Atmospheric Environment 184 (2018) 129–139

Contents lists available at ScienceDirect

Atmospheric Environment

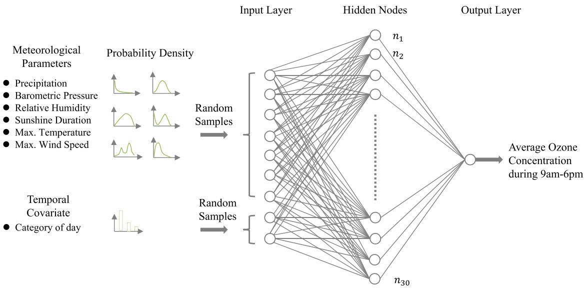
journal homepage: www.elsevier.com/locate/atmosenv

Artificial neural network model for ozone concentration estimation and Monte Carlo analysis

Meng Gao[a], Liting Yin[a,b], Jicai Ning[a,*]

[a] *Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai, 264003, China*
[b] *University of Chinese Academy of Sciences, Beijing, 100049, China*

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Air pollution
Artificial neural network
Monte Carlo simulation
Uncertainty analysis
Sensitivity analysis

ABSTRACT

Air pollution in urban atmosphere directly affects public-health; therefore, it is very essential to predict air pollutant concentrations. Air quality is a complex function of emissions, meteorology and topography, and artificial neural networks (ANNs) provide a sound framework for relating these variables. In this study, we investigated the feasibility of using ANN model with meteorological parameters as input variables to predict ozone concentration in the urban area of Jinan, a metropolis in Northern China. We firstly found that the architecture of network of neurons had little effect on the predicting capability of ANN model. A parsimonious ANN model with 6 routinely monitored meteorological parameters and one temporal covariate (the category of day, i.e. working day, legal holiday and regular weekend) as input variables was identified, where the 7 input variables were selected following the forward selection procedure. Compared with the benchmarking ANN model with 9 meteorological and photochemical parameters as input variables, the predicting capability of the parsimonious ANN model was acceptable. Its predicting capability was also verified in term of warming success ratio during the pollution episodes. Finally, uncertainty and sensitivity analysis were also performed based on Monte Carlo simulations (MCS). It was concluded that the ANN could properly predict the ambient ozone level. Maximum temperature, atmospheric pressure, sunshine duration and maximum wind speed were identified as the predominate input variables significantly influencing the prediction of ambient ozone concentrations.

1. Introduction

Tropospheric ozone, a major air pollutant in urban areas, has adverse effects on human health (Fuhrer et al., 1997; Fontes et al., 2014; Liu et al., 2018). Most ozone in the troposphere is not directly emitted to the atmosphere (Munir et al., 2013), but produced in the atmosphere by the photochemical oxidation of volatile organic compounds (VOCs) in the presence of nitrogen oxides (NO_x) (Jenkin and Clemitshaw, 2000;

[*] Corresponding author.
E-mail address: jcning@yic.ac.cn (J. Ning).

https://doi.org/10.1016/j.atmosenv.2018.03.027
Received 26 July 2017; Received in revised form 7 March 2018; Accepted 11 March 2018
Available online 12 March 2018

García et al., 2011). Previous studies revealed that ambient ozone concentrations were strongly connected with road-traffic and meteorology in urban areas (Revlett, 1978; Lelieveld and Crutzen, 1990; Yi and Prybutok, 1996; Baur et al., 2004; Munir et al., 2013; Zanis et al., 2014). Vehicle emission is considered as one of the major sources of ozone precursors, particularly in large cities (Jenkin, 2008). Solar radiation has the greatest effect on the photochemical reactions generating tropospheric ozone (García et al., 2011). Consequently, it was observed that ambient concentrations were the highest during hot and sunny summer episodes characterized by low ventilation (a result of low winds and low vertical mixing) (Luna et al., 2014; Zanis et al., 2014; Biancofiore et al., 2015). On the contrary, precipitation and high relative humidity result in low ozone concentration due to the reduction of the photochemical production efficiency and an increase of wet deposition (Lelieveld and Crutzen, 1990; García et al., 2011). Atmospheric movements of the air (i.e. winds) causes the spread of high concentrations of pollutants (in this case the ozone and its precursors); therefore, wind speed and direction are also highly correlated with ambient ozone concentration (Revlett, 1978; García et al., 2011; Luna et al., 2014; Biancofiore et al., 2015).

During the last few decades, various modeling techniques, either mechanistic or statistical, have been applied to predict ambient ozone concentration. Mechanistic models explicitly simulate the life cycle of air pollutants including formation, emission, transport and disappearance in numerical way (Brasseur et al., 1998; Russell and Dennis, 2000; Fusco and Logan, 2003; Schaap et al., 2008). The performance of these mechanistic models was usually constrained by the quantity and quality of input datasets (Han et al., 2008). Alternatively, a few of statistical techniques such as regression or other data-driven methods that requires less information were also commonly used. Multiple linear regression (MLR) and artificial neural network (ANN) were the two popular methods (Yi and Prybutok, 1996; Gardner and Dorling, 1999; Chaloulakou et al., 2003; Wang et al., 2003; Heo and Kim, 2004; García et al., 2011; Luna et al., 2014; Biancofiore et al., 2015; Taylan, 2018), where the ambient ozone concentration was expressed as a function of photochemical and/or meteorological parameters. ANNs can simulate human learning and pattern recognition allowing the information to be extracted from imprecise and nonlinear data sets (Hagan et al., 1996). As they are flexible and less assumption-dependent, there is no need to define the underlying physical process between the inputs and outputs (García et al., 2011). In the tropospheric ozone formation, the relationships between photochemical and meteorological variables involved are complex and non-linear (Gardner and Dorling, 1999; Jenkin and Clemitshaw, 2000; García et al., 2011), for this reason the ANN is more preferred to MLR in predicting the ozone concentration (Yi and Prybutok, 1996; Chaloulakou et al., 2003).

There have been numerous studies about the modeling of ozone concentration using ANN models in different locations and seasons (Biancofiore et al., 2015 and reference therein). The model architecture or input parameters were not unique either (García et al., 2011; Luna et al., 2014; Biancofiore et al., 2015; Taylan, 2018). The time resolution was either day or hour; and the output variable was either daily maximum values or average values. The commonly used photochemical parameters as inputs of ANN models include but not limited to NO, NO_2, total hydrocarbon (THC) and non-methane hydrocarbon (NMHC) (Yi and Prybutok, 1996; Heo and Kim, 2004; García et al., 2011; Luna et al., 2014). Although adding more photochemical parameters might increase the capability of ANN in predicting the ozone level (Biancofiore et al., 2015), the cost of obtaining such dataset is not less than that obtaining the ozone concentration directly in practice. On the contrary, meteorological parameters are routinely monitored in many urban areas so that the data availability is not a problem especially in developing countries. Based on the parsimony principle, it is tempted to develop an ANN model to predict ozone concentration using only a few of meteorological parameters.

ANN development is stochastic in nature, and no identical results can be reproduced on different occasions unless carefully devised (Elshorbagy et al., 2010). Therefore, it is necessary to analyze prediction uncertainty and identify the contribution of individual sources of uncertainty to total prediction uncertainty (Wagener and Gupta, 2005). Generally, ANN prediction uncertainty stems from two aspects, the uncertainty in ANN structures as well as the uncertainty in ANN inputs, weights and biases (Chitsazan et al., 2015). The Monte Carlo Simulation (MCS) technique is a widely used method for prediction uncertainty analysis in ANN modeling and it allows the quantification of the model prediction uncertainty (Shrestha et al., 2009; Kasiviswanathan et al., 2016). In addition, MCS technique can also be used for sensitivity analysis to determine how much "sensitive" is the ANN output to the changes in the value of the input parameters (Nourani and Fard, 2012). It is worth to note that results of uncertainty analysis and sensitivity analysis about ANN modeling of ambient ozone level were less reported in previous literatures.

The objective of this study is to develop a parsimonious ANN model for ambient ozone as a function of meteorological parameters and other temporal covariate as predictors. The dataset including ozone, NO_2, and other meteorological parameters, which was collected in Jinan (a metropolis in Northern China), was used for training, validating and testing the ANN model. The performance of the created ANN models was analyzed using multiple statistical metrics; and the MCS technique was applied for uncertainty and sensitivity analysis.

2. Materials and methods

2.1. Study area and datasets

In this paper, Jinan, the capital city of Shandong province in Northern China, was selected as the study area. Jinan has a humid continental climate with four distinctive seasons (dry and nearly rainless in spring, hot and rainy in summer, crisp in autumn and dry and cold in winter). The average annual temperature is 14.70 °C, and the annual precipitation is around slightly above 670 mm. January is the coldest and driest month, with a mean temperature of −0.4 °C and 5.7 mm of equivalent rainfall. July is the hottest and wettest month; and the mean temperature and precipitation are 27.5 °C and 201 mm, respectively. Due to the mountains to the south of the city, temperature inversions are common, occurring on about 200 days per year. The urban population in Jinan is about 4.69 million. Like several northern cities in China, Jinan also faces the problem of severe air pollution, especially in winter and spring.

The air quality and meteorological data was provided by Ministry of Environmental Protection (MEP) and Meteorological Administration (CMA) of China, respectively. In this study, we used the meteorological dataset collected at the only national meteorological station affiliated to CMA (N36°36′, E117°03′) in the urban area of Jinan. The meteorological dataset can be accessed from the Climate Data Center of CMA (CDC-CMA, 2017). The original dataset at this station includes daily observations of maximum/average/minimum temperature, maximum/average/minimum atmospheric pressure, average/maximum wind speed, wind direction, relative humidity and sunshine duration. At the beginning of 2013, MEP of China gradually published the hourly air quality data (PM2.5, PM10, NO_2, SO_2, O_3, CO) to the public. We collected the dataset of air quality at the nearest air quality monitoring site (N36°37′, E116°59′) from the meteorological station. The time range of air quality data was from 19 January 2013 to 31 October 2017. The air quality data can be accessed from the data center of MEP of China (DC-MEP, 2017). Fig. 1 shows the geographic locations of the meteorological station and the air quality monitoring site. The urban area of Jinan approximately ranges from E116°53′ to E117°12′ and N36°34′ to N36°41'.

Previous studies reveal that ambient ozone concentration gradually increases in the morning, because solar radiation promotes the formation of photochemical oxidants (García et al., 2011). When

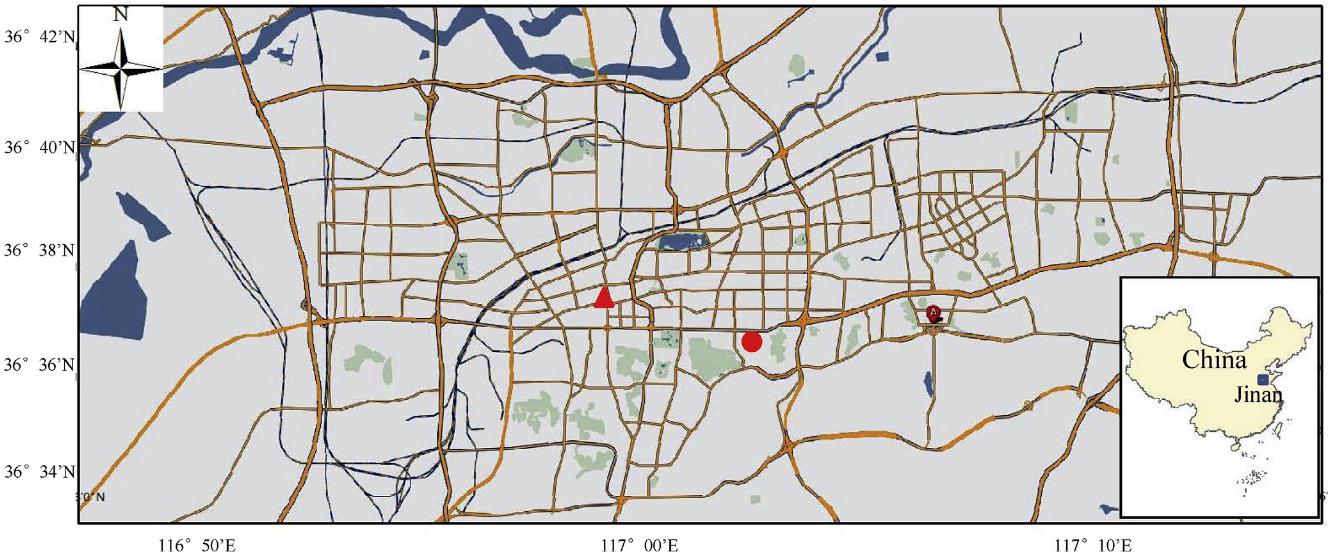


Fig. 1. Geographic locations of the meteorological and air quality monitoring stations. Circle represents the meteorological station affiliated to Meteorological Administration (CMA) of China, while triangle represents the air quality monitoring station affiliated to Ministry of Environmental Protection (MEP) of China.

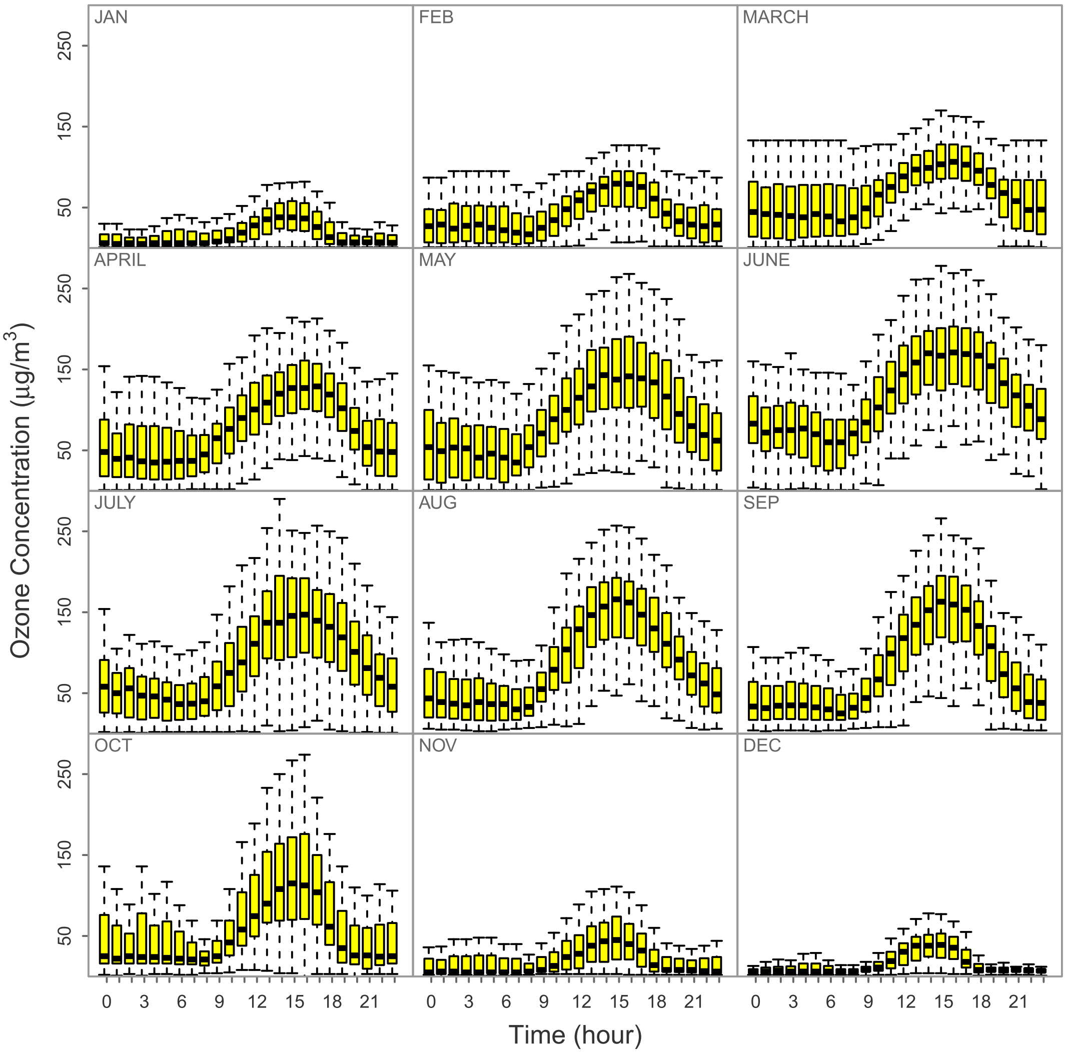


Fig. 2. Diurnal variations of ozone concentration in each month within a year. The bottom and top of the box indicate the first and third quartiles, and the band inside the box indicates the second quartile (the median). The ends of the whiskers represent the minimum and maximum values.

concentrations of precursors in the atmosphere are lowered, the formation of ozone stops and its concentration decrease as the day progresses. Hence, ambient ozone concentrations exhibit significant diurnal variation in urban areas (Fig. 2). From Fig. 2, the ozone pollution level is more serious in the daytime. In this study, only the ozone concentrations between 9:00 a.m. to 6:00 p.m. are used, because ambient ozone concentration in the daytime could properly reflect the ambient ozone pollution level. Days with more than three missing values of hourly ozone concentrations were considered to be unavailable. For the purpose of model development, only records with both pollution and meteorological observations were taken in to account. The final dataset with complete meteorological and air quality data covers 1658 days in the study period.

In this study, the output of ANN model was the average ozone concentration in daytime (9:00 a.m. - 6:00 p.m.). Not all routinely monitored meteorological parameters were used as input parameters of ANN model, because our objective was to develop a parsimonious ANN model. Based on previous studies, 7 meteorological parameters including precipitation (PRE), maximum barometric pressure (MaxPRS), relative humidity (HUM), sunshine duration (SD), maximum temperature (MaxTemp), maximum wind speed (MaxWind), and wind direction (WD) are selected as the potential predictors. Since vehicular source emissions were the major source of ozone precursors, the data obtained on the weekends and holidays need to be treated separately. Fig. 3 showed the probability distribution of daily average ozone concentration on working days, weekends, and China's legal holidays in May, June, September and October (there was not legal holidays in July and August). From Fig. 3, we found that ozone level was the higher on regular weekends but the lower on working days. This surprising phenomenon was mainly due to difference of road traffic on different days. Most shopping and business centers are located in the urban areas of Jinan; therefore, traffic jam is more serious on weekends and holidays than that on working days resulting in more emission of some primary pollutants in urban areas. On legal holidays, a part of private vehicles left from the urban areas for leisure travel, traffic jam was not as serious compared with regular weekends. On working days, public transportation was more preferred by citizens resulting in less emission of air pollutants. To present this difference, we added a temporal covariate, the category of day (CD), where 1, 2 and 3 represent working day, legal holiday, and regular weekend, respectively. In addition, data of one photochemical parameter (NO_2) was also collected for the purpose of model comparison.

2.2. Artificial neural network

Artificial neural network (ANN) model is an essentially simple mathematical model defining a function $F: X \rightarrow Y$, where the nonlinear relationships between variables in inputs X and variables in output Y can be determined (Antanasijević et al., 2013). Before developing a ANN model, network topology, neuron characteristics, and training or learning rules with inputs, output(s) and hidden layers with interconnections should be specified (Mehrotra et al., 2000). Generally, a 3-layer ANN (input, hidden and output) model is capable to produce acceptable performance in predicting the ozone concertation (Yi and Prybutok, 1996; Prybutok et al., 2000; García et al., 2011; Luna et al., 2014; Biancofiore et al., 2015). In this study, the layer number of ANN model was also set to be 3. Neuron is the fundamental processing unit that computes a weighted sum of its input signals and then applies a nonlinear activation function to produce an output signals (García et al., 2011). Neurons can use any differentiable transfer function to generate their output, and logarithmic sigmoid (log-sigmoid) and tangent sigmoid (tan-sigmoid) are two most commonly used as hidden transfer functions while the linear transfer function is more applied as output transfer function (Deo and Sahin, 2015). The number of hidden layer neurons is another ANN parameter needed to be determined, although there is no rule for the optimum number of neurons in the hidden layer (Yetilmezsoy and Demirel, 2008). ANNs are also sensitive to the number of neurons in their hidden layers. Too few neurons can

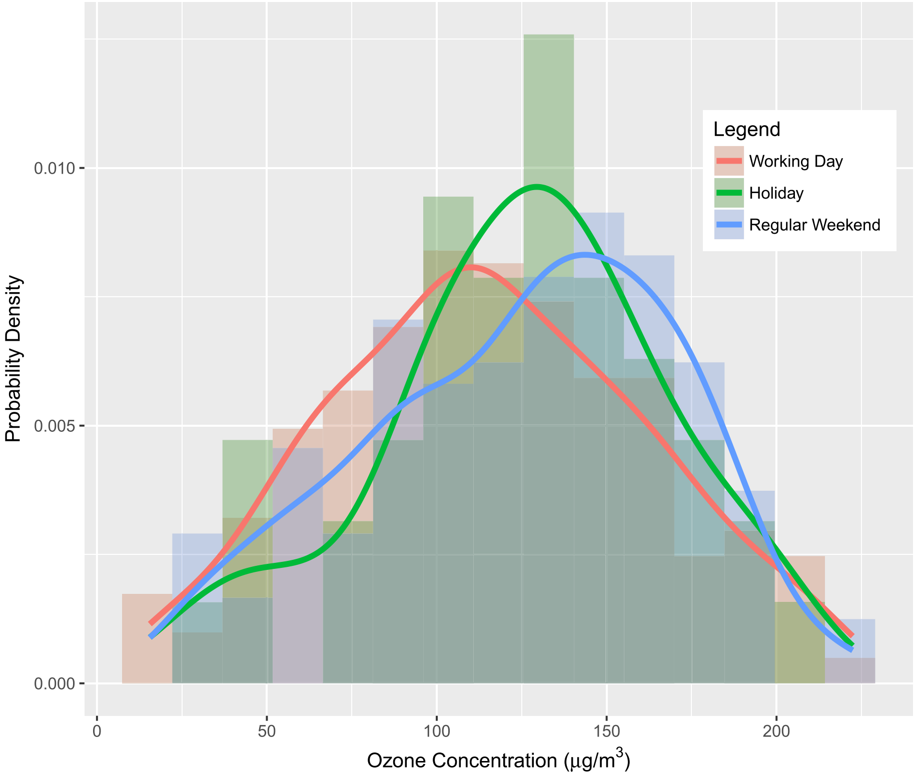


Fig. 3. Probability distribution of daytime average of ozone concentration on working days, legal holidays, and regular weekends in May, June, September, and October (2013–2016).

Table 1
Candidate ANN models with different network architecture and the model performance in the testing phase.

Model ID	Hidden transfer function*	Output transfer function	Network structure§	*IA*	*RMSE*	*MAE*	R^2
M1	Tansig	Linear	8–10–1	0.9414	23.1441	17.1356	0.8119
M2	Tansig	Linear	8–20–1	0.9425	23.0558	17.1574	0.8163
M3	Tansig	Linear	8–30–1	0.9439	22.2102	17.6091	0.8226
M4	Logsig	Linear	8–10–1	0.9393	23.3738	17.3657	0.8203
M5	Logsig	Linear	8–20–1	0.9386	23.4334	17.4500	0.8217
M6	Logsig	Linear	8–30–1	0.9397	23.3997	17.4225	0.8210

*Tansig = tangent sigmoid, Logsig = logarithmic sigmoid.
§Number of input parameters–number of hidden neurons–number of output.

Table 2
Candidate ANN models with different combinations of input parameters in forward selection procedure and the model performance in the testing phase. The last row is the benchmarking ANN model for input selection.

Input subset	*IA*	*RMSE*	*MAE*	R^2
MaxTemp	0.8828	30.2980	23.8120	0.6396
MaxTemp, MaxPRS	0.9057	27.9530	22.0880	0.7025
MaxTemp, MaxPRS, MaxWind	0.9090	27.9190	21.9900	0.7105
MaxTemp, MaxPRS, MaxWind, SD	0.9280	24.8600	19.3070	0.7575
MaxTemp, MaxPRS, MaxWind, SD, HUM	0.9437	22.4740	17.4830	0.7963
MaxTemp, MaxPRS, MaxWind, SD, HUM, PRE	0.9428	22.8570	17.6980	0.8008
MaxTemp, MaxPRS, MaxWind, SD, HUM, PRE, WD	0.9428	22.6400	17.6070	0.8050
MaxTemp, MaxPRS, MaxWind, SD, HUM, PRE, WD, CD	0.9439	22.2102	17.6091	0.8226
MaxTemp, MaxPRS, MaxWind, SD, HUM, PRE, WD, CD, NO_2[a]	0.9451	21.9290	17.5960	0.8429

[a] The correlation between NO_2 and O_3 was not ranked.

lead to underfitting. Too many neurons can contribute to overfitting, in which all training points are well fitted, but the fitting curve oscillates wildly between these points. As long as training methods, 2 s-order training methods, primarily the Levenberg–Marquardt (LM) or the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton back-propagation learning algorithms were considered to be computationally efficient (Marquardt, 1963; Dennis and Schnabel, 1983). In this study, the LM algorithm was chosen as the default one due to its rapid convergence (Yetilmezsoy and Demirel, 2008; Deo and Sahin, 2015).

The number of neuron in hidden layer and the transfer and output functions of the network were usually decided by trial and error (Deo and Sahin, 2015). By setting different neuron characteristics and training algorithms, a few of candidate ANN models can be developed and evaluated to select the optimum one. The ANN models were developed under MATLAB environment running under Intel 4-core i7, 3.4 GHz CPU. After training and validating the networks, the meteorological parameters and covariate in the testing phase are used to predict the ozone concentrations, and then the predictions are compared with the observed values of the ozone concentrations.

2.3. Evaluation of model performance

The performance of ANN model was evaluated using four metrics including: (1) Willmott's Index of Agreement (*IA*) (Willmott, 1982); (2) Root-Mean Square Error (*RMSE*), (3) Mean Absolute Error (*MAE*), (4) Coefficient of Determination (R^2). *IA* represents the ratio between the mean square error and the potential error. *RMSE* and *MAE* measure residual errors and they are valuable to the model as they indicate the error in the output units. R^2 equals the square of the Pearson correlation coefficient between predicted and observed values of ozone concentrations in linear least squares regression (Antanasijević et al., 2014). A high R^2 implies a good model performance, and vice-versa. These four metrics are sufficient for evaluation of the ANN model's performance and used herein for selection of the best architectures or combination of input parameters of the networks.

Moreover, the ANN model could also be used in environmental warning system. Therefore, the capability of ANN model in predicting air pollution episodes is also one very important metric of model performance (Noori et al., 2010). According to the China's ambient air quality standards, threshold limit value for 8-h average of ambient ozone concentration in urban area is 160 $\mu g/m^3$ (MEP, 2012). In this study, the average ozone concentration in daytime above 160 $\mu g/m^3$ is selected as a representative value for episodes, and the final ANN model will be further evaluated during pollution episodes. We define warning success ratio as:

$$WSR = \frac{count(Observation > 160,\ Prediction > 160)}{count(Observation > 160)} \quad (1)$$

WSR represents the warning ability in environmental warming system.

2.4. Inputs selection

In this study, the dataset consists of 9 potential input variables including 7 meteorological parameters, one temporal covariate, and one photochemical parameter (NO_2). The objective of this study is to explore the feasibility of predicting ozone concentration using meteorological observations and temporal covariate. We firstly create an ANN model using all 9 input variables, and it will be used for benchmarking the performance of models with different combinations of input parameters. The 7 meteorological parameters and one temporal covariate are then selected using forward selection (FS) technique (Khan et al., 2007; Noori et al., 2010; Dehghani et al., 2013). FS begins with ordering all potential input variables according to their correlation with the output variable (from the most to the least correlated variable). Then, the most correlated variable is chosen as the first input variable, and the remaining input variables are added sequentially. In each trail, the model performance in the testing phase is evaluated by using the four metrics, *IA*, *RMSE*, *MAE*, and R^2.

2.5. Uncertainty analysis and sensitivity analysis

Once the architecture, neuron characteristics, and training algorithm of the optimum ANN model were determined, Monte-Carlo simulations (MCS) were further conducted to examine the uncertainty of the final ANN model in predicting daily ozone concentration. MCS involves the repeated generation of random parameters from their probability distributions, and then computing the statistics of the output (Antanasijević et al., 2014). The first step is to determine the probability density functions (PDFs) of the input variables (e.g. Gaussian, log-normal, Weibul etc.). Kolmogorov–Smirnov test is used to compare the probability distribution of input variables with a reference probability distribution with a 5% significance level (Dehghani et al., 2013; Antanasijević et al., 2014). With these fitted probability distributions, we randomly resample the input dataset without replacement for 1000 times, keeping the ratios of training, validation and testing sets unchanged. Unrealistic samples exceeding the maximum or

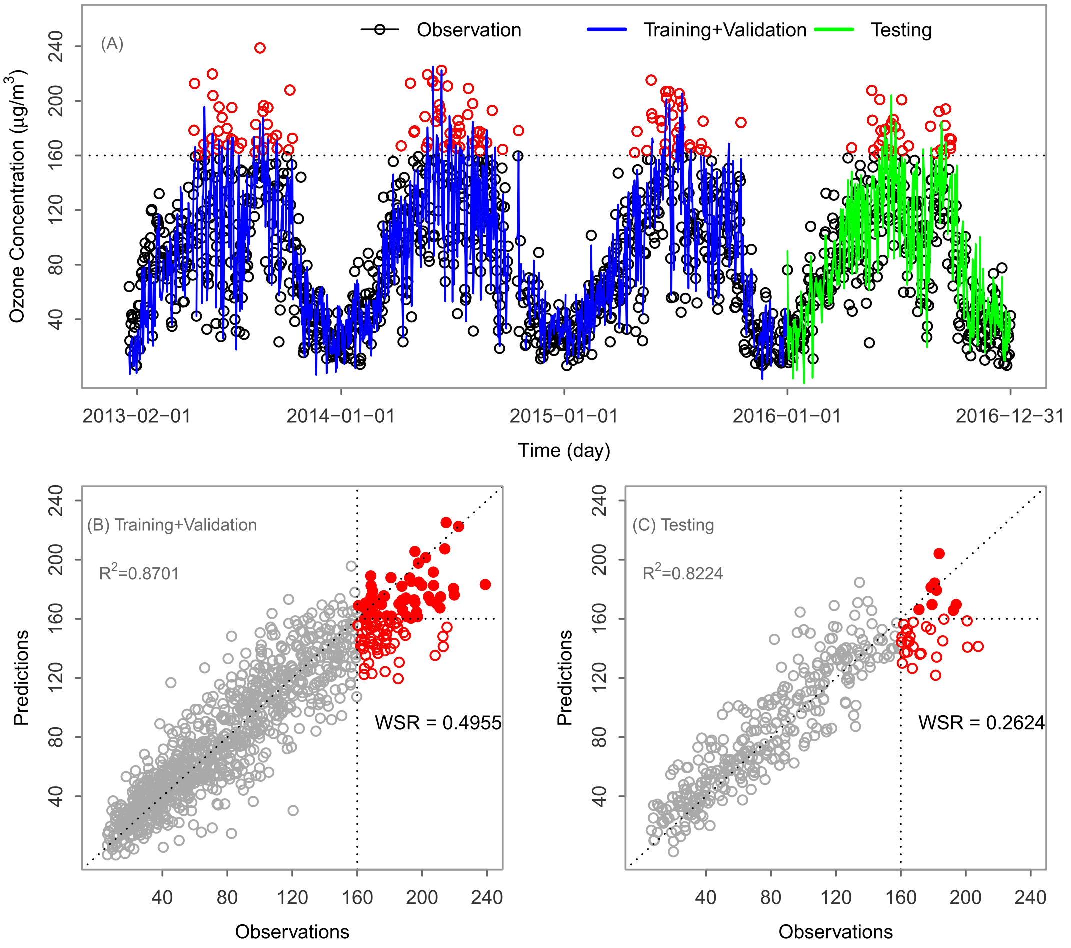


Fig. 4. (A) Predicted and observed ozone concentrations in training, validation, and testing phases. The dashed line indicates the threshold limit value of ambient ozone concentration in urban area (160 $\mu g/m^3$). (B) Scatterplots of predicted vs. observed ozone concentrations in training and validation phases (9 January 2013 to 31 December 2015). Observed ozone concentration larger than 160 $\mu g/m^3$ marked using red circles was considered as pollution episodes, while the filled circles represent successful warning ozone pollution. *WSR* represents warning success ratio defined by Eq. (1). (C) Scatterplots of predicted vs. observed ozone concentrations in testing phase (1 January 2016 to 31 December 2016). The interpretation of circles are the same as that in (B). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

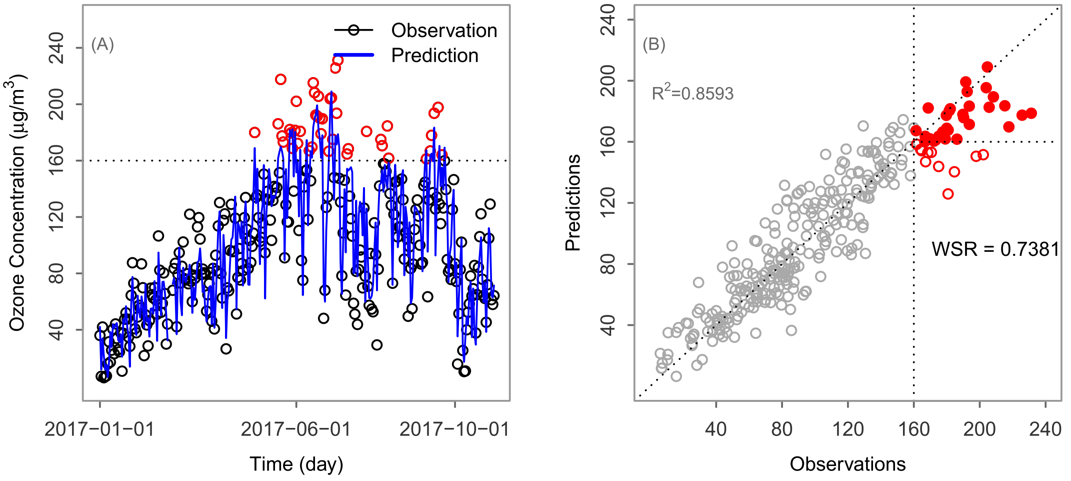


Fig. 5. (A) Predicted and observed ozone concentrations in the extended testing phase (from 1 January 2017 to 31 October 2017). (B) Scatterplots of predicted vs. observed ozone concentrations in the extended testing phase.

minimum limits should be excluded (Antanasijević et al., 2014). Then, the final ANN model generates 1000 time series of ozone concentration. At each day, the 95% confidence intervals are determined by finding the 2.5th and 97.5th percentiles of the constructed distribution. This 95% confidence interval provides more information than other statistics about the range of predictions associated with the optimum ANN model (Noori et al., 2009, 2010). The wider the interval, the smaller is the accuracy of the forecast and vice versa. The ratio of observed values

Table 3
Meteorological parameters and temporal inputs and their fitted probability distribution function (PDF) with the input ranges and obtained O_3 ranges using the optimum ANN model (Monte Carlo sensitivity analysis).

	Input	Unit	PDF	Kolmogorov-Smirnov test		ΔI	ΔO_3
				Stat.	Sig.(p)		
1	MaxTemp	°C	Gaussian mixture	0.0393	0.0372	47	165.05
2	MaxPRS	hPa	Lognormal	0.0081	< 0.001	45	59.43
3	SD	hour	Weibull mixture	0.1465	< 0.001	13.3	36.33
4	MaxWind	m/s	GEV	0.0330	0.0126	12.2	33.69
5	HUM	%	Beta	0.0902	< 0.001	87	22.75
6	CD	–	–	–	–	3	16.42
7	PRE	mm	Exponential	0.0675	0.096	127.1	15.17

that lie within the 95% confidence interval relative to all datasets is also calculated as the robustness metric of the final ANN model (Noori et al., 2009, 2010).

Also with the final ANN model, the sensitivity of the ozone concentration to the meteorological and temporal inputs is investigated. Sensitivity analysis is implemented for all input variables respectively (Antanasijević et al., 2014). For each input variable, a dataset consisting of 1000 samples of inputs are firstly generated, where the values of this input variable are sampled from its fitted probability distribution while the other input variables are set to be constant with the mean values. Finally, the range of predicted ozone concentration is computed to quantify the influence of each input variable on the output variable.

3. Results and discussions

3.1. Model development and input selection

First, the network architecture was determined by evaluating the model performance in the testing phase. In this study, a few of candidate ANN models with different number of hidden neurons or hidden transfer function and 7 meteorological parameters and one temporal covariate as input variables are developed (e.g. Şahin, 2012; Deo and Sahin, 2015). Specifically, the hidden transfer function was either “log-sigmoid” or “tan-sigmoid”, while output transfer function was always “linear”. The number of hidden neurons might be 10, 20, and 30, respectively. This resulted in a total of 6 ANN models (Table 1). The main dataset from 9 January 2013 to 31 December 2016 (1461 days) was used for model development. More specifically, dataset from 9 January 2013 to 31 December 2015 was used for training (used for ANN model training) and validation (extracted from the training dataset and used in training process to prevent ANN overtraining and to enable better generalization of the ANN model on new data). Other dataset in 2016 was used as the testing dataset (used to evaluate ANN model generalization after the training process). The optimum ANN model was selected based on the values of the 4 statistical metrics (IA, $RMSE$, MAE, and R^2) calculated in the testing phase. Table 1 summarized the model performance. We found that these six candidate ANN models performed almost equally well. The training time was not too long either (less than 50 s). Consequently, the specifications of ANN mode with “tan-sigmoid” equation as hidden transfer function and 30 hidden neurons were adapted due to its better performance.

Next, the performances of ANN model with different combinations of input parameters were also evaluated. The order of correlations between the output variable and the 8 input variables from highest to lowest were MaxTemp, MaxPRS, MaxWind, SD, HUM, PRE, WD and CD. The lowest correlation between output variable and WD or CD could be explained by the discrete essence of WD (16 states: $1,2,\cdots,16$) and CD (3 states: $1,2,3$). Therefore, the variable with the highest correlation (i.e. MaxTemp) was selected as the first and most important variable. Table 2 presented the prediction performance of ANN model in testing phase with different combinations of input variables in the forward selection procedure. The performance of the benchmarking model with all 9 input variables was also shown in the last row of Table 2. The combination of model input variables was also selected based on the values of the 4 metrics. From Table 2, we found that the input variable, WD, contributed little to improving the model performance; therefore, WD was eliminated in the following analysis. It is worth to note that adding CD significantly improve the model output, although CD was not highly correlated with the model output. This finding verified the reasonability of adding the temporal covariate (CD) in ANN modeling of ozone. Although CD is called as temporal covariate, it actually represents the vehicular source emissions on different categories of days as explained in previous section.

In addition, we also found that adding the only photochemical parameter (NO_2) could improve the model output. Specifically, R^2 could be improved from 0.8226 to 0.8429 in the testing period. Since our objective is to explore the feasibility of predicting ozone concentration solely using meteorological observations, we noted that the model performance of ANN model with only meteorological parameters and temporal covariate as input variables were acceptable. Thus, the final ANN model had 6 meteorological parameters (MaxTemp, MaxPRS, MaxWind, SD, HUM, PRE) and one time covariate (CD). The following analyses were confined to this final ANN model in the next two subsections.

3.2. Model performance

We firstly illustrated the results of ANN modeling in Fig. 4. In the training and validation phase (9 January 2013 to 31 December 2015), R^2 was equal to 0.8701, while in the testing phase (2016), R^2 became 0.8224. There was a very good agreement between the predicted and the observed ozone concentration. In the testing phase, $IA = 0.9415$, $RMSE = 22.2157$, $MAE = 17.6010$. The warning success ratios in the two phases were 0.4955 and 0.2624, respectively. In addition, we further extended the testing period to 2017 (from 1 January 2017 to 31 October 2017), and the prediction results were shown in Fig. 5. The ANN model performed better in 2017 than in 2016. R^2 became 0.8593, and WSR became 0.7381. The prediction capability of the final ANN model in 2017 was satisfactory.

3.3. Uncertainty analysis

Uncertainty analysis of the predicted daily ozone concentration during the whole study period (from 19 January 2013 to 31 October 2017) has been quantified by estimating the confidence intervals of the simulation results. The 6 meteorological parameters used in the final ANN model were fitted to pre-assumed probability distribution functions (PDFs), respectively. The other input variable, CD, followed a discrete probability distribution. Table 3 summarized the results of Kolmogorov–Smirnov test, and Fig. 6 presented the empirical and fitted PDFs. Fig. 7 showed the 95% confidence intervals for the estimates of zone concentration. In total, 81% (1341) of all 1658 observations fall within the 95% confidence intervals. Also from Fig. 7, we found that lower extremes were located within the 95% confidence interval, while many higher extremes were beyond the upper bound. That means the current ANN model has limitation to predict extreme high ozone concentration using the meteorological parameters and temporal covariate as predictors.

3.4. Sensitivity analysis

The final ANN model for sensitivity analysis was trained using the main dataset of 1461 observations. Then, 7 blocks of 1000 input vectors were generated. In each block, one input was randomly generated

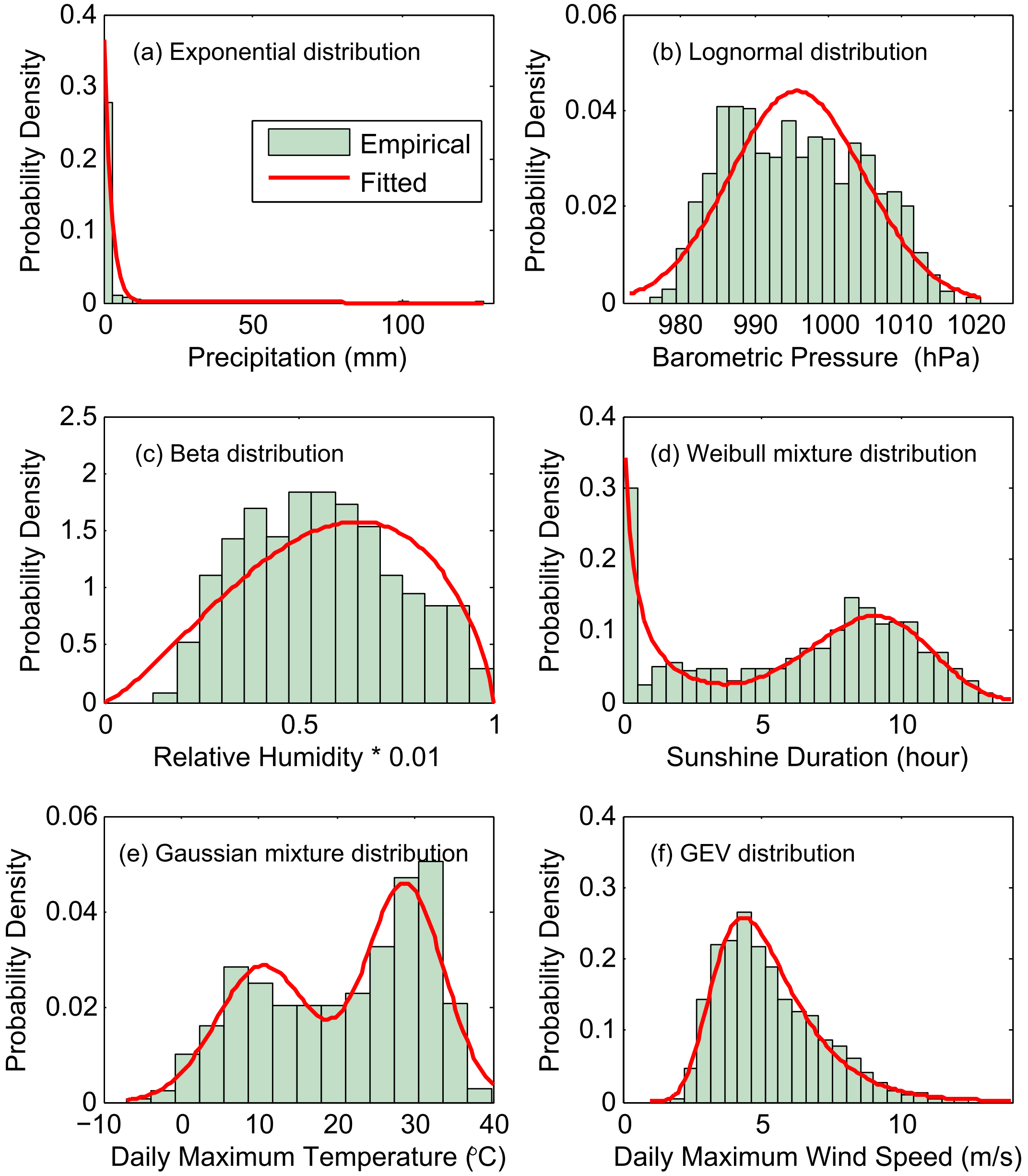


Fig. 6. Empirical and fitted probability density functions for 6 continuous meteorological parameters. (a) Precipitation: Exponential distribution, $f(x) = \frac{1}{\mu}e^{\frac{-x}{\mu}}$, $\hat{\mu} = 2.1828$; (b) Maximum atmospheric pressure: Lognormal distribution, $f(x) = \frac{1}{x\sigma\sqrt{2\pi}}e^{\frac{-(lnx-\mu)^2}{2\sigma^2}}$, $\hat{\mu} = 6.9035$, $\hat{\sigma} = 0.0091$; (c) Relative humanity: Beta distribution, $f(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$, $0 \le x \le 1$, $B(a, b)$ is the Beta function, $\hat{a} = 2.3787$, $\hat{b} = 1.7333$; (d) Sunshine duration: Weibull mixture distribution, $f(x) = \theta\frac{b_1}{a_1}\left(\frac{x}{a_1}\right)^{b_1-1} e^{-\left(\frac{x}{a_1}\right)^{b_1}} + (1-\theta)\frac{b_2}{a_2}\left(\frac{x}{a_2}\right)^{b_2-1} e^{-\left(\frac{x}{a_2}\right)^{b_2}}$, $\hat{a}_1 = 1.1698$, $\hat{b}_1 = 0.69$, $\hat{a}_2 = 9.5498$, $\hat{b}_2 = 4.7777$, $\hat{\theta} = 0.3686$; (e) Maximum temperature: Gaussian mixture distribution, $f(x) = \theta\frac{1}{x\sigma_1\sqrt{2\pi}}e^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}} + (1-\theta)\frac{1}{x\sigma_2\sqrt{2\pi}}e^{\frac{-(x-\mu_2)^2}{2\sigma_2^2}}$, $\hat{\mu}_1 = 10.5262$, $\hat{\sigma}_1 = 6.0001$, $\hat{\mu}_2 = 28.7390$, $\hat{\sigma}_2 = 5.0213$, $\hat{\theta} = 0.4291$; (f) Maximum wind speed: Generalized extreme value

according to its continuous (or discrete) probability distribution, while other inputs had measured mean (or mode) values. Maximum and minimum limits on each input had been adopted to prevent unrealistic selection of extreme values (Antanasijević et al., 2014). The input ranges and obtained ozone ranges were presented in Table 3. From Table 3, we found that the order of inputs' influences on ozone concentration from highest to lowest were maximum temperature, maximum atmospheric pressure, sunshine duration, maximum wind speed, relative humanity, temporal covariate category of day, and precipitation.

Moreover, the influence of maximum temperature, atmospheric pressure, sunshine duration and maximum wind speed were further closely examined by splitting their ranges to smaller intervals. The obtained ozone concentrations for each interval were shown in Fig. 8. With the increasing of maximum temperature, the predicted ozone concentration increased accordingly, and the output ozone concentration was more sensitive to high values of maximum temperature (Fig. 8a). The basic reasoning is that photochemical reaction rates are sensitive to temperature, so that increasing the temperature in the troposphere stimulates a series of interlinked reactions that contribute to ozone formation (García et al., 2011). Fig. 8b showed that the predicted ozone concentration decreased with the increase of the

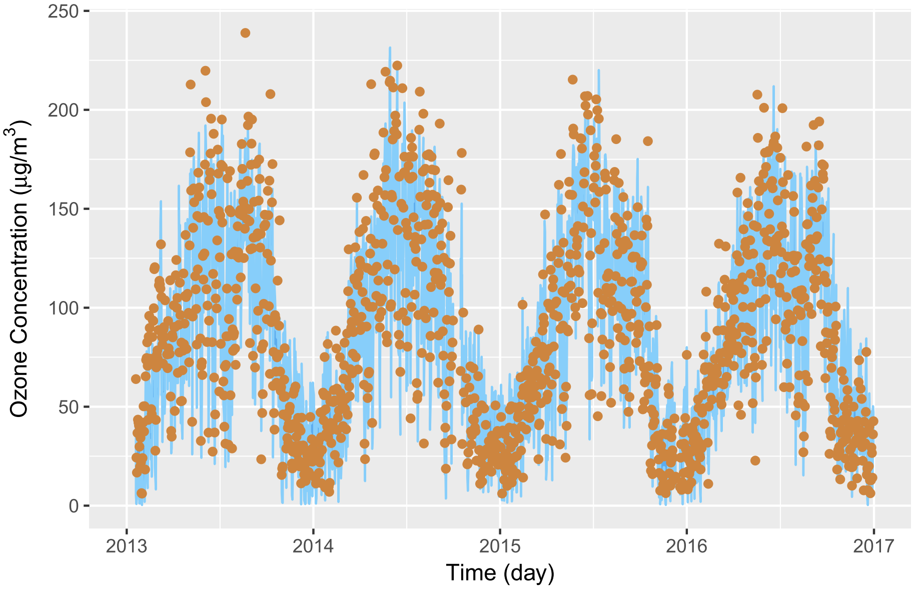


Fig. 7. Observations of ozone concentration and the 95% confidence intervals estimated by Monte Carlo simulation with randomly sampled input vectors.

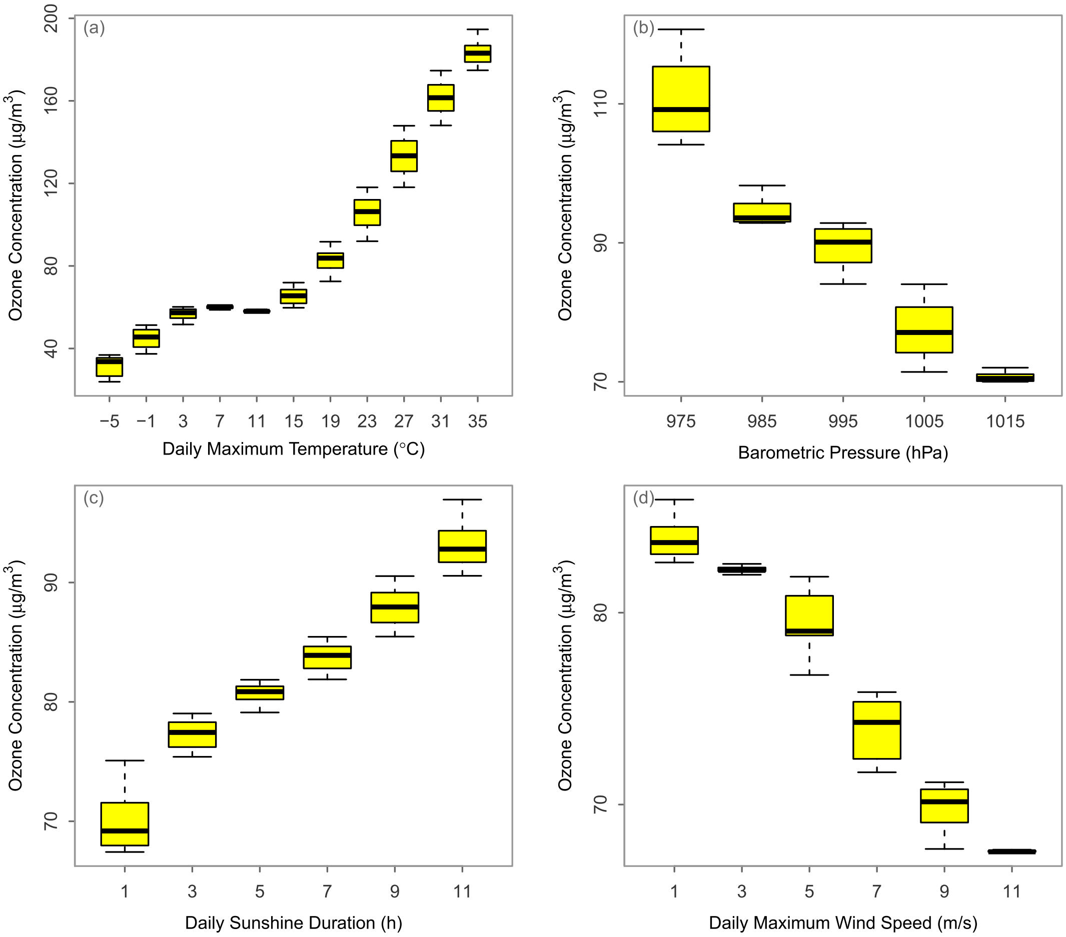


Fig. 8. Sensitivity analysis of predicted ozone concentration on 4 predominate input variables. For each input variable, 1000 samples was generated using the fitted probability distribution functions and other input variables were assumed to be constant to predict the ozone concentration. The generated samples of the four meteorological parameters were split to smaller intervals, and the predicted ozone concentrations for each interval were summarized as boxplots. The interpretation of boxplots was the same as that in Fig. 1. (a) daily maximum temperature, (b) daily maximum atmospheric pressure, (c) daily sunshine duration, (d) daily maximum wins speed.

maximum atmospheric pressure. This phenomenon could be explained by the seasonal variations of ozone pollution level and atmospheric pressure. In winter, the atmospheric pressure is higher but the ozone level is lower; while in summer the atmospheric pressure is relatively lower but the ozone level is higher. Fig. 8c showed that a growth of the daily sunshine duration resulted in a significant increase of predicted ozone concentration. Solar radiation has the great effect on photochemical reactions, i.e., it is involved in the formation and destruction of the various compounds involved in the increase of tropospheric ozone (Sun et al., 2013). In this study, we have no data of solar radiation in the original meteorological dataset; instead, the daily sunshine duration released by CMA have been used to represent solar radiation. Fig. 8d showed that a growth of the wind speed resulted in a sharp decrease of ozone concentration, because winds are responsible for the dispersion of air pollutants (in this case the ambient ozone and its precursors). If wind speeds are high, the pollutants tend to disperse quickly. The results of sensitivity analysis were consistent with the findings revealed in some previous literatures (García et al., 2011; Luna et al., 2014).

In Table 2, sensitivity analysis showed that relative humidity ranked fifth and had obvious influence on ozone concentration. In Jinan, the most humid days are mainly distributed in late July and early August. This is also the rainy season and solar radiation is not so strong due to cloud and aerosol in tropospheric atmosphere. The climate in most part of China is strongly influenced by the East Asian monsoon (Gao et al., 2016). Strong winds usually occurs in spring and autumn, northwest and southeast are two major wind directions. Therefore, the predicted ozone concentration was not so sensitive to wind direction. Sensitivity analysis also revealed the difference of predicted ozone concentrations on different types of day (Table 2). This difference was consistent with that showed in Fig. 3. The least sensitive parameter in our ANN model was precipitation. We noted that the real relationship between precipitation and ozone concentration was not properly reflected in our ANN model, because the output ozone concentration was the average value in daytime while the precipitation was the summation within 24 h.

4. Conclusions

In this study, we investigated the feasibility of using ANN model with meteorological parameters as input variables to predict ozone concentration in the urban area of Jinan, China. Before creating the ANN model, the hourly ozone concentration data was statistically analyzed. It was found that the ambient ozone concentration exhibited significant diurnal and seasonal variations. Therefore, the average of ozone concentration in daytime was used to represent ozone pollution level. Moreover, we found that the probability distributions of ozone concentration on working days, regular weekends, and holidays were different. In this study, the category of day was also used as a potential input variable of the ANN model.

The performance of ANN model was evaluated using four statistical metrics in the testing period. Primarily, we found that the architecture of network of neurons had little effect on the predicting capability of ANN model. Then the input variables were selected using forward selection procedure, and wind direction was eliminated in the final ANN model. When the temporal covariate (CD) was added, the model performance was also obviously improved, although the temporal covariate was not highly correlated with the output variable. Compared with the benchmarking ANN model with all meteorological and photochemical parameters as input variables, the predicting capability of the final ANN model with the 6 meteorological parameters and one temporal covariate as input variables was acceptable. Its predicting capability was also verified in term of warming success ratio during the pollution episodes.

Monte-Carlo simulations were conducted to examine the uncertainty of the final ANN model in predicting daily ozone concentration. The uncertainty analysis showed that the ANN model could properly predict the ozone level, while a few of the observed extreme high values fell outside of the 95% confidence interval. Furthermore, the Monte Carlo simulation technique was also used to investigate the sensitivity of the output ozone concentration to the meteorological and temporal input variables. Maximum temperature, atmospheric pressure, sunshine duration and maximum wind speed were identified as the predominate input variables that significantly influence the range of predicted ozone concentration. The importance of this study is that we have explored the feasibility of using ANN model to predict ambient ozone concentration using a few of meteorological parameters and as predictors. This approach is very useful especially in developing countries where atmospheric chemistry data are sparse.

Acknowledgments

This work was partly supported by the Youth Innovation Promotion Association of CAS (2016195), CAS Knowledge Innovation Project (KZCX2-EW-QN209), and National Natural Science Foundation of China (31570423). The helpful comments from the two anonymous reviewers were also acknowledged.

References

Antanasijević, D., Pocajt, V., Perić-Grujić, A., Ristić, M., 2014. Modelling of dissolved oxygen in the Danube River using artificial neural networks and Monte Carlo Simulation uncertainty analysis. J. Hydrol 519, 1895–1907.

Antanasijević, D., Pocajt, V., Povrenović, D., Ristić, M., Perić-Grujić, A., 2013. PM10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci. Total Environ. 443, 511–519.

Baur, D., Saisana, M., Schulze, N., 2004. Modelling the effects of meteorological variables on ozone concentration—a quantile regression approach. Atmos. Environ. 38, 4689–4699.

Biancofiore, F., Verdecchia, M., Di, C.P., Tomassetti, B., Aruffo, E., Busilacchio, M., Bianco, S., Tommaso, S.D., Conlangeli, C., 2015. Analysis of surface ozone using a recurrent neural network. Sci. Total Environ. 514, 379–387.

Brasseur, G., Hauglustaine, D., Walters, S., Rasch, R., Müller, J.F., Granier, C., Tie, X., 1998. MOZART, a global chemical transport model for ozone and related chemical tracers 1. Model description. J. Geophys. Res. 103, 28265–28289.

Chaloulakou, A., Saisana, M., Spyrellis, N., 2003. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Sci. Total Environ. 313, 1–13.

Chitsazan, N., Nadiri, A.A., Tsai, T.C., 2015. Prediction and structural uncertainty analyses of artificial neural networks using hierarchical Bayesian model averaging. J. Hydrol 528, 52–62.

CDC-CMA, 2017. Daily Datasets of Ground Meteorological Parameters in China. http://data.cma.cn/site/index.html.

DC-MEP, 2017. Daily Datasets of Air Quality in China. http://datacenter.mep.gov.cn/index.

Dehghani, M., Saghafian, B., Nasiri Saleh, F., Farokhnia, A., Noori, R., 2013. Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. Int. J. Climatol. 34, 1169–1180.

Dennis, J.E., Schnabel, R.B., 1983. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, Englewood Cliffs, NJ.

Deo, R.C., Sahin, M., 2015. Application of the Artificial Neural Network model for prediction of monthly Standardized Precipitation and Evapotranspiration Index using hydrometeorological parameters and climate indices in eastern Australia. Atmos. Res. 161–162, 65–81.

Elshorbagy, A., Corzo, G., Srinivasulu, S., Solomatine, D.P., 2010. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology—Part 2: concepts and methodology. Hydrol. Earth Syst. Sci. 14, 1931–1941.

Fuhrer, J., Skärby, L., Ashmore, M., 1997. Critical levels for ozone effects on vegetation in Europe. Environ. Pollut. 97, 91–106.

Fontes, T., Silva, L.M., Silva, M.P., Barros, N., Carvalho, A.C., 2014. Can artificial neural networks be used to predict the origin of ozone episodes? Sci. Total Environ. 488–489, 197–207.

Fusco, A.C., Logan, J.A., 2003. Analysis of 1970-1995 trends in tropospheric ozone at Northern Hemisphere midlatitudes with the GEOS-CHEM model. J. Geophys. Res. 108, 1988–1997.

Gao, M., Mo, D., Wu, X., 2016. Nonstationary modeling of extreme precipitation in China. Atmos. Res. 182, 1–9.

García, I., Rodríguez, J.G., Tenorio, Y.M., 2011. Artificial Neural Network models for prediction of ozone concentrations in Guadalajara, Mexico. In: Popović, D. (Ed.), Air Quality-models and Applications. InTech, pp. 35–52.

Gardner, M.W., Dorling, S.R., 1999. Neural network modeling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmos. Environ. 33, 709–719.

Hagan, M.T., Demuth, H.B., Beale, M.H., 1996. Neural Network Design. University of

Colorado, B, PWS Pub.

Han, Z., Ueda, H., An, J., 2008. Evaluation and intercomparison of meteorological predictions by five MM5-PBL parameterizations in combination with three land-surface models. Atmos. Environ. 42, 233–249.

Heo, J., Kim, D., 2004. A new method of ozone forecasting using fuzzy expert and neural network systems. Sci. Total Environ. 325, 221–237.

Jenkin, M.E., 2008. Trends in ozone concentration distributions in the UK since 1990: local, regional and global influences. Atmos. Environ. 42, 5434–5445.

Jenkin, M.E., Clemitshaw, K.C., 2000. Ozone and other secondary photochemical pollutants: chemical processes governing their formation in the planetary boundary layer. Atmos. Environ. 34, 2499–2527.

Kasiviswanathan, K.S., Sudheer, K.P., He, J., 2016. Quantification of prediction uncertainty in artificial neural network models. In: Shanmuganathan, S., Samarasinghe, S. (Eds.), Artificial Neural Network Modelling. Springer International Publishing, pp. 145–159.

Khan, J.A., Aelst, S.V., Zamar, R.H., 2007. Building a robust linear model with forward selection and stepwise procedures. Comput. Stat. Data Anal. 52, 239–248.

Liu, H., Liu, S., Xue, B., Lv, Z., Meng, Z., Yang, X., Xue, T., Yu, Q., He, K., 2018. Ground-level ozone pollution and its health impacts in China. Atmos. Environ. 173, 223–230.

Lelieveld, J., Crutzen, P.J., 1990. Influence of cloud and photochemical processes on tropospheric ozone. Nature 343, 227–233.

Luna, A.S., Paredes, M.L.L., Oliveira, G.C.G.D., Corrêa, S.M., 2014. Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at rio de janeiro, Brazil. Atmos. Environ. 98, 98–104.

Marquardt, D., 1963. An algorithm for least-squares estimation of nonlinear parameters. SIAM J. Appl. Math. 11, 431–441.

Mehrotra, K., Mohan, C.K., Ranka, S., 2000. (Second Printing) Elements of Artificial Neural Networks. MIT Press, Cambridge MA.

MEP, 2012. Ambient Air Quality Standards of People's Republic of China (GB3095-2012). (in Chinese).

Munir, S., Chen, H., Ropkins, K., 2013. Quantifying temporal trends in ground level ozone concentration in the UK. Sci. Total Environ. 458–460, 217–227.

Noori, R., Abdoli, M.A., Farokhnia, A., Abbasi, M., 2009. Results uncertainty of solid waste generation forecasting by hybrid of wavelet transform-ANFIS and wavelet transform-neural network. Expert Syst. Appl. 36 (6), 9991–9999.

Noori, R., Hoshiyaripour, G.A., Ashrafi, K., Araabi, B.N., 2010. Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration. Atmos. Environ. 44, 476–482.

Nourani, V., Fard, M.S., 2012. Sensitivity analysis of the artificial neural network outputs in simulation of the evaporation process at different climatologic regimes. Adv. Eng. Softw 47, 127–146.

Prybutok, V.R., Yi, J., Mitchell, D., 2000. Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. Eur. J. Oper. Res. 122, 31–40.

Revlett, G.H., 1978. Ozone forecasting using empirical modeling. J. Air Pollut. Control Assoc. 28, 338–343.

Russell, A., Dennis, R., 2000. NARSTO critical review of photochemical models and modeling. Atmos. Environ. 34, 2283–2324.

Şahin, M., 2012. Modelling of air temperature using remote sensing and artificial neural network in Turkey. Adv. Space Res. 50, 973–985.

Schaap, M., Timmermans, R.M.A., Roemer, M., Boersen, G., Builtjes, P., Sauter, F., Velders, G., Beck, J., 2008. The LOTOS EUROS model: description, validation and latest developments. Int. J. Environ. Pollut. 32, 270–290.

Shrestha, D.L., Kayastha, N., Solomatine, D.P., 2009. A novel approach to parameter uncertainty analysis of hydrological models using neural networks. Hydrol. Earth Syst. Sci. 13, 1235–1248.

Sun, W., Zhang, H., Palazoglu, A., 2013. Prediction of 8 h-average ozone concentration using a supervised hidden Markov model combined with generalized linear models. Atmos. Environ. 81, 199–208.

Taylan, O., 2018. Modelling and analysis of ozone concentration by artificial intelligent techniques for estimating air quality. Atmos. Environ. 150, 356–365.

Wagener, T., Gupta, H.V., 2005. Model identification for hydrological forecasting under uncertainty. Stoch. Environ. Res. Risk Assess. 19, 378–387.

Wang, W., Lu, W., Wang, X., Leung, A.Y., 2003. Prediction of maximum daily ozone level using combined neural network and statistical characteristics. Environ. Int. 29, 555–562.

Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63, 1309–1313.

Yetilmezsoy, K., Demirel, S., 2008. Artificial neural networks (ANN) approach for modeling of Pb (II) adsorption from aqueous solution by Antep pistachio (Pistacia Vera L.) Shells. J. Hazard Mater. 153, 1288–1300.

Yi, J., Prybutok, V.R., 1996. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environ. Pollut. 92, 349–357.

Zanis, P., Hadjinicolaou, P., Pozzer, A., Tyrlis, E., Dafka, S., Mihalopoulos, N., Lelieveld, J., 2014. Summertime free-tropospheric ozone pool over the eastern Mediterranean/Middle East. Atmos. Chem. Phys. 14, 115–132.