510045, China;
E-Mails: zhuqr2006@163.com (Q.Z.); xuwjia@mail.sysu.edu.cn (W.X.)

[3] Institute of Advanced Technology, Sun Yat-Sen University, Guangzhou 510275, China

* Author to whom correspondence should be addressed; E-Mail: liu_its@163.com.

Academic Editor: Rebecca Sheesley

Received: 28 April 2015 / Accepted: 24 June 2015 / Published: 9 July 2015

Abstract: In this paper, based on a sample selection rule and a Back Propagation (BP) neural network, a new model of forecasting daily SO_2, NO_2, and PM_{10} concentration in seven sites of Guangzhou was developed using data from January 2006 to April 2012. A meteorological similarity principle was applied in the development of the sample selection rule. The key meteorological factors influencing SO_2, NO_2, and PM_{10} daily concentrations as well as weight matrices and threshold matrices were determined. A basic model was then developed based on the improved BP neural network. Improving the basic model, identification of the factor variation consistency was added in the rule, and seven sets of sensitivity experiments in one of the seven sites were conducted to obtain the selected model. A comparison of the basic model from May 2011 to April 2012 in one site showed that the selected model for PM_{10} displayed better forecasting performance, with Mean Absolute Percentage Error (MAPE) values decreasing by 4% and R^2 values increasing from 0.53 to 0.68. Evaluations conducted at the six other sites revealed a similar performance. On the whole, the analysis showed that the models presented here could provide local authorities with reliable and precise predictions and alarms about air quality if used at an operational scale.

Keywords: daily SO_2; NO_2; PM_{10} concentration; a selection sample rule; BP neural network; meteorological similarity; variation consistency

1. Introduction

Air quality has recently become a serious issue in several of the large cities in China. This problem has significant potential for adverse impacts on human health and the environment [1–3]. Therefore, it is extremely important to accurately forecast the concentrations of pollutants to provide guidance for travel advice and governmental policies.

Forecasting the concentrations of air pollutants represents a difficult task due to the complexity of the physical and chemical processed involved. However, many researchers have been focusing on these types of forecasts [4–8]. The most common forecasting approaches are numerical models and statistical models. Numerical models do not require a large quantity of measured data, but they demand sound knowledge of pollution sources, the chemical composition of the exhaust gases, and the physical processes in the atmospheric boundary layer. This crucial knowledge is often limited. Thus, approximations and simplifications are often employed in the modeling process.

In contrast, statistical models usually necessitate a large quantity of measurement data under a large variety of atmospheric conditions. By applying regression and machine learning techniques, a number of functions can be used to fit the pollution data in terms of selected predictors. Neural networks, a subset of statistical models, are usually presented as systems of interconnected neurons that can compute values from inputs by feeding information through the network. Unlike other statistical models, neural networks make no prior assumptions concerning the data distribution. They can model highly nonlinear functions and can be trained for accurate generalization. These features of the neural network make it an attractive alternative to numerical and other statistical models [9–12].

There have been many applications of neural networks in air quality forecasting since the 1990s, and researchers have obtained fairly good results [13–16]. Despite the successful applications of neural networks in the area of atmospheric science, the method has its own weakness and limitations. Studies have shown that there are three main factors that affect neural network effectiveness: network topology, learning algorithm, and learning samples [17,18]. Previous research mainly concentrated on the network structure and learning algorithm, which improved the forecasting accuracy of the network [19–24]. However, when improvements in the network structure and learning algorithm reach a certain degree, improvements in the accuracy of the air quality forecasting models plateau. Therefore, the selection of learning samples has become a vital factor that determines the mapping ability and generalization of the network. This is because the selection can ensure the representativeness of the learning samples and remove unnecessary interference, and thereby improve the forecasting accuracy of the model. Harri Niska *et al.* [21] used a genetic algorithm for selecting the inputs and designing the high-level architecture of a multi-layer perceptron model for forecasting NO_2 concentrations. Sousa *et al.* [22] predicted hourly ozone concentrations based on feed-forward artificial neural networks using principal components as inputs, and they improved the predictions of models by reducing their complexity and eliminating data collinearity.

The main objectives of this paper are to develop a sample filter method for the prediction of the daily NO_2, SO_2, and PM_{10} concentration in the Guangzhou Pearl River Delta region based on a similarity principle of weather and pollutant background concentration. During the development of the prediction models, the selection of parameters is conducted by means of sensitivity experiments and the Back Propagation (BP) neural network is used for data-driven computation. The above actions are all part of an integrated environmental strategy designed and run by the local authorities of Guangzhou, according to the demands of the Action Plan on Prevention and Control of Air Pollution. Currently, this action plan is the most rigorous and systematic framework for improving air quality in China.

2. Data

A significant quantity of observational data under a wide variety of atmospheric conditions was required for this study. The dataset in this paper includes meteorological parameters and pollutant concentrations in Guangzhou, which is located in the south central part of Guangdong Province, China (23°06′ N Latitude, 113°15′ E Longitude).

Real-time monitoring meteorological parameters, including temperature, wind speed, wind direction, rainfall, atmospheric pressure, relative humidity, and solar radiation intensity, were obtained from an automatic air quality monitoring station at Sun Yat-Sen University, located in the Haizhu District of Guangzhou City. Forecasting meteorological data, including temperature, wind speed, wind direction, and rainfall, were obtained from Guangzhou Weather Forecasts [25]. All the data were processed into the daily mean value as needed, according to the National Ambient Air Quality Standards (GB 3095-2012) issued by Environment Protection Administration (EPA) of China [26]. The monitoring meteorological data were used as historical meteorological data in the model, and the forecasting meteorological data were used as the meteorological data of the forecasting day. To reduce the interference of different geographic locations on the monitoring meteorological data, pollutant concentration forecasting of seven state-controlled air quality monitoring sites in urban Guangzhou was performed. Thus, the applied monitoring data of the atmospheric environment were derived from the daily pollutant concentration data from seven state-controlled air quality monitoring sites as reported by the Guangzhou Environmental Protection [27]. These state-controlled air quality monitoring sites are the Guangya Middle School (Num. 1), the Guangzhou No. 5 Middle School (Num. 2), the Guangzhou Environmental Monitor Station (Num. 3), the Experimental Kindergarten of Tianhe Vocational School (Num. 4), Luhu Park (Num. 5), Guangdong University of Business Studies (Num. 6), and the Guangzhou No. 86 Middle School (Num. 7). The data span the period from January 2006 to April 2012, and a total of 23,195 valid samples were used for the paper.

3. Methods

In view of the small variation in weather during our study period, a similarity principle of weather and concentration parameters was applied. The multilayer selection rule for historical samples from Guanghzhou was then constructed. This step is very important for the development of predictive models. The selection of historical samples can improve the similarity between the occurrence of historical pollution and future pollution, and a proper selection can improve the efficiency of

data-driven models (e.g., BP neural networks). This is also in line with the pollution formation, where the main factor affecting the diffusion and transport of pollutants is the different meteorological parameters, and every meteorological parameter has a different influence on NO_2, SO_2, and PM_{10} [28,29]. Thus, the sample selection was based on meteorological similarity and the consistency of the variation trend. The rule was divided into two parts, namely the identification of meteorological parameter similarity and the consistency of the variation trend, *i.e.*, the identification of similarity in background concentrations.

First, a comprehensive correlation analysis of pollutant concentration and meteorological parameters was performed to determine the key factors of the selection rule, and these parameters were also used as inputs into the BP neural network. Next, the three-layer selection sample rule was applied. Finally, we utilized the improved BP neural network for data-driven computation to establish the air quality forecasting model of urban Guangzhou.

3.1. Identification of the Key Factors

A comprehensive correlation analysis of pollutant concentration and meteorological factors was conducted. The number of related days was set to two: the meteorology for the forecasting day and for the day before the forecasting day. Meanwhile, the daily mean value of pollutant concentration two days before the forecasting day was used as an input factor in an attempt to counteract the lack of pollutant emission source data.

A comprehensive analysis of pollutant concentration and meteorological factors was conducted for different pollutants, mainly through correlation analysis and weight analysis of the influencing factors in each pollution scenario. The analysis was intended to identify the degree of influence of each meteorological factor on pollutants, thus resulting in the selection of the factors with the greatest impact on pollutants and the allocation of the corresponding influencing weights. The correlation analysis started with the comparison of two typical pollution scenarios, namely, the ascending or descending periods of each pollutant, and the serious pollution or slight pollution periods. In this way, the degree of influence that the meteorological factors had on pollutants under these two situations was obtained. The average value of the two scenarios was calculated and multiplied with a correlation coefficient to obtain the comprehensive weight of the influence of each meteorological factor on different pollutants.

The ascending and descending periods of each pollutant are defined as the periods when the change in the pollutant concentration between consecutive days exceeds 0.05 mg/m^3. Serious pollution or slight pollution are defined as periods when the Air Pollution Index of the pollutant exceeds 100 or is lower than 20, respectively.

The identification of the influencing weight of each meteorological factor under the above-mentioned periods was achieved using the following steps:

(a) Obtaining the representative data for the meteorological factor

The specific data include the average value of the ascending period M_{iu}, the average value of the descending period M_{id}, the maximum value of the analysis period $M_{i\max}$, the minimum value $M_{i\min}$ of the analysis period, and the overall average value $M_{i\mathrm{adv}}$. The i represents the specific meteorological factor.

(b) Numerical normalization

(c) Variation analysis of the meteorological factor (D_i)

$$D_i = \frac{M_{iu}' - M_{id}'}{M_{iadv}'} \tag{1}$$

(d) Computation of the influencing weight

$$w_i = \frac{D_i}{\sum_{i=1}^{n} D_i} \tag{2}$$

Finally, the comprehensive influencing weights between meteorology factors and pollutant concentrations were determined by the following equation:

$$r = R \times (w_1 + w_2) / 2 \tag{3}$$

where r is the comprehensive influencing weight between the meteorology factor and the pollutant concentration; R is the correlation coefficient between the meteorology factor and the pollutant concentration; w_1 is the influencing weight in the ascending or descending period; and w_2 is the influencing weight in the serious or slight pollution periods.

3.2. A Selection Sample Rule Based on the Similarity Principle

Multiple meteorological factors create a variety of meteorological parameter spaces that impose different impacts on the transport and diffusion of pollutants. During air quality forecasting, if the appropriate meteorological space is found, the intrinsic relationship between multiple physical quantities and the pollutant will have a reference. An appropriate set of samples was selected for the main influencing factors such that forecasting could be targeted, and the mapping ability and generalization of the network could be improved. Thus, three-layer sample screening principles based on meteorological similarity criteria were proposed.

3.2.1. The Basic Description

The first level of screening identifies samples where the similarity of each meteorological factor reaches a certain threshold value range. The screened samples should conform to the following formula:

$$\Delta y_j \le y_{j\,set}, \text{ where, } \Delta y_j = | y_{j\,pre} - y_{j\,sam} | \tag{4}$$

where $y_{j\,pre}$ is the meteorological factor on the day of forecasting; $y_{j\,sam}$ is the meteorological factor of the sample; Δy_j is the meteorological similarity of the meteorology factors between the sample and the day of forecasting; j is the specific meteorological factor; and $y_{j\,set}$ is the threshold value screened by the meteorological factor, forming a primary threshold matrix Y. In this matrix, the threshold value can change dynamically according to the sample size demanded.

The second level of screening applies a threshold value range for total weighted meteorological similarity. The screened samples should conform to the following formula:

$$S \le S_{set}, \text{ where, } S = \sum_{j \le Mnum} (w_j \cdot \Delta y_j) \quad (5)$$

where S is the entire meteorological similarity; S_{set} is the threshold value screened by the entire meteorological similarity; w_j is the weight of each meteorological factor, forming the weight matrix W; and M_{num} is the number of meteorological factors.

The third level of screening identifies the n samples with the highest meteorological similarity. The screened samples should conform to the following formula:

$$Q_{num} \le n \quad (6)$$

where Q_{num} is the number of samples in the sequenced sample column, and n is the number of samples needed.

Among these criteria, the selection of the weight matrices and the threshold matrices is key to obtaining high quality samples. Hence, the following identification approaches for weight matrices and threshold matrices were adopted.

3.2.2. Identification of w_j

The establishment of the weight matrix w_j was integrated with the selection of model input factors, and a comprehensive correlation analysis of pollutant concentration and meteorological factors was performed. While choosing the input parameters of the neural network, the weight matrix of the selection sample rule was also established.

3.2.3. Identification of $y_{j\,set}$

The establishment of the threshold matrix $y_{j\,set}$ was accomplished via the orthogonal test method, which is a highly efficient experimental design method used for the arrangement of multi-factor experiments and the search for optimal horizontal combinations [30]. For the different pollutants, we set different levels of factors and selected some representative experimental points (horizontally mixed) for the experiments. The optimal horizontal combination was selected to generate the threshold matrix of the selection sample rule [31].

Based on the results of the above weight matrix w_j, the tested experimental factors were identified. In accordance with prior knowledge, the level of each experimental factor was confirmed. The minimum absolute error of the forecasting model was adopted as the experimental objective to seek the optimal combination and finally identify the sample optimization threshold matrix.

3.3. Identification of the Variation Trend Consistency

There will be some scenarios in which wind speed decreases in history but increases on the prediction day compared with the previous day, based on the selection rule stated above in Section 3.2. Such a scenario will lead to an error in the prediction model for use in the BP neural network. Therefore, it is necessary to identify the variation trend consistency.

The factors considered were deduced according to the weight matrix of the selection rule (see Section 3.1) and the principles of the pollution formation. The chosen factors were rainfall, wind speed, and background concentration. However, sensibility experiments were still needed to determine the key

factor for NO_2, PM_{10}, and SO_2. The details of the experimental results will be introduced in the following section.

3.3.1. Variation Trend Consistency for Wind Speed

Because wind speed is a vector, wind speed is described as w_x, w_y.

$$w_x = w_s \cdot \cos(w_d) \text{ and } w_y = w_s \cdot \sin(w_d)$$

where w_s is the recorded wind speed and w_d is the recorded wind direction.

Thus, the steps for the identification of the variation trend consistency for wind speed are as follows:

(1) Calculate the variation between the forecasting day and the day before.

$$\Delta(ws_1)^2 = \left[(w_{x-p})^2 + (w_{y-p})^2\right] - \left[(w_{x-p-1})^2 + (w_{y-p-1})^2\right] \quad (7)$$

where $\Delta(ws_1)^2$ is the difference between the squared values of wind speed on the day of forecasting and the day before; w_{x-p} and w_{y-p} are the two wind vectors on the day of forecasting; and w_{x-p-1} and w_{y-p-1} represent the two wind vectors before the day of forecasting.

(2) Calculate the variation between the two adjacent days in the samples selected in Section 3.1,

$$\Delta(ws_2)^2 = \left[(w_{x-t})^2 + (w_{y-t})^2\right] - \left[(w_{x-t-1})^2 + (w_{y-t-1})^2\right] \quad (8)$$

where $\Delta(ws_2)^2$ is the difference between the squared values of wind speed on the forecasting day and the day before; w_{x-t} and w_{y-t} are the two wind vectors on the forecasting day; and w_{x-t-1} and w_{y-t-1} are the two wind vectors on the day before the forecasting day.

(3) Identify whether the wind speed in the forecasting data shows the same tendency of ascending or descending as that in the selected samples. If the tendency is the same, the samples are reserved; otherwise, the samples are removed.

3.3.2. The Variation Trend Consistency Identification of Rainfall

The variation in the rainfall levels in the forecasting data was calculated using the following formula:

$$\Delta RF_1 = RF_p - RF_{p-1} \quad (9)$$

The variation in the historical rainfall levels was calculated using the following formula:

$$\Delta RF_2 = RF_t - RF_{t-1} \quad (10)$$

We then identified whether the rainfall level in the forecasting data showed the same tendency of ascending or descending as that in the sample data. If similar, the samples are reserved; otherwise, the samples are removed.

3.3.3. Similarity Identification of Background Concentration

The following steps were used to conduct the similarity identification of the background concentration:

(1) The background concentration on the day of forecasting is calculated as follows:

$$BC_1=0.6BC_{P\text{-}1}+0.4BC_{P\text{-}2} \tag{11}$$

(2) The background concentration in the sample data is calculated as follows:

$$BC_2=0.6BC_{t\text{-}1}+0.4BC_{t\text{-}2} \tag{12}$$

(3) Identify whether the background concentration in the forecasting data and the absolute difference of the background concentration on the day of forecasting is in the range of the threshold value. If they are in the range, the samples are reserved; otherwise, they are removed.

$$ABS(BC1\text{-}BC2)<=Set \tag{13}$$

3.4. Improvements in BP Neural Network

Due to its strong learning and generalization ability, a BP neural network was used as the data-driven computation method [32]. In this paper, a BP neural network with three layers was applied to predict the daily concentrations of NO_2, PM_{10}, and SO_2. The layers included an input layer, a hidden layer, and an output layer. The data described in Section 2 were divided into training, validation and test sets. The training and validation sets were from January 2006 to April 2011 in seven air quality monitoring sites, of which 80% of these data were randomly selected for the training set; the remaining 20% of the data comprised the validation set. In addition, the data from May 2011 to April 2012 were used for the test set, aiming to test and compare the model performance in seven air quality monitoring sites. There are two main components affecting pollutant concentration: emission sources and pollutant transmission and diffusion conditions. The key factor that affects pollutant transmission and diffusion in a city is the meteorological conditions. Therefore, the meteorological factors identified in Section 3.1 were considered as the major input factors for the BP neural network. According to the conclusions in the literature [33,34], the daily concentrations of NO_2, PM_{10}, and SO_2 for the two days before the forecasting day were also used as input factors for the BP neural network to reduce the influencefor lacking emissions data. The final number of variables used in the input layer (NInput) in each forecast model is shown in Table 1.

The neuron number of the hidden layer is half that of the input layer [35]. Different neural network structures were established for NO_2, PM_{10}, and SO_2. The neuron in the output layer was regarded as the forecasted daily concentration of NO_2, PM_{10}, and SO_2.

The training termination conditions in the BP neural network were also changed to improve the overall accuracy of the forecasting model. When the average relative error of all training samples reached a specified error value, the training would cease. The specified error value was determined by experiments for different error. For NO_2, PM_{10}, and SO_2, the optimal specified error values were 0.5, 0.4, and 0.35, respectively. Every group training sample was processed five times, which means that five groups of models were developed. The model with the least average relative error was selected as the prediction model, reducing the randomness of the BP neural network.

Table 1. Forecasting results of the seven groups of sensitivity experiments.

Pollutants	Experiments	N Input	Mean (mg/m^3)	MAE (mg/m^3)	MAPE	R	TFA	Ef	Af
SO_2	Basic (Group 1)	10	0.027	0.009	37.4	0.422	0500	−0.322	1.513
	RF * (Group 2)	10	0.027	0.009	36.6	0.510	0.536	0.010	1.543
	WS (Group 3)	10	0.027	0.010	43.2	0.304	0.464	−0.583	1.693
	BC (Group 4)	10	0.027	0.009	40.3	0.345	0.483	−0.937	1.577
	RF + WS (Group 5)	10	0.027	0.009	38.5	0.430	0.482	−0.192	1.501
	RF + BC (Group 6)	10	0.027	0.011	49.7	0.118	0.464	−1.726	1.575
	WS + BC (Group 7)	10	0.027	0.012	52.8	0.178	0.393	−1.174	1.716
PM_{10}	basic(Group 1)	7	0.105	0.025	26.6	0.536	0.492	0.210	1.297
	RF (Group 2)	7	0.105	0.026	28.8	0.476	0.433	0.108	1.319
	WS (Group 3)	7	0.105	0.025	26.2	0.527	0.483	0.190	1.289
	BC (Group 4)	7	0.105	0.024	24.6	0.563	0.500	0.225	1.280
	RF + WS (Group 5)	7	0.105	0.025	27.8	0.479	0.417	0.159	1.315
	RF + BC * (Group 6)	7	0.105	0.023	22.7	0.672	0.550	0.348	1.269
	WS + BC (Group 7)	7	0.105	0.024	26.9	0.581	0.417	0.317	1.290
NO_2	Basic (Group 1)	10	0.073	0.020	25.0	0.680	0.550	0.261	1.340
	RF (Group 2)	10	0.073	0.020	24.1	0.660	0.533	0.199	1.345
	WS (Group 3)	10	0.073	0.018	22.7	0.702	0.533	0.352	1.291
	BC (Group 4)	10	0.073	0.019	23.7	0.715	0.517	0.337	1.315
	RF + WS (Group 5)	10	0.073	0.018	23.7	0.723	0.617	0.386	1.298
	RF + BC (Group 6)	10	0.073	0.019	24.3	0.716	0.483	0.380	1.306
	WS + BC * (Group 7)	10	0.073	0.018	22.5	0.688	0.567	0.397	1.271

Note: * the Selected Model determined by making experiments.

3.5. Indices of Model Evaluation

We used the following indicators to evaluate the models: Mean absolute error (MAE), Mean Absolute Percentage Error (MAPE), Correlation coefficient (R), tendency forecasting accuracy (TFA), Nash–Sutcliffe coefficient of efficiency (Ef), and Accuracy factor (Af) [36]. The TFA is the forecasting accuracy rate determination for the upward or downward trend of pollutant concentrations over two consecutive days on the basis of monitoring results. Ef, an indicator of the model fit, is a normalized measure ($-\infty$ to 1) that compares the mean square error generated by a particular model simulation to the variance of the target output sequence. An Ef value closer to 1 indicates better model performance; an Ef value of zero indicates that the model is, on average, performing only as good as the use of the

mean target value for prediction, and an Ef value < 0 indicates an altogether questionable choice of the model. Af is a simple multiplicative factor indicating the spread of the results around the prediction. The larger the Af value, the less accurate the average estimate.

The MAE, MAPE, TFA, Af and Ef are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_{pre,i} - y_{mon,i}\right| \tag{14}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left|y_{pre,i} - y_{mon,i}\right|}{y_{mon,i}} \times 100\right) \tag{15}$$

$$TFA = \frac{A}{N} \tag{16}$$

$$E_f = 1 - \frac{\sum_{i=1}^{N}(y_{pre,i} - y_{mon,i})^2}{\sum_{i=1}^{N}(y_{pre,i} - \overline{y}_{mon})^2} \tag{17}$$

$$A_f = 10^{\sum_{i=1}^{N}\frac{\left|\log\left(\frac{y_{pre,i}}{y_{mon,i}}\right)\right|}{N}} \tag{18}$$

where y_{pre} and y_{mon} are the predicted and measured values, respectively, and $\overline{y}_{mon}$ is the mean of the measured values of the response variable. N is the total number of the observations. A is the number of correct forecasts for the upward or downward trend of pollutant concentrations over two consecutive days.

4. Results and Discussion

4.1. The Results of the Sensitivity Experiments in Guangzhou No. 5 Middle School (Num. 2)

As described in Section 3.3, sensitivity experiments were performed to determine the key factors. The data were obtained from the Guangzhou No. 5 Middle School site. Seven group experiments were performed for SO_2, PM_{10}, and NO_2. The first experiment (called "Group 1") was made by the model based on the selection rules described in Section 3.2. That is to say, Group 1 was run using the Basic Model. Besides these selection rules, the second to fourth experiments were conducted based on the variation trend consistency identification of rainfall (RF), wind speed (WS), and background concentration (BC), while the fifth to seventh experiments were considerations of RF + WS, RF + BC, and WS + BC. These experiments were referred to as Group 2, Group 3, Group 4, Group 5, Group 6, and Group 7, respectively. Table 1 summarizes the results of the seven groups of sensitivity experiments. The models with the best performance were selected (termed the Selected Models).

For PM_{10}, the value of Ef and Af of Group 6 were much closer to 1.0 compared with the other models. Compared with Group 1, the Mean Absolute Percentage Error (MAPE) of Group 6 was 4% lower (0.227), R increased by almost 14%, and TFA increased by nearly 6% (0.550). For NO_2, Group 7 had the best results with an MAPE of only 0.225, an R value of 0.688, and a TFA value of 0.567. The Ef and Af of Group 7 were 0.397 and 1.271, respectively, which were much closer to 1.0 than the other experiments. Group 2 had the most ideal experimental results for SO_2; the MAPE was

0.366, R and TFA were both higher than 0.5, and Ef was the only positive value. In contrast to the PM_{10} and NO_2 results, the SO_2 experiments based on BC did not produce the best results. This scenario is perhaps due to a non-obvious variation in daily SO_2 concentrations.

4.2. Errors of the Selected Models of Num. 2 for May 2011 to April 2012

The forecasting results are shown in a scatter diagram of the predicted *versus* the observed concentrations (Figure 1). The distribution of SO_2 is relatively dispersed, which is due to the diversity of the influencing factors and the complexity of dynamic processes. Singh *et al.* [36] forecast respirable suspended particulate matter (RSPM), SO_2, and NO_2. The results showed that compared with the two other pollutants, the degree of dispersion in the scatter diagram of the monitored and predicted SO_2 values was higher. Kurt *et al.* [37] used a neural network to build models of SO_2, PM_{10}, and CO. The error distribution of the SO_2 forecasting model based on the data from two days prior ranged from 37% to 40%, and the model was the least accurate of the three. The distributions of PM_{10} and NO_2 were relatively better, *i.e.*, the line fitted the correlation at 0.5 and above. The forecasting results were stable and the model performed well.

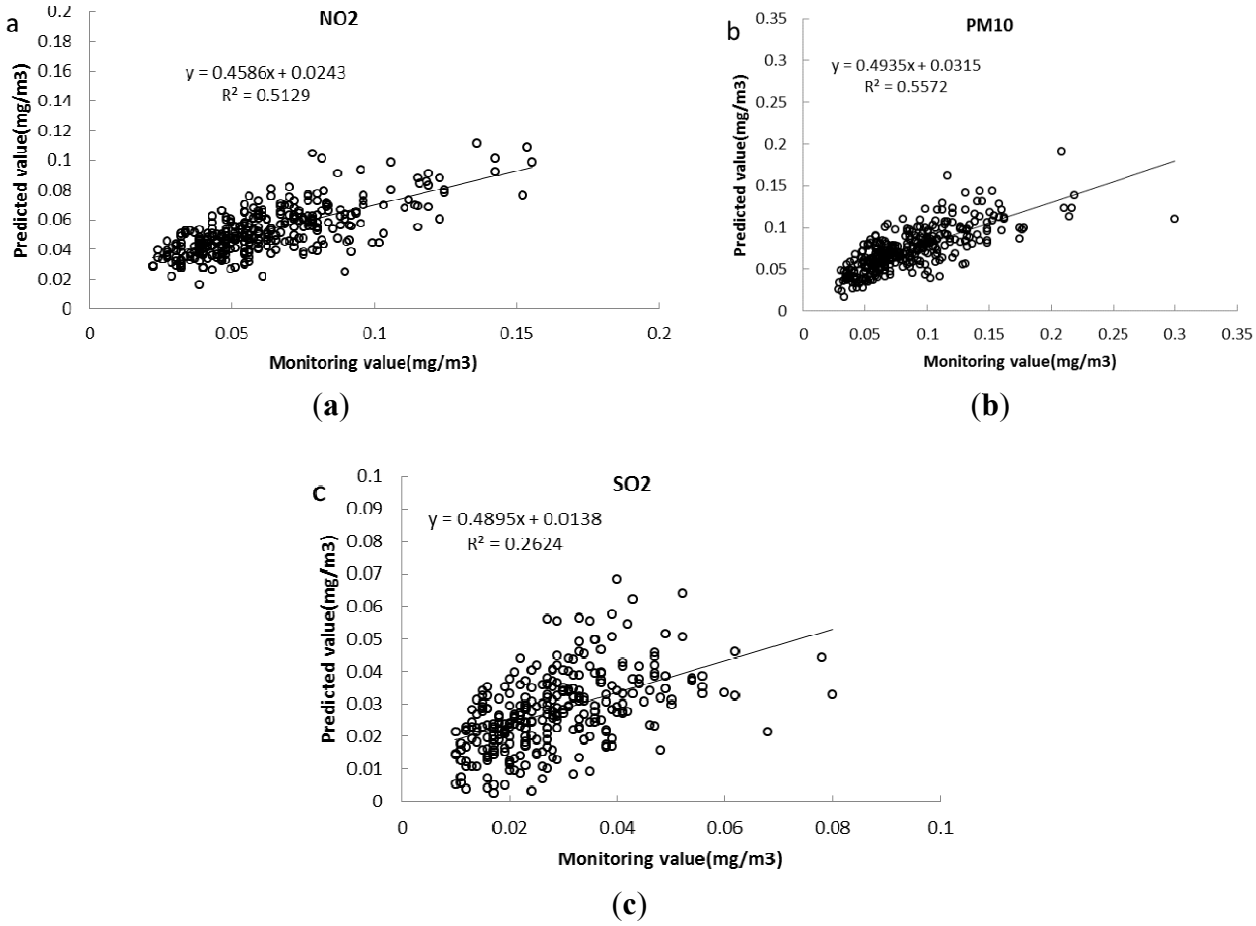

Figure 1. (a-c) Scatter plots of predicted *versus* observed NO_2, PM_{10}, SO_2 concentrations for Num. 2.

Errors in the selected model for Guangzhou No. 5 Middle School (Num. 2) from May 2011 to April 2012 are shown in Figure 2. Overall, the monthly prediction accuracy of SO_2 was higher than PM_{10} and NO_2, and the NO_2 model performed better than the PM_{10} model. The highest errors for SO_2,

PM_{10}, and NO_2 were observed in February, where the daily concentrations were almost the highest due to the bad weather; the BP neural network is not sensitive to extremely high or low values [33,34]. However, the MAPE of the SO_2, PM_{10}, and NO_2 models were 0.383, 0.353, and 0.290, respectively. These MAPE values are acceptable for operational forecasts.

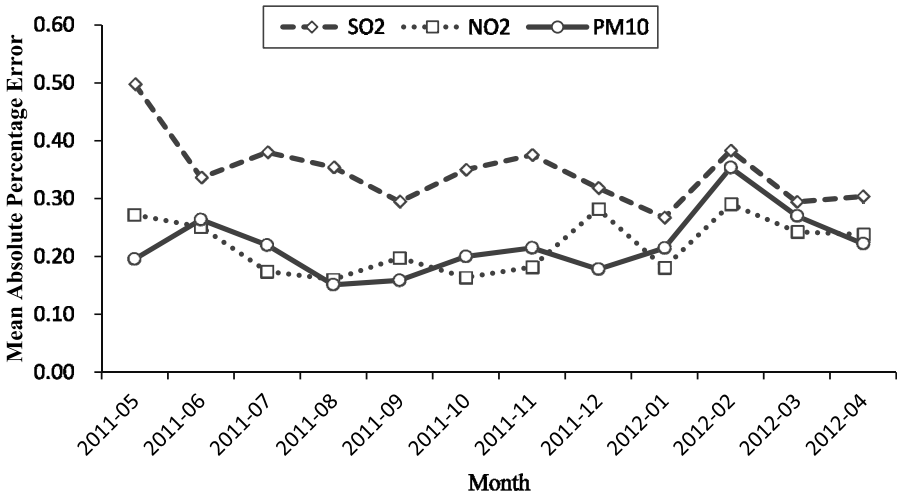

Figure 2. MAPE of models for Num. 2 from May 2011 to April 2012.

4.3. Errors in the Selected Models for Others Sites

The selected model for SO_2, NO_2, and PM_{10} was tested in the remaining six sites (detailed description in Section 2) in the urban district of Guangzhou, and a comparison was made between the Selected Model and the Basic Model. The results are shown in Table 2. On the whole, the Selected Model was equal to or better than the Basic Model for SO_2, NO_2, and PM_{10}. As for SO_2, the MAPE of the Selected Model decreased from 0.417 to 0.377, the correlation increased from 0.409 to 0.477, the TFA increased from 0.490 to 0.517. In addition, the Ef and Af were closer to 1 compared with the Basic Model. Adding the sample optimization rules to the variation tendency identification of the rainfall level changes improved the forecast accuracy of the different pollutants to different degrees at every site. For PM_{10}, the MAPE of the Selected Model was 0.250 for the six sites, which was almost 0.10 lower than that of the Basic model. The correlation was greater than 0.7, and the TFA increased by 24%, from 0.421 to 0.523. Adding the variation tendency identification of the rainfall level changes and the similarity identification of the background concentrations to the model resulted in an effective improvement of the forecast accuracy of PM_{10}. Regarding NO_2, adding the variation tendency identification of the wind speed changes and the similarity identification of the background concentrations did not greatly improve the forecast results. The Selected Model is useful for the six sites, and the errors of the model are acceptable for application purposes.

Table 2. Comparisons between the Selected and Basic Model in the remaining six sites.

Pollutant	Site	Model	Mean (mg/m^3)	MAE (mg/m^3)	MAPE	R	TFA	Ef	Af
SO_2	Num. 1	Basic	0.024	0.008	36.8	0.525	0.506	0.159	1.459
		Selected	0.024	0.008	34.9	0.614	0.525	0.237	1.451
	Num. 3	Basic	0.027	0.010	43.6	0.418	0.511	−0.164	1.539
		Selected	0.027	0.010	40.4	0.409	0.475	−0.181	1.548
	Num. 4	Basic	0.023	0.009	44.2	0.394	0.509	−0.301	1.567
		Selected	0.023	0.009	41.3	0.456	0.527	−0.332	1.541
	Num. 5	basic	0.022	0.007	35.6	0.441	0.455	−0.019	1.468
		Selected	0.022	0.007	31.6	0.472	0.515	0.059	1.408
	Num. 6	Basic	0.027	0.011	42.8	0.355	0.466	0.055	1.587
		Selected	0.027	0.010	39.6	0.451	0.508	0.019	1.551
	Num. 7	basic	0.036	0.015	47.8	0.298	0.527	−0.580	1.662
		Selected	0.036	0.013	41.2	0.422	0.561	−0.239	1.563
PM_{10}	Num. 1	Basic	0.083	0.023	26.2	0.656	0.438	0.348	1.328
		Selected	0.083	0.022	24.9	0.713	0.509	0.397	1.335
	Num. 3	Basic	0.067	0.018	32.1	0.604	0.132	0.348	1.354
		Selected	0.067	0.018	26.8	0.694	0.542	0.459	1.322
	Num. 4	basic	0.061	0.017	31.6	0.680	0.493	0.454	1.350
		Selected	0.061	0.017	26.4	0.741	0.506	0.523	1.317
	Num. 5	basic	0.067	0.016	24.7	0.742	0.465	0.537	1.268
		Selected	0.067	0.016	22.7	0.729	0.531	0.487	1.267
	Num. 6	Basic	0.063	0.018	30.4	0.583	0.493	0.301	1.358
		Selected	0.063	0.019	29.2	0.589	0.492	0.247	1.390
	Num. 7	basic	0.087	0.022	25.4	0.682	0.467	0.408	1.308
		Selected	0.087	0.022	23.3	0.717	0.525	0.431	1.288
NO_2	Num. 1	basic	0.061	0.013	20.9	0.688	0.483	0.392	1.248
		Selected	0.061	0.013	20.5	0.715	0.500	0.448	1.243
	Num. 3	Basic	0.068	0.016	22.2	0.596	0.463	0.226	1.272
		Selected	0.068	0.015	21.5	0.676	0.557	0.320	1.266
	Num. 4	Basic	0.052	0.010	21.9	0.685	0.511	0.456	1.232
		Selected	0.052	0.010	19.3	0.722	0.541	0.502	1.215
	Num. 5	Basic	0.038	0.009	25.8	0.613	0.454	0.363	1.285
		Selected	0.038	0.009	23.0	0.599	0.462	0.308	1.267
	Num. 6	Basic	0.053	0.014	26.8	0.757	0.497	0.405	1.337
		Selected	0.053	0.015	24.6	0.728	0.528	0.310	1.334
	Num. 7	Basic	0.041	0.010	27.4	0.668	0.476	0.435	1.305
		Selected	0.041	0.009	23.1	0.700	0.505	0.465	1.269

5. Conclusions

In this paper, based on a selection sample rule and BP neural network, a new model of forecasting daily SO_2, NO_2, and PM_{10} concentrations in seven Guangzhou sites was developed.

(1) A meteorological similarity principle was applied in the development of the selection sample rule. Key meteorological factors influencing the daily SO_2, NO_2, and PM_{10} concentrations were determined and weight matrices and threshold matrices were generated. A basic model was

then developed based on the improved BP neural network. The selection sample rule consisted of three layers.

(2) In improving the basic model, identification of the variation consistency of some factors was added in the rule, and seven sets of sensitivity experiments (one in each of the seven sites) were conducted to obtain the selected model. These experiments determined that the variation consistency of the rainfall level added to the SO_2 forecast model, the rainfall level variation tendency and the background concentration similarity identification added to the PM_{10} forecast model, while wind speed variation identification and background concentration similarity identification added to the NO_2 forecast model. The improved BP neural network was also used for data-driven computation.

(3) Evaluations in the site by comparison of the basic model from May 2011 to April 2012 showed the selected model for PM_{10} displayed better forecasting performance, with MAPE values decreasing by 4% and R^2 values increasing from 0.53 to 0.68. The selected model for NO_2 had little improvements compared with the basic model, while the MAPE values of the selected model for SO_2 were as high as 36.6% with R^2 values of 0.51.

(4) Evaluations conducted at the six other sites revealed similar performances. The MAPE values of the selected models for SO_2, PM_{10}, and NO_2 were 37.7%, 25.0%, and 22.0%, respectively. Of course, the above results showed that the SO_2 model may be further improved in future research, by developing a combined model or by considering the interaction of atmospheric pollutants.

Acknowledgments

This work was completely supported by the National Natural Science Foundation of China (No. 51108471).

Author Contributions

Yonghong Liu, Qianru Zhu conceived and designed the model and experiments; Dawen Yao and Weijia Xu collected and analyzed the data; Yonghong Liu wrote the paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Dimitriou, K.; Kassomenos, P.A.; Paschalidou, A.K. Assessing air quality with regards to its effect on human health in the European Union through air quality indices. *Ecol. Indic.* **2013**, *27*, 108–115.
2. Pope, C.A., III; Burnett, R.T.; Thun, M.J.; Calle, E.E.; Krewski, D.; Ito, K.; Thurston, G.D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA* **2002**, *287*, 1132–1141.

3. Li, G.; Sang, N. Delayed rectifier potassium channels are involved in SO2 derivative-induced hippocampal neuronal injury. *Ecotoxicol. Environ. Saf.* **2009**, *72*, 236–241.
4. Juhos, I.; Makra, L.; Tóth, B. Forecasting of traffic origin NO and NO2 concentrations by Support Vector Machines and neural networks using Principal Component Analysis. *Simul. Model. Pract. Theory* **2008**, *16*, 1488–1502.
5. Finardi, S.; de Maria, R.; D'Allura, A.; Cascone, C.; Calori, G.; Lollobrigida, F. A deterministic air quality forecasting system for Torino urban area, Italy. *Environ. Model. Softw.* **2008**, *23*, 344–355.
6. Dong, M.; Yang, D.; Kuang, Y.; He, D.; Erdal, S.; Kenski, D. PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining. *Expert Syst. Appl.* **2009**, *36*, 9046–9055.
7. Pai, T.Y.; Ho, C.L.; Chen, S.W.; Lo, H.M.; Sung, P.J.; Lin, S.W.; Lai, W.J.; Tseng, S.C.; Ciou, S.P.; Kuo, J.L.; Kao, J.T. Using seven types of GM (1, 1) model to forecast hourly particulate matter concentration in Banciao City of Taiwan. *Water Air Soil Pollut.* **2011**, *217*, 25–33.
8. Pai, T.Y.; Hanaki, K.; Chiou, R.J. Forecasting Hourly Roadside Particulate Matter in Taipei County of Taiwan Based on First-Order and One-Variable Grey Model. *CLEAN Soil Air Water* **2013**, *41*, 737–742.
9. Comrie, A.C. Comparing neural networks and regression models for ozone forecasting. *J. Air Waste Manag. Assoc.* **1997**, *47*, 653–663.
10. Schlink, U.; Dorling, S.; Pelikan, E.; Nunnari, G.; Cawley, G.; Junnine, H.; Greig, A.; Foxall, R.; Eben, K.; Chatterton, T.; *et al*. A rigorous inter-comparison of ground-level ozone predictions. *Atmos. Environ.* **2003**, *37*, 3237–3253.
11. Kukkonen, J.; Partanen, L.; Karppinen, A.; Ruuskanen, J.; Junninen, H.; Kolehmainen, M.; Niska, H.; Dorling, S.; Chatterton, T.; Foxall, R.; *et al*. Extensive evaluation of neural network models for the prediction of NO_2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmos. Environ.* **2003**, *37*, 4539–4550.
12. Diaz-Robles, L.A.; Ortega, J,C.; Fu, J.S.; Reed, G.D.; Chow, J.C.; Watson, J.G.; Moncada-Herrera, J.A. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmos. Environ.* **2008**, *42*, 8331–8340.
13. Yi, J.; Prybutok, V.R. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environ. Pollut.* **1996**, *92*, 349–357.
14. Grivas, G.; Chaloulakou, A. Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. *Atmos. Environ.* **2006**, *40*, 1216–1229.
15. Hooyberghs, J.; Mensink, C.; Dumont, G.; Fierens, F.; Brasseur, O. A neural network forecast for daily average PM10 concentrations in Belgium. *Atmos. Environ.* **2005**, *39*, 3279–3289.
16. Paschalidou, A.K.; Karakitsios, S.; Kleanthous, S.; Kassomenos, P.A. Forecasting hourly PM10 concentration in Cyprus through artificial neural networks and multiple regression models: Implications to local environmental management. *Environ. Sci. Pollut. Res.* **2011**, *18*, 316–327.
17. Zhang, G.; Eddy Patuwo, B.; Hu, Y.M. Forecasting with artificial neural networks: The state of the art. *Int. J. Forecast.* **1998**, *14*, 35–62.

18. Gardner, M.W.; Dorling, S.R. Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmos. Environ.* **1998**, *32*, 2627–2636.
19. Kolehmainen, M.; Martikainen, H.; Ruuskanen, J. Neural networks and periodic components used in air quality forecasting. *Atmos. Environ.* **2001**, *35*, 815–825.
20. Lu, W.Z.; Fan, H.Y.; Lo, S.M. Application of evolutionary neural network method in predicting pollutant levels in downtown area of Hong Kong. *Neurocomputing* **2003**, *51*, 387–400.
21. Niska, H.; Hiltunen, T.; Karppinen, A.; Ruuskanen, J.; Kolehmainen, M. Evolving the neural network model for forecasting air pollution time series. *Eng. Appl. Artif. Intell.* **2004**, *17*, 159–167.
22. Sousa, S.I.V.; Martins, F.G.; Alvim-Ferraz, M.C.M.; Pereira, M.C. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environ. Model. Softw.* **2007**, *22*, 97–103.
23. Al-Alawi, S.M.; Abdul-Wahab, S.A.; Bakheit, C.S. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environ. Model. Softw.* **2008**, *23*, 396–403.
24. Pires, J.C.M.; Gonçalves, B.; Azevedo, F.G.; Carneiro, A.P.; Rego, N.; Assembleia, A.J.B.; Silva, P.A.; Lima, J.F.B.; Alves, C.; Martins, F.G. Optimization of artificial neural network models through genetic algorithms for surface ozone concentration forecasting. *Environ. Sci. Pollut. Res.* **2012**, *19*, 3228–3234.
25. Guangzhou Weather Forecasts. Available online: http://www.tqyb.com.cn/ (accessed on 1 January 2013).
26. Ministry of Environmental Protection of China. *Ambient Air Quality Standards*; China Environmental Science Press: Beijing, China, 2012
27. Guangzhou Environmental Protection. Available online: http://www.gzepb.gov.cn/comm/apidate.asp (accessed on 1 January 2013).
28. Elminir, H.K. Dependence of urban air pollutants on meteorology. *Sci. Total Environ.* **2005**, *350*, 225–237.
29. Pearce, J.L.; Beringer, J.; Nicholls, N.; Hyndman, R.J.; Tapper, N.J. Quantifying the influence of local meteorology on air quality using generalized additive models. *Atmos. Environ.* **2011**, *45*, 1328–1336.
30. Yu, Z.Y.; Yuan, J.Y.; Yu, Y.; Zhang, W.; Wu, Z.H. Research on Relationship of Control Parameters of Cement Concrete Strength by Orthogonal Test Method. *J. Huangshi Inst. Technol.* **2012**, *3*, 38–41.
31. Lin, Y.; Yang, X.G.; MA, Y.Y. An Analysis of Factors Causing Congestion with the Application of Orthogonal Experimental Design Method. *Syst. Eng.* **2005**, *10*, 39–43.
32. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366.
33. Li, L. Study on urban air quality forecast model based on adaptive artificial neural network. M.Sc. Thesis, Sun Yet-Sen University, Guangzhou, China, 2011.
34. Zhu, Q.R. Study on combined urban air quality forecast model. M.Sc. Thesis, Sun Yet-sen University, Guangzhou, China, 2013.

35. Cai, M.; Yin, Y.; Xie, M. Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. *Transp. Res. Part D Transp. Environ.* **2009**, *14*, 32–41.
36. Singh, K.P.; Gupta, S.; Kumar, A.; Shukla, S.P. Linear and nonlinear modeling approaches for urban air quality prediction. *Sci. Total Environ.* **2012**, *426*, 244–255.
37. Kurt, A.; Oktay, A.B. Forecasting air pollutant indicator levels with geographic models 3days in advance using neural networks. *Expert Syst. Appl.* **2010**, *37*, 7986–7992.

© 2015 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).