Atmospheric Pollution Research 12 (2021) 101066

Contents lists available at ScienceDirect

Atmospheric Pollution Research

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

Adjusting prediction of ozone concentration based on CMAQ model and machine learning methods in Sichuan-Chongqing region, China

Hua Lu [a,f], Min Xie [b,*], Xiaoran Liu [a,f], Bojun Liu [c], Minzhi Jiang [d], Yanghua Gao [a,f], Xiaoli Zhao [e]

[a] Chongqing Institute of Meteorological Sciences, Chongqing, 401147, China
[b] School of Atmospheric Sciences, Nanjing University, Nanjing, 210023, China
[c] Chongqing Meteorological Observatory, Chongqing, 401147, China
[d] Department of Applied Physical Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27514, USA
[e] Sichuan Meteorological Disasters Prevention Technology Center, Chengdu, 610072, China
[f] Chongqing Engineering Research Center of Agrometeorology and Satellite Remote Sensing, Chongqing, 401147, China

ARTICLE INFO

Keywords:
Ozone prediction
Machine learning
WRF-CMAQ
Sichuan-chongqing region

ABSTRACT

With increasing ozone pollution and deeper understanding of its harm to humans and climate, it is important to accurately forecast ozone. In this study, training and testing data sets were constructed with hourly numerical models forecasts and monitoring station observation for the year 2018 for Sichuan-Chongqing region, China. Three machine learning methods including Lasso, random forest and long short-term memory recurrent neural network (LSTM-RNN) coupled with CMAQ model were trained to forecast the ozone concentrations. The Lasso regression and random forest were used to realize feature optimization in four sub-regions separately. Coupled model with Lasso-random forest coupled feather selection schemes showed the best performance among different models. The main conclusions of adjusting results showed that deviations of hourly ozone prediction by CMAQ alone forecasts can be significantly reduced after machine learning coupled model adjusting, and correlation coefficients can be remarkably improved. Adjusting effects varied with different sub-regions and seasons. In three basin sub-regions, adjusting with random forest had the best performance, while in the plateau sub-region, adjusting with LSTM-RNN was most satisfactory, where root mean squared error decrease rate was 80.2% and correlation coefficient reached 91%. Machine learning methods performed better in summer and autumn for the three basin sub-regions, while in the plateau sub-region, adjusting was more significant in summer compared to other seasons.

1. Introduction

Ground-level ozone is mainly generated by a series of complicated photochemical reactions involving abundance of ozone precursors, such as nitrogen oxides (NOx) and volatile organic compounds (VOCs), under certain meteorological conditions. Severe ozone pollution is harmful to human health, ecological balance, climate, and has negative impact on regional air quality as the primary component of photochemical smog and greenhouse gas (Maji et al., 2019; Xie et al. 2014, 2016c). In recent years, with rapid industrialization and urbanization, complex air pollution characterized by high particulate matter and ozone is of great concern in many megacities of China, and has been gaining extensive attention of the public, researchers, and policymakers. Due to strict measures to improve air quality in megacities since 2013 by the Chinese government, fine particulate matter of 2.5 μm or less aerodynamic diameter ($PM_{2.5}$) has significantly reduced. However, ozone and its precursors have had an increased trend recently (Xie et al., 2016b; Zhao et al., 2018; Zhan et al., 2020), suggesting that more efforts should be paid to ozone pollution control.

With aggravated ozone pollution, it is important to provide reliable forecasting, which would allow more efficient countermeasures to prevent the air pollution crisis and protect public health (Arhami et al. 2013; Su et al., 2020). Traditional prediction methods, based on experience and statistical methods are simple in calculation, but cannot meet the demand of accuracy and high resolution for its lack in theory and efficiency (McKeen et al., 2009; Zhang et al., 2012). In recent years, the

Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.
* Corresponding author.
E-mail addresses: vibgyor0113@163.com (H. Lu), minxie@nju.edu.cn (M. Xie), liuxiaoran8283@126.com (X. Liu), 5inklbj@163.com (B. Liu), minzhi@unc.edu (M. Jiang), gaoyanghua@sina.com (Y. Gao), 49672734@qq.com (X. Zhao).

https://doi.org/10.1016/j.apr.2021.101066
Received 21 January 2021; Received in revised form 24 April 2021; Accepted 25 April 2021
Available online 27 April 2021

third-generation air quality model of U.S. Environmental Protection Agency Community Multiscale Air Quality system (CMAQ) was developed to integrate the latest research achievement of atmospheric chemistry and physics. CMAQ was designed for applications ranging from investigating complex formation and transport mechanisms of pollution, to regulatory and policy analysis, and has become one of the most widely used model in regional and urban air quality forecast, for its multiscale and highly flexible prediction ability (Wong et al., 2011; Wang et al., 2016; Lightstone et al., 2017; Qiao et al., 2019).

However, large bias error for O_3 forecast is a characteristic not only of the CMAQ model but also found in many other air quality forecast models, because of multiple reasons, such as inaccuracy of emission inventory, lack of description of extreme pollution events, and imperfection in chemistry and physics parameterization schemes (McKeen et al., 2009; Djalalova et al., 2015; Lightstone et al., 2017; Zhang et al., 2017). Although the essential cause of air pollution is the excessive emissions, objective factors like meteorological conditions play a crucial role in the occurrence of pollution, through controlling the dispersion of atmospheric pollutants and providing driving force for regional pollution variations (Yin et al., 2017; Ning et al., 2018; Zhao et al., 2018; Zhan et al., 2019). Uncertainties in meteorological input may contribute to the biases in air quality predictions (Huang et al., 2018). Therefore, bias-correction techniques applied to air quality forecast models have experienced considerable development. Previous studies have used various methods, including linear regression, kalman filtering, Bayesian method, to revise the model results respectively (Borrego et al., 2011; Djalalova et al., 2015; Huang et al., 2016; Mok et al., 2017). These methods rely on the analysis of simulated and observed air quality and meteorological data, and improve the accuracy of numerical simulation effectively. But these methods do not consider complex processes of pollutants transport, mixing, and feedback. Machine learning methods have been shown to be quite powerful in capturing the hidden non-linear relationships between air pollutant concentrations and meteorological factors by processing huge amounts of data and building the model more objectively and flexibly. Additionally, the performance of machine learning model does not degrade much with noisy data (Arhami et al., 2013). For its ideal performance in prediction with less input data and higher computational efficiency, considerable progress has been made in the development of machine learning applied in meteorological and air quality concentration forecast (Feng et al., 2015; Biancofiore et al., 2017; Freeman et al., 2017; Zamani, 2019).

Despite remarkable achievement, there are still big challenges for the application of machine learning for air quality predictions. An important step in developing machine learning models is to select proper input variables that have most significant impact on model performance, which could decrease network size, increase processing speed and efficiency (Arhami et al., 2013). The characteristics of air pollution in different regions vary with diverse regional meteorological and pollutant emission conditions (Zhao et al., 2017; Ning et al., 2018; Zhan et al., 2019). Therefore, it is essential to have deep research on feature selection, model training, and testing in different regions. The Sichuan-Chongqing region is located on the leeward slope of east side of Tibetan Plateau in the southwestern China, surrounded by high mountains, with basin topography in the middle, which is not conducive to diffusion of pollutants (Fig. 1). Additionally, rivers in the basin are vertical and horizontal, and this region experience high humidity conditions due to abundant water vapor. Covering about 100 million population, 11.3 million vehicle number and advanced industry (Zhao et al., 2017), heavy pollutants emissions are found in the region (Ning et al., 2018). Due to strong anthropogenic emissions, complex terrain, and unique atmospheric characteristics, the Sichuan-Chongqing region has been faced serious air pollution like other megacity clusters such as the Beijing-Tianjin-Hebei region, the Pearl River Delta, and the Yangtze River Delta, which is a sign of significant pollution in China (Zhao et al., 2017; Ning et al., 2018; Zhan et al., 2019). Although, SO_2 and PM_{10} were the primary air pollutants in the past, nowadays severe $PM_{2.5}$ and O_3 pollution have been exhibited in this region after optimizing energy mix (Tian et al., 2016; Wang et al., 2017a; Cai et al., 2018). Hence, to improve accuracy of forecasts and provide scientific support for regional air quality control, it is imperative to study prediction and adjusting high resolution air quality numerical models for the Sichuan-Chongqing region. Chongqing Meteorological Bureau has introduced and localized the CMAQ model, based on existing numerical weather forecast system, which could provide technical support and reference value forecasts of atmospheric pollutants.

The objective of this study was to determine reliable adjusting methods for forecasting O_3 concentrations in the Sichuan-Chongqing region. Three machine learning methods, including Lasso regression, random forest regression, and long short-term memory recurrent neural network (LSTM-RNN), were applied for adjusting O_3 concentration prediction by the Weather Research and Forecasting (WRF) coupled with CMAQ model. Referencing (Zhao et al., 2017), the regional 22 cities and autonomous monitoring stations were divided into four sub-regions according to geographic and climate features (Fig. 1 and Table 1). Lasso regression and random forest were used to realize the selection of input variables in four separate sub-regions, and three machine learning models were optimized with selected data sets to achieve more reliable prediction for sub-regions. Using prediction of WRF-CMAQ model as a baseline, this study explored the performances of different machine learning models, adjusting O_3 concentrations in different sub-regions, stations, and seasons. We present the data and methods in section 2, describe the construction of data sets, model training and the results of optimization input feature selection in four sub-regions, and discuss adjusting performances based on three machine learning methods in section 3. Finally, a summary is provided in section 4.

2. Data and methods

2.1. Datasets

Hourly O_3 data, from January to December 2018, was used as the target dataset for machine learning algorithms, which was collected from 22 air quality monitoring stations in Sichuan-Chongqing region. Additionally, six criteria pollutants of the previous day at 22 stations, including O_3, $PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO, were collected to construct the training and testing datasets. The real-time concentrations of six pollutants were downloaded from the website of China National Environmental Monitoring Centre (http://106.37.208.233:20035/). The quality assurance and quality control of these data, as well as the sampling methods of the 22 stations follow those in the Chinese national standard HJ/T 193–2005. The meteorological monitoring data, including pressure, temperature, dew point temperature, relative humidity, wind direction, wind speed, and precipitation, were obtained for the same period from China Weather Website Platform, maintained by the China Meteorological Bureau. The distance of each meteorological monitoring stations to the certain air quality monitor was calculated and the most nearest one was chosen. Index numbers and locations of the chosen meteorological monitoring stations could be found in Fig. 1 (b).

2.2. WRF-CMAQ model system

The air quality prediction data from WRF-CMAQ system of Chongqing Meteorological Bureau were used, including every hourly prediction concentrations of O_3, $PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO. To obtain more meteorological information, data near ground and on different pressure levels from WRF forecast were collected. The WRF-CMAQ model has a configuration with release time of 0800 local time zone (GMT+8) and providing a 72-h forecasts. Considering of the spin-up period to reach equilibrium (Ferreira et al., 2013; Jerez et al., 2020), the 0–12 h forecast results were commonly not used. Testing result in Fig. 2 including RMSE of WRF forecast temperature and wind speed

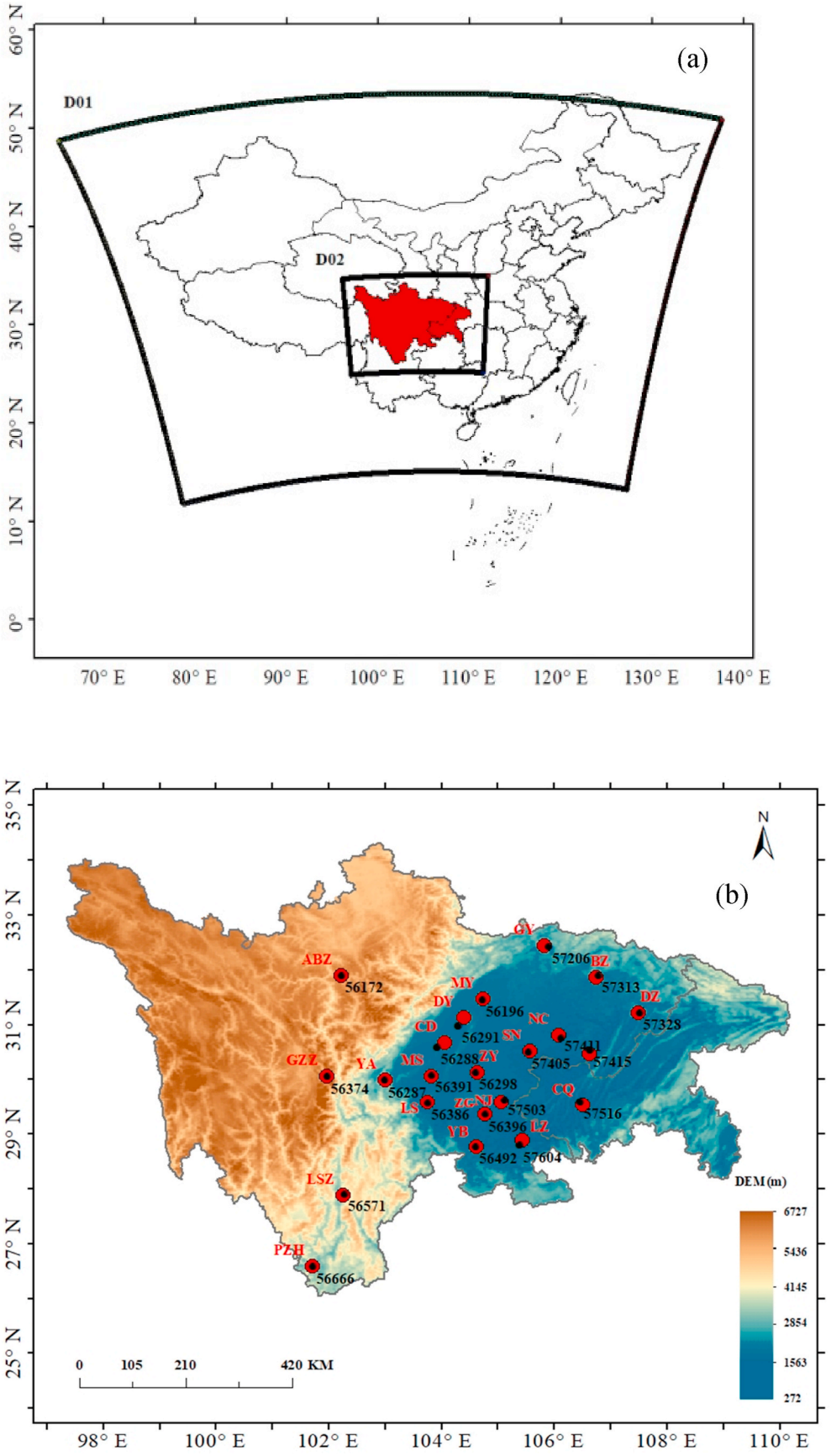

Fig. 1. Nesting domain of numerical model and location of the monitoring stations. (a) Location of the Sichuan-Chongqing region (shaded in red); (b) position of environmental (red dot) and meteorological (black dot) monitoring stations (note: topography of Sichuan-Chongqing region was shown on the map).

Table 1
Four sub-regions in Sichuan-Chongqing region.

Region	City
West Sichuan Basin (WSB)	Chengdu (CD), Mianyang (MY) , Deyang (DY) , Leshan (LS) ,Meishan (MS) ,Yaan (YA) ,Ziyang (ZY)
South Sichuan Basin (SSB)	Zigong (ZG), Yibin (YB), Luzhou (LZ), Neijiang (NJ), Chongqing (CQ)
Northeast Sichuan Basin (NESB)	Guangan (GA), Nanchong (NC) , Suining (SN), Guangyuan (GY), Dazhou (DZ), Bazhong (BZ)
Plateau of West Sichuan Basin (PWSB)	Abazhou (ABZ) , Ganzizhou (GZZ) , Liangshanzhou (LSZ) , Panzhihua (PZH)

showed forecast hours 24–48 of meteorological forecast by WRF were more reliable. As a result the forecast results of 24–48 h were used in the study. Model forecast data at the location of 22 air quality monitoring stations in Sichuan-Chongqing region, which were interpolated from the WRF and CMAQ 3 km resolution forecasts respectively. In general, we used 36 variables as initial input to construct machine learning models of adjusting hourly O_3 forecasts bias, which could be found in Table 2. Meteorological factors on both 850 and 700 hPa layer were chosen to construct the initial datasets, as representative of weather conditions of mixing layer and lower troposphere layer. Local air pollution emissions are mainly permeated by convection within the mixing layer, and weather system at 700 hPa may play a key role in the formation of air pollution in this region (Ning et al., 2018). The complex terrain topography with high mountains around makes unique relationship between meteorological factors and pollutants in the Sichuan-Chongqing region. Double counting these meteorological factors for 850 and 700 hPa pressure level can better represent the convection characteristics in the lower atmosphere above this region.

The CMAQ model (Binkowski et al., 2003) version 4.7.1 and WRF model (Skamarock et al., 2008) version 3.5.1 were used to forecast atmospheric composition concentrations. A 2-nested domain was applied for model set up, with horizontal grid spacing of 9 km and 3 km respectively. The 9 km domain (600 × 480 grid cells) covers China and its surrounding countries, while the 3 km domain (480× 360 grid cells) covers the Sichuan Basin region and adjacent regions. The model domain design could be found in Fig. 1(a). The hourly forecasts datasets of Global Forecast System from National Centers for Environmental Prediction with 0.5°x0.5°horizontal resolution were used to create initial and border conditions for WRF. Configuration features of WRF were as following parameterization: physics of the model include the Yonsei University scheme (Hong et al., 2006) for the planetary boundary layer, RRTMG radiation physics (Mlawer et al., 1997) option, Noah land-surface model (Dudhia, 2001), Thompson microphysics options (Thompson et al., 2008), Monin-Obukhov surface layer scheme (Monin and Obukhov 1954). The CMAQ was setup using the same horizontal grid and domain regions with the WRF as shown in Fig. 1 (a). And for the vertical direction, we set 15 sigma layers, in which the boundary layer is divided into 8 layers. The details of CMAQ configuration are including PPM advection, multiscale horizontal diffusion, eddy vertical diffusion (Appel et al., 2008), CB05cl chemistry solver (Yarwood et al., 2005) and aero5 aerosol modules. The biogenic emissions were calculated by Model of Emissions of Gases and Aerosols from Nature (Guenther et al., 2006). The anthropogenic emissions were results from Inversion of Multi-resolution Emission Inventory (MEIC) for China. It was developed by Tsing Hua University with a resolution of 0.25 ° × 0.25 °, including emissions of agriculture, transport, industry, power plants and human residence (Wang et al. 2016, 2017b). In this study, MEIC was interpolated to the model domain with consideration of population and gross domestic product. Besides, the anthropogenic emission was managed with ensemble square root Kalman filter (Wu et al., 2018).

2.3. Lasso regression

Least absolute shrinkage and selection operator (Lasso) is a biased estimation method, which could select elements with higher relevancy of the concerned target, while dismissing the rest to simplify the input datasets (Tibshirani 2011). The level of relevancy can be determined by construction of a penalty function and compression of some coefficients

Table 2
Variables used to construct the initial datasets.

Data source	City	
Atmospheric component monitering	O_3, $PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO of the previous day	
Weather monitoring station	pressure, temperature, dew point temperature, relative humidity, wind direction, wind speed, and precipitation	
WRF forecast	Surface	10 m U and V wind component, 2 m dew point temperature, 2 m air temperature, boundary layer height, surface solar radiation, sunshine duration, total precipitation
	Pressure level	850 hPa temperature, 850 hPa U and V wind component, 850 hPa vertical velocity, 850 hPa vorticity, 700 hPa U and V wind component, 700 hPa vertical velocity, and 700 hPa vorticity
CMAQ forecast	O_3, $PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO	

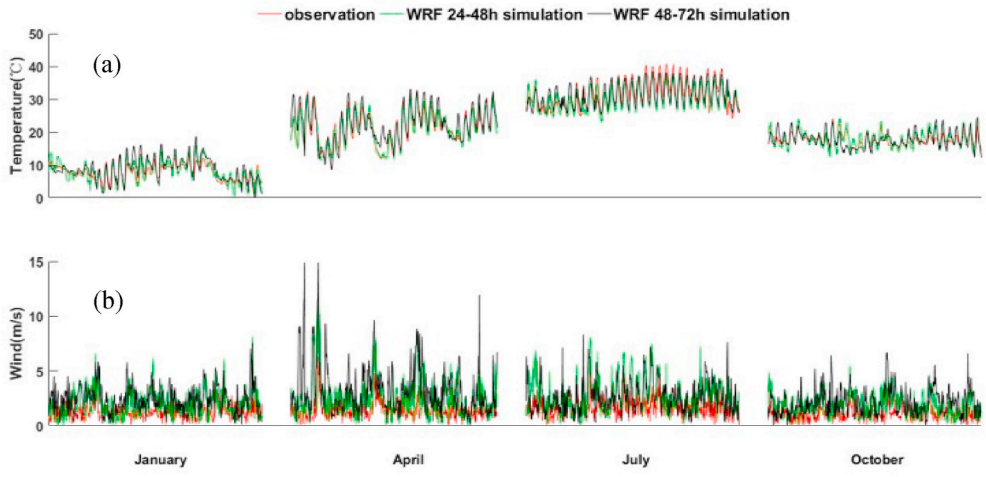

Fig. 2. Comparisons of observation and WRF simulation in the representative months of 4 seasons.

by setting a threshold. It is a kind of biased estimation for processing data with complex multi-collinearity. Lasso regression is a more realistic regression method to obtain the regression system at the cost of losing some information by giving up the unbiasedness of least square method.

In this study, we predicted hourly O_3 concentrations through linear combination of given n input features $X = (x_1, x_2 \ldots x_n)$; the equation for calculating is as follows:

$$f(x) = \sum_{i=1}^{n} \lambda_i x_i + \lambda_0. \quad (1)$$

By minimizing the loss function:

$$loss(\lambda) = y - \left(\sum \lambda_i x_i + \lambda_0 \right)^2 + \alpha\lambda. \quad (2)$$

$\lambda_i$is the weight matrix of the input features and λ_0 is the deviation matrix. λ_i and λ_0 are calculated, giving the Lasso regression model that could predict the O_3 concentration. $\alpha\lambda$ is a regularization term that not only helps to reduce the risk of overfitting, but also enables the selection of characteristic variables, y is the observation of O_3 concentration, and x_i is i_{th} input feature.

2.4. Random forest regression

Random forests are a combination of tree predictors, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman 2001). Since it is relatively robust to noise, random forests are not prone to overfitting, so that it is carried out in various fields of data mining (Zamani 2019).

To predict the O_3 concentration, the same target and input feature matrix were used in random forest regression. The main steps to construct the model are as follows:

Step 1: The basis feature data can be constructed as follows:

$$D = \{(x_m, y_m), m = 1, 2 \ldots, n\}, (X, Y) \in R^i * R, \quad (3)$$

where Y is the observation of O_3 concentration, and X is input feature matrix.

Step 2: To grow each tree of h_i, random subspace D_j must be generated through a random selection with replacement from D, among which the optimal feature was selected and divided. Then, an ensemble of N trees h_i are grown because of repeated training.

Step 3: The prediction result is an average of N trees h_i. Random forest regression is a multiple non-linear regression model, using the idea of double random, resulting in random forest not prone to overfitting, and diversity among classifiers as well.

2.5. Deep learning

Deep learning is defined as neural networks composed of multiple layers, which has advantage of persisting information of previous events compared to traditional neural networks. RNN is one of such system that can process information in a loop. Each network in the loop takes information from the previous network, performs the specified operation and produces output with passing information to the next network meanwhile (Schmidhuber 2015). However, some applications may require information from a more previous time period that exceeds what RNN could provide. Fortunately, LSTM networks are capable in such situations. LSTM, which is a special form of RNN, is designed to solve the problem of long term dependency issue of RNN (Kumar et al. 2018). A typical LSTM structure has three gates, including an input gate, a forget gate, and an output gate. Each gate consists of an activation function, *sigmoid,* and a concatenation operation (•). Gates can decide the amount and flow of previous information to the cell sate, which is the fundamental component of LSTM as shown in Fig. 3.

The steps for utilizing the LSTM are as follows:

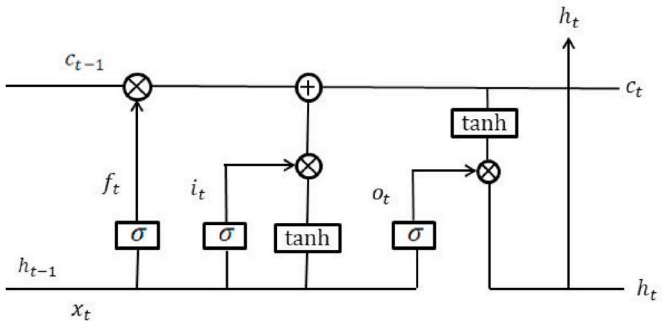

Fig. 3. Concept of LSTM

Step 1: The discarded information needs to be judged through a forget gate based on input h_{t-1} and x_t, shown in Equation (4), where W_{fx}, W_{fh}, W_{fc} are the weight matrix of the input, output of the previous moment and memory cell to the forget gate, respectively, b_f is the deviation matrix of the forget gate, and f_t is the output of the forget gate.

Step 2: The information to add to the cell state through the input gate must be decided, including W_{ix}, W_{ih}, and W_{ic} as weight matrices for the input, output of the last moment, and memory cells to input gate, respectively, b_i as deviation matrix of the input gate, and i_t as output of the input gate (Equation (5)). The cell state must also be updated from c_{t-1} to c_t, which is done using Equation (6).

Step 3: The output information to flow out from output gate is then controlled using Equation (7), where W_{ox}, W_{oh}, and W_{oc} are weight matrix of the input, output of the last moment, and memory cells to input gate, respectively, b_o is the deviation matrix of the output gate, and o_t is information from output gate. Finally, the output of LSTM h_t module can be computed by activation function ϕ and o_t, using Equation (8).

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f\right) \quad (4)$$

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i) \quad (5)$$

$$c_t = f_t^{\circ}c_{t-1} + i_t^{\circ}\phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (6)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_t + b_o) \quad (7)$$

$$h_t = o_t^{\circ}\phi(c_t) \quad (8)$$

We used an LSTM-RNN model to handle prediction of the hourly O_3 concentration of time series to adjusting the numerical forecasts. The model was constructed with an input layer, a LSTM network layer, and adjusted as 10 cell numbers in hidden layer, tanh as activation function and *Adam* optimizer. The target feature matrix used was the same as in Lasso and the random forest model. We trained the model with a learning step of 20, batch size dynamically adjusting with different subregions, diving time series into 20, and 200 times iterative training. Finally, validation of the model was performed with the testing data set.

2.6. Evaluation methods

Mean bias (MB) and root mean squared error (RMSE) were adopted as indicators (Equations (9) and (10)) for evaluation, which could compare absolute and relative deviation between prediction and observation, respectively. Additionally, Pearson correlation coefficient (R) was calculated to show the linear relationship, as shown in Equation (11).

$$\text{MB} = \frac{1}{n}\sum_{i=1}^{n}(m(i) - o(i)) \quad (9)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(m(i) - o(i))^2}{n}} \quad (10)$$

$$R = \frac{cov(m,o)}{\sqrt{cov(m,m)*cov(o,o)}} \quad (11)$$

In the above equations, $m(i)$ is prediction, $o(i)$ is observation, $cov(m,o)$ stands for the covariance, and $cov(m,m)$ and $cov(o,o)$ are variance of prediction and observation, respectively.

3. Results and discussions

3.1. Model construction

The technical flowchart of the model construction is presented in Fig. 4. Firstly, data set consisting of numerical model prediction and observation data was used. A random selection was performed on the basis data set; 85% of the data set was selected to be used as training data, while the remaining 15% was used as the testing data set. Secondly, Lasso regression and random forest model were trained with the training data set, which in turn optimized the input feature selection. Lasso, random forest regression and LSTM-RNN alone and coupled with the CMAQ model were trained with the different selected feature schemes to construct the prediction model. Best schemes for different algorithms were selected. Finally, the testing data set was used in the three machine learning models with the best schemes to realize the prediction of hourly O_3 concentration and adjusting of the CMAQ forecasts.

3.1.1. Data set construction

The basis input feature data set for machine learning methods included simulated hourly air quality data from CMAQ forecasts of 2018, simulated hourly meteorological data from WRF forecasts of the same period, the previous day's observed hourly air quality and meteorological data from 22 ground monitoring stations, while the target data set was the observed hourly ozone concentration on the day. Data preprocessing was performed to the basis data set, which included discarding samples that contained null values or messy codes, and data standardization was carried out, which not only avoided the influence to training results by different magnitude of each feature, but could also improve the calculation efficiency. There were 8760 times of hourly data for 2018, while the four sub-regions of WSB, SSB, NESB and PWSB in Sichuan-Chongqing region originally contained 61320, 43800, 52560, and 35040 samples, respectively. After data preprocessing, there were 50677, 35950, 42355, and 28465 samples for the four sub-regions separately, where each sample contained 36 input and 1 target features. To guarantee the spatial and time representativeness of the training and testing data sets and avoid the impact of some observation instrument and environment on training and testing results, as well as to obtain sufficient samples that may affect the tree or neural network training (Halevy et al. 2009), we disrupted the data for all regions and times in each sub-region and randomly selected 85% samples as training data set. Therefore there were, respectively, 43073, 30557, 35999, and 24194 samples for WSB, SSB, NESB and PWSB, while the remaining 15% were testing data set, where 7604, 5393, 6356, and 4271 samples were included in four sub-regions separately.

3.1.2. Input feature optimization

An important step in developing machine learning models is to select input feature that have most significant impact on the model (Arhami et al., 2013; Kang et al., 2020). Feature optimization could reduce the dimensions of the input data set, thereby cutting down computational cost and improving the interpretability of the machine learning model. In this study, we performed feature optimization in the way of Lasso regression training. With the enhancement of punishment, weight coefficients of some features were set to be 0, hence their features were discarded and feature optimization was realized. Lasso regression is an embedded selection method, which integrates feature optimization and model training (Tibshirani 2011). In our study, we obtained weight coefficients of 36 input features through Lasso regression training in four sub-regions respectively, sorted the absolute value of the weight coefficients and selected features with most significant impact on ozone prediction in different regions. Besides, feature optimization with random forest algorithm was also conducted. The features were sorted by the importance calculated by the random forest algorithm. When dimension of the selected input feature matrix reaches a certain value, the RMSE between prediction and observation will become stable, implying that the RMSE does not reduce significantly (Fig. 5). Therefore, we select different dimensions of input feature matrix due to the four sub-regions, with 17 input features for WSB, 14 for SSB, 12 for NESB, and 13 for PWSB with the Lasso regression. While 16 features for WSB, 17 for SSB, 15 for NESB and 13 for PWSB were selected by the random forest. The theories of feature selection with the 2 algorithms are different, resulting in various optimization consequences.

The feature optimization results of four sub-regions in the Sichuan-Chongqing region are presented in Table 3. Ground-level ozone is mainly generated by complex photochemical reactions between precursors NO_x ($NO_x = NO + NO_2$) and VOCs (Xie et al., 2016c; Zhan et al., 2020). Increased UV radiation could photolyze NO_2 into NO. Moreover, the interaction between NO_2 and SO_2 could lead to variation of sulfate and nitrate aerosol, and consequently affect ozone formation (Ma et al., 2017). Besides, considering the similar sources of CO and VOCs, CO can be regarded as proxy for VOCs. Interaction between particulate matter and ozone concentrations varied with different conditions. In winter, there was a significant decrease in ground-level ozone along with increasing high particulate matter that may reduce solar radiation. However, in summer, increased ozone could contribute to secondary aerosol formation (Wang et al., 2020). Feature optimization results show that the simulation and observation atmospheric composition are generally effective predictors of ozone, but specific composition differ between regions. For example, because of the non-linear dependence of

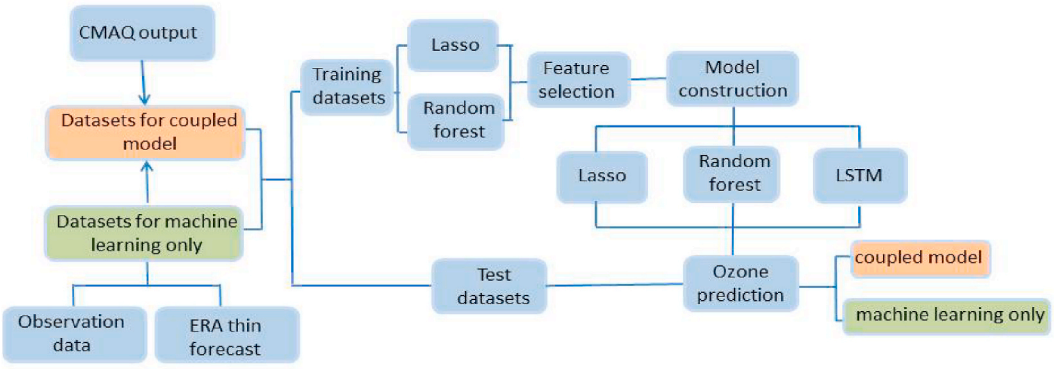

Fig. 4. Technical flowchart of O_3 prediction and adjusting.

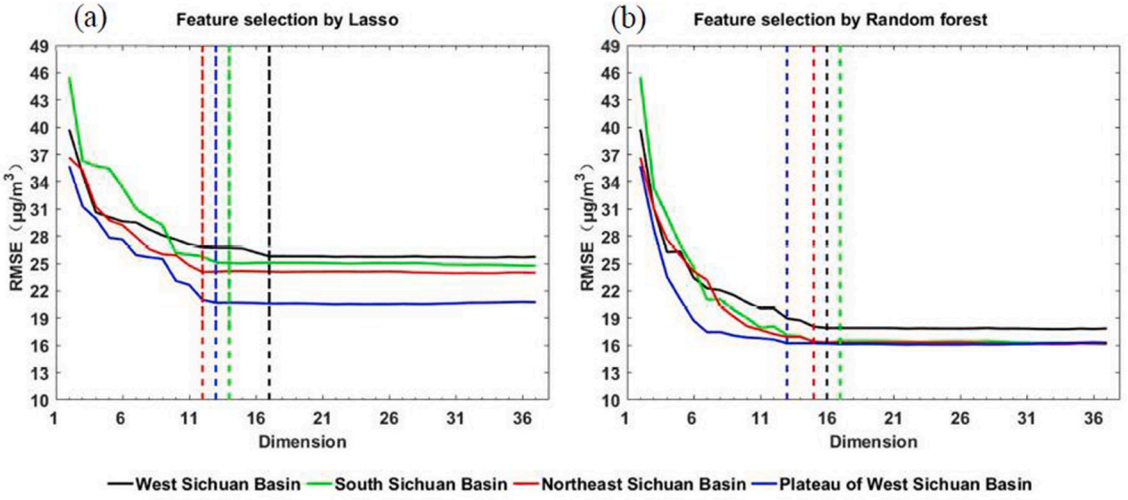

Fig. 5. RMSE between the observation and the prediction of hourly ozone concentrations based on Lasso regression and random forest.

Table 3
Feature optimization of 4 sub-regions in Sichuan-Chongqing region through Lasso and Random forest.

Sub-region	Model	Feature optimization
WSB	Lasso	O_3, NO_2, CO, SO_2, $PM_{2.5}$, O_3 of CMAQ forecasts, $PM_{2.5}$ of CMAQ forecasts, pressure, wind speed, relative humidity, 850 hPa temperature, surface solar radiation, boundary layer height, 850 hPa U wind component, 850 hPa V wind component, 700 hPa vertical velocity, 700 hPa V wind component
	Random forest	O_3, NO_2, CO, SO_2, $PM_{2.5}$, O_3 of CMAQ forecasts, pressure, wind speed, relative humidity, $PM_{2.5}$ of CMAQ forecasts, surface solar radiation, boundary layer height, 850 hPa U wind component, 850 hPa V wind component, 700 hPa vertical velocity, 700 hPa V wind component
SSB	Lasso	O_3, NO_2, CO, SO_2, $PM_{2.5}$, O_3 of CMAQ forecasts, NO_2 of CMAQ forecasts, SO_2 of CMAQ forecasts, $PM_{2.5}$ of CMAQ forecasts, 2 m temperature, surface solar radiation, boundary layer height, 850 hPa V wind component, 700 hPa U wind component
	Random forest	O_3, NO_2, CO, SO_2, $PM_{2.5}$, O_3 of CMAQ forecasts, NO_2 of CMAQ forecasts, SO_2 of CMAQ forecasts, $PM_{2.5}$ of CMAQ forecasts, 2 m temperature, surface solar radiation, boundary layer height, 850 hPa V wind component, 700 hPa U wind component, 700 hPa V wind component, total precipitation, 700 hPa vertical velocity
NESB	Lasso	O_3, NO_2, SO_2, O_3 of CMAQ forecasts, $PM_{2.5}$ of CMAQ forecasts, temperature, 2 m temperature, 850 hPa temperature, surface solar radiation, boundary layer height, 700 hPa V wind component, 700 hPa vertical velocity
	Random forest	O_3, NO_2, SO_2, O_3 of CMAQ forecasts, $PM_{2.5}$ of CMAQ forecasts, temperature, 2 m temperature, 850 hPa temperature, surface solar radiation, boundary layer height, 10 m V wind component, 700 hPa V wind component, 700 hPa vertical velocity, 850 hPa V wind component, 850 hPa U wind component
PWSB	Lasso	O_3, NO_2, SO_2, $PM_{2.5}$, PM_{10}, O_3 of CMAQ forecasts, NO_2 of CMAQ forecasts, temperature, 2 m temperature, 850 hPa temperature, 850 hPa V wind component, surface solar radiation, 700 hPa vertical velocity
	Random forest	O_3, NO_2, SO_2, $PM_{2.5}$, PM_{10}, O_3 of CMAQ forecasts, NO_2 of CMAQ forecasts, temperature, 2 m temperature, 850 hPa temperature, 850 hPa V wind component, surface solar radiation, 700 hPa vertical velocity

ozone formation on its precursors, for urban and industrial areas like WSB and SSB with Chengdu and Chongqing as representatives, respectively, ozone production is known as VOCs-limited, hence lower VOCs will suppress ozone formation, and lower NO_x with stable VOCs will lead to increasing to ozone production (Wang et al. 2016, 2020). Therefore, ozone concentration significantly related to CO and NO_2, which are commonly proxy for VOCs and NO_x. While in NESB and PWSB, that are relatively less urbanized, ozone formation is mainly affected by NO_x for VOCs emission is low, as a result ozone prediction is related to NO_2 in this 2 regions. Besides, pollution in WSB, SSB, and NESB were mainly fine particulate matter pollution (Zhao et al., 2018). While PWSB is near to Tibetan Plateau, thus air quality of PWSB is affected by dust aerosol transported from Takla Makan Desert. Study showed that dust have mostly influence on air quality and climate feedback compared to other aerosols (Jia et al., 2015). Therefore, as opposed to the other three regions, ozone prediction in PWSB was mainly related to PM_{10}.

Numerous previous studies have revealed the impact of meteorology conditions on ozone (Jasaitis et al., 2016; Tao et al., 2016; Xie et al., 2016a). In this study, feature optimization results also showed ozone prediction to have a relationship with related meteorological features. Solar radiation has an influence on photolysis of ozone precursors, and affects photochemical reaction rate and activity (Wang et al., 2020). Hence, solar radiation is an inevitable factor in ozone concentration prediction. As verified in feature optimization results in all regions (Table 3), ozone prediction showed strong correlation with surface solar radiation and sunshine duration. The seasonal variations of surface ozone in Sichuan-Chongqing region showed peak concentrations in late spring and summertime with higher temperature and stronger radiation (Zhao et al., 2018). Therefore, temperature was selected as an important factor in ozone concentration prediction as well. Moreover, some other meteorological parameters can affect the horizontal and vertical transportation that relate to ozone prediction (Tao et al., 2016). For instance, in the three sub-regions WSB, SSB and NESB that primarily have basin topography (Fig. 1), feature optimization results showed that boundary layer height and wind cannot be neglected in ozone concentration prediction. In addition, when the weather system underwent a holistic change, such as with the arrival of cold air accompanied by wind direction turning northerly, observable increase in pressure, and decrease in temperature and humidity, surface air pollutants concentrations was generally lower (Zhao et al., 2017; Zhong et al., 2018). It is worth noting that not low importance of surface wind in calculate the air pollutant could be found with the feature optimization in some sub-regions. Studies show the reliance of pollutants on surface wind is not significant in 4 sub-regions of Sichuan-Chongqing region, with R^2 of only 0.12, 0.04, 0.04 and 0.01 respectively (Zhao et al., 2017). The result may be related to the local special terrain topography with high mountains surrounded, which confine the air flow in the terrain. Low wind speed prevails all the year (Fig. 2) and averaged wind speed is below 1.5 m/s

according to the Sichuan and Chongqing statistical yearbook of recent years.

Lastly, in this study, we obtained training and testing dataset of various sub-regions based on feature optimization using the Lasso, random forest and Lasso-random forest (L-R) coupled methods. L-R coupled methods adopted all the variables of the two selection results. Sample numbers of different seasons in each station and sub-region are shown in Fig. 6, which establishes that training and testing the dataset are representative of space and time, since samples of training and testing distribute evenly. Then, we used the three machine learning methods with training dataset to construct Lasso regression, random forest, and LSTM-RNN models, and validated the model results with the testing dataset.

3.2. Machine learning model evaluation

3.2.1. Evaluation in different sub-regions

To revealed the effects of the coupling between machine learning and numerical model, we compared the machine learning only and coupled model with 3 feature optimization schemes respectively. The RMSE of each scheme in different sub-regions was given in Fig. 7. L-R coupled selection scheme showed best results in various models compared with single Lasso or random forest selection schemes. In addition, bias of coupled model was generally lower than the corresponding machine learning only model. As for the various machine learning algorithms, Lasso only and coupled with the CMAQ model had high bias compared with the other two algorithms. Generally speaking, coupled models with the L-R coupled feature optimization schemes showed the best result and we chose these to test ozone forecast results in different sub-regions and seasons below.

Table 4 summarizes statistical parameters as representation of deviation between hourly ozone concentration observation and prediction by different models for each sub-regions for 2018. The values of MB for CMAQ were generally negative, implying that the ozone concentrations from the numerical simulations were lower than the observed values, and the forecast results from the 3 machine learning coupled models decreased MB to about ± 1 μg/m^3, indicating that coupled model showed better result than the CMAQ model only. It is worth noting that prediction deviation of CMAQ in PWSB was obviously larger than that in the three basin regions, as the values of RMSE in the three basin regions were around 60 μg/m^3, while RMSE in PWSB reached 69.72 μg/m^3. Additionally, CMAQ-MB in PWSB was 1.5–2 times of that in three other regions. Studies have shown that the PWSB region may be an important ozone source for the cities in the Sichuan basin, especially the western Sichuan basin (Zhao et al., 2018). Although particulate matter pollution in PWSB was less than other regions in the Sichuan-Chongqing region, ozone concentration was high in this region (Zhao et al., 2017), which was one of the reason why prediction deviation in PWSB was more significant. Moreover, the parameterization setting of CMAQ in this study was more suitable for cities in basin region like Chongqing and Chengdu. Hence, higher prediction deviation was present in plateau regions like PWSB. Adjustment results from the three machine learning models dramatically decreased the values of RMSE from CMAQ forecasts. For WSB, SSB, and NESB, the adjustment based on Lasso regression, random forest and LSTM-RNN coupled with CMAQ model decreased RMSE by approximately 60%, 70% and 70% on average, respectively. The results from random forest were better than those from LSTM-RNN in these basin regions. Additionally, for the PWSB region, there was a decrease of approximately 71.2% based on Lasso regression, 77.7% for random forest, and 80.0% for LSTM-RNN. In all, adjusting in PWSB was more significant than those in the other three basin regions. The result based on LSTM-RNN was better than the other two machine learning methods in PWSB, while the opposite was the case in the other three basin regions.

The scatter plots between observation and prediction of hourly ozone concentrations is illustrated in Fig. 8 for four sub-regions. The prediction effect of ozone change trend by CMAQ forecasts and adjusted by the three machine learning methods can be compared. Prediction using CMAQ forecasts was generally lower than observations with the scatter distribution near the X axis (the correlation coefficients were 0.31, 0.26, 0.36 and 0.26 in WSB, SSB, NESB and PWSB, respectively). Compared with MB and RMSE of CMAQ forecasts in Table 4 and scatter plots, we found that the prediction deviation by numerical model in SSB was low but ozone change trend prediction was poor, while in PWSB the deviation was high and correlation coefficient was small at the same time. As we see from all figures, points after adjusting showed better goodness-of-fit, which verified the usefulness of the three machine learning methods. Among that, correlation coefficients between observation and prediction adjusted by Lasso regression model reached 0.78, 0.85, 0.79, and 0.83 in the four sub-regions. In addition, adjusting effects were more significant in NESB and PWSB, where the original prediction by CMAQ for ozone change trend was poorer. It is obvious that random forest and LSTM-RNN models had better adjusting effects compared to Lasso regression. For the three basin regions, adjusting effect by random forest model was slightly better than LSTM-RNN. However, in the PWSB region that is located on the edge of plateau, LSTM-RNN was superior to random forest.

3.2.2. Evaluation in different stations

The statistical parameters, including MB, RMSE and R, of observations and predictions (by original CMAQ forecasts and after adjusting using the three machine learning methods) were calculated for 22

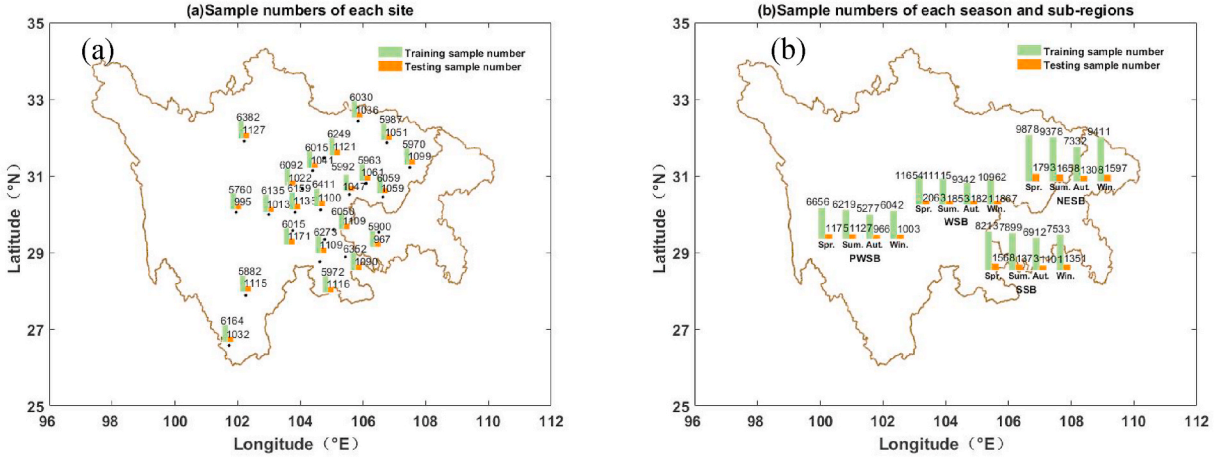

Fig. 6. Sample number distribution of each season and sub-regions in the Sichuan-Chongqing region.

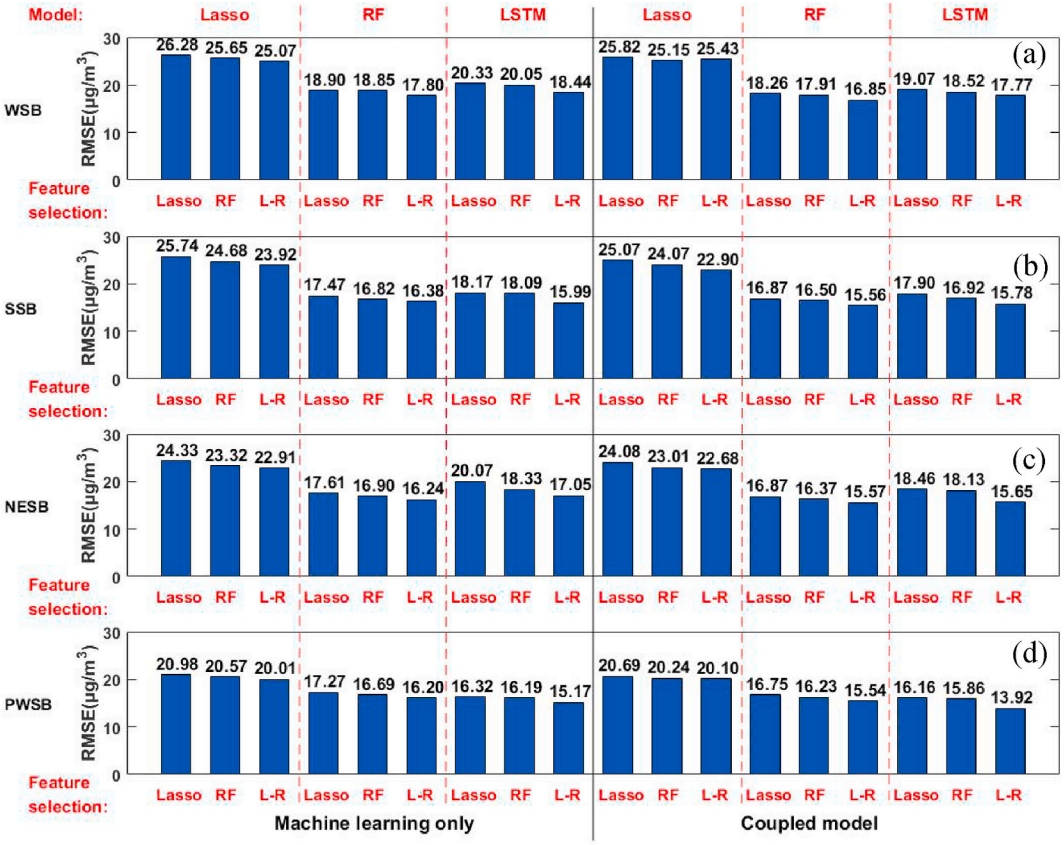

Fig. 7. RMSE for ozone forecasts and adjusting at stations throughdifferent machine learning only and machine learing-CMAQ coupled models with various feature selection schemes in 4 sub-regions of Sichuan-Chongqing region, RF is short for random forest and L-R is for Lasso-random coupled.

Table 4
Summary of statistical parameters.

Region	MB(μg/m^3)				RMSE (μg/m^3)				Decreasing rate of RMSE (%)		
	CMAQ	Lasso	RF	RNN	CMAQ	Lasso	RF	RNN	Lasso	RF	RNN
WSB	−31.57	0.61	0.42	−0.05	61.49	25.43	16.85	17.77	58.6	72.6	71.1
SSB	−21.49	−0.35	−0.13	0.73	58.25	22.90	15.56	15.78	60.7	73.3	72.9
NESB	−39.27	0.20	0.13	0.21	54.95	22.68	15.57	15.65	58.7	71.7	71.5
PWSB	−58.43	−0.46	0.15	−0.37	69.72	20.10	15.54	13.92	71.2	77.7	80.0

stations and is shown in Fig. 9. MB of CMAQ forecasts in 20 stations (except Chongqing and Chengdu) were commonly significant negative deviation. Studies have shown that ozone concentration can increase under high temperature and low wind speed (Tao et al., 2016). In this study, model evaluation results show that temperature estimated by WRF was generally 1–2 °C lower than the observed values at the peak value period, and wind speed was significantly higher. These bias in meteorological factors could be one of the reasons for the underestimation in ozone forecasts. Qiao et al. (2019) also reached a similar conclusion in numerical model study in the Sichuan basin. Moreover, in recent years, anthropogenic emissions have changed remarkably with rapid economic development. The used emission inventories in models generally do not account extreme events, which can result in distinction of inventory from reality, and thereby contribute to deviation of ozone and precursor concentration from CMAQ forecasts (Lightstone et al., 2017). RMSE of ozone predictions by CMAQ forecasts in Chongqing and Chengdu were as remarkable as in other 20 stations, which implies that low absolute values of MB in the two stations were due to the offset. After adjusting by using the three machine learning methods, absolute values of MB for all stations decreased to zero. Additionally, RMSE reduced to 20–30 μg/m^3 and R increased to 0.75–0.85 by using Lasso regression, while RMSE reduced to 15–20 μg/m^3 and R increased to 0.85–0.95 by the methods of random forest and LSTM-RNN.

Fig. 10 shows the ozone concentrations results of the three machine learning coupled models by calculating the decreasing rate of RMSE, MB and increasing rate of R at 22 stations in the Sichuan-Chongqing region. As for the MB decreasing rate, all the 22 stations showed similar adjusting effects, because of the cancelling out of positive and negative values. We discussed the adjusting effects with RMSE and R as below. For the 18 stations in the basin areas, the RMSE decreased 55–80%. With the aid of random forest and LSTM-RNN coupled model, however, the RMSE decreased even more, with the reduction reached 70–85%. Increasing rate of R shower higher value in GY and GZZ, where lower R value could be found according to Fig. 9. Among the three machine learning algorithms, the results from random forest adjusting were the best in the basin cities. Similar to Fig. 8, adjusting effect was more significant at four stations in the PWSB region according to RMSE and R. For example, at the stations of ABZ, GZZ and LSZ, the RMSE values decreased around 80%. But for PZH, it is located on the border of a basin and a plateau. Thus, its characteristics are close to the stations in the basin areas rather than ABZ, GZZ and LSZ. In general, adjusting through the three machine learning coupled models remarkably reduced the

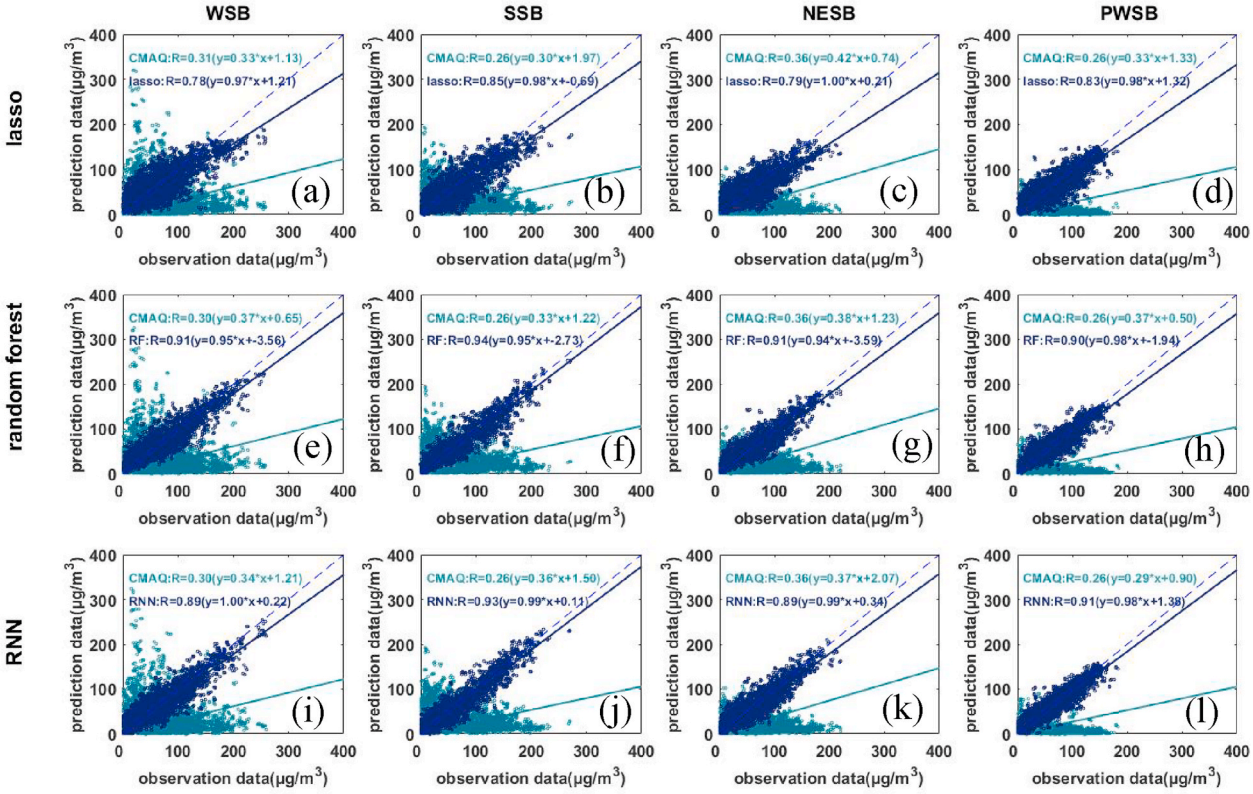

Fig. 8. Scatter plot between observations and predicted hourly ozone concentrations by CMAQ and three adjusting methods in different sub-regions, the solid lines represent the best fitting lines (light blue is for CMAQ and dark blue for adjusting models).

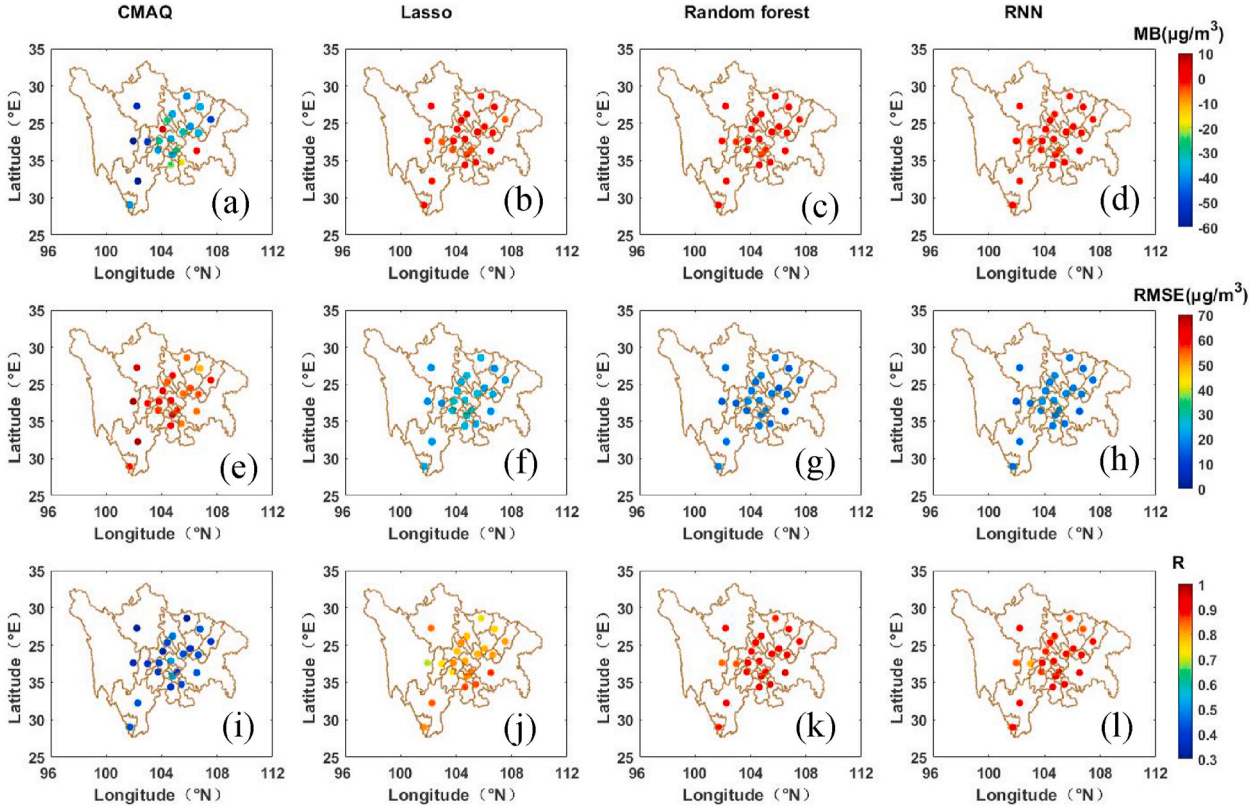

Fig. 9. MB, RMSE and R for ozone forecasts and adjusting at stations in the Sichuan-Chongqing region.

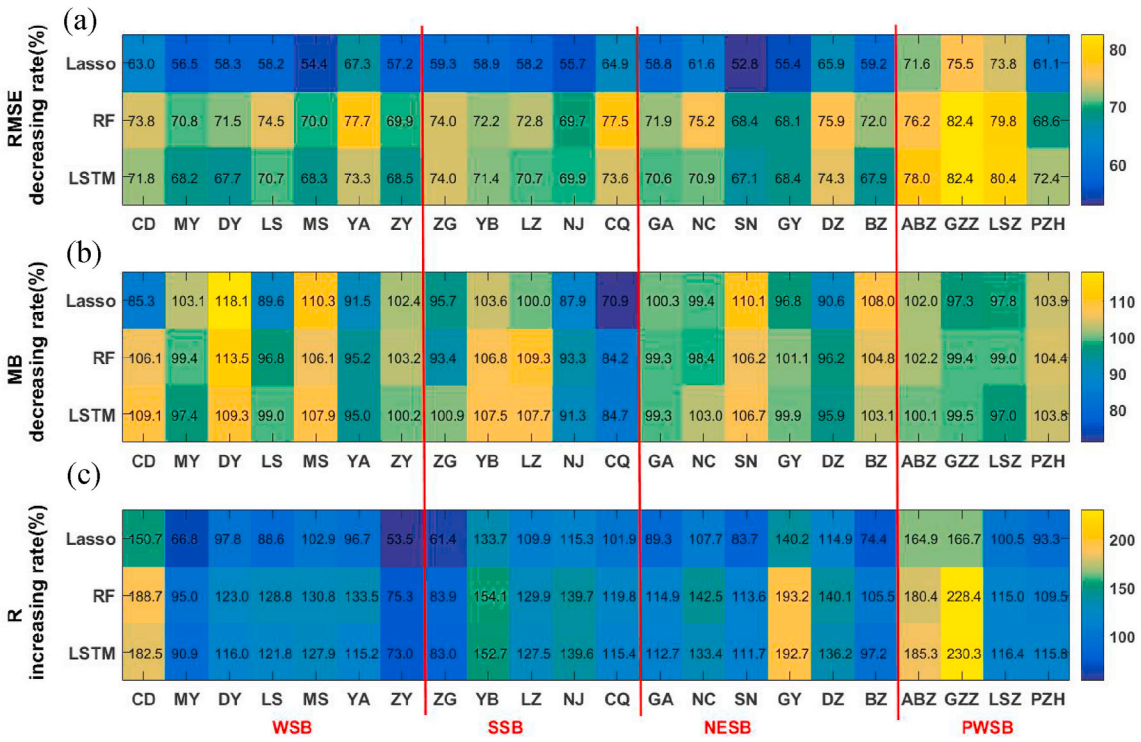

Fig. 10. Decreasing rate of RMSE, MB and increasing rate of R for ozone adjusting effect of the three machine learning methods at stations in Sichuan-Chongqing region.

deviations of hourly simulated ozone concentrations by CMAQ alone for 22 stations, and increased correlation coefficients as well. Adjusting effect by random forest model was better than the other two methods for the 18 stations with basin topography, while RNN-LSTM was better in the four cities in PWSB with plateau topography. The correlation relationship of ozone prediction and relative input feature parameters was obviously different between the 18 stations in basin and the four stations in PWSB. The random forest has better prediction effect for linear

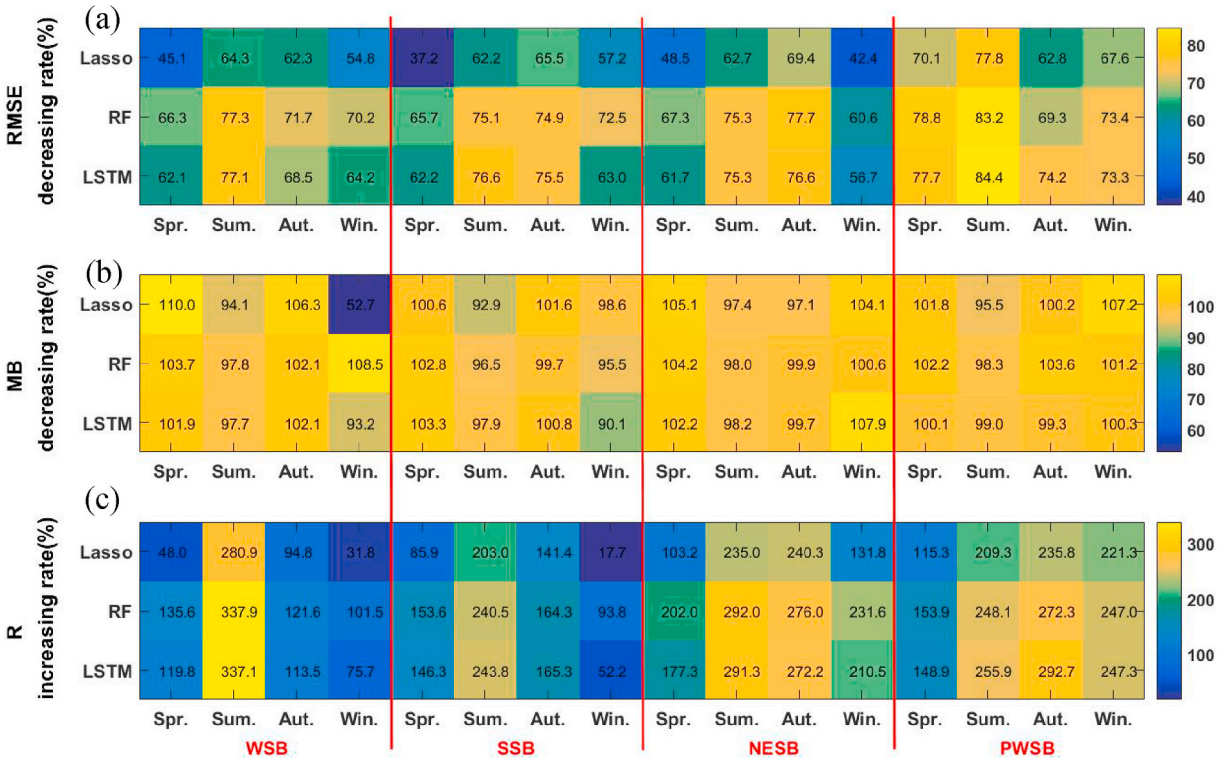

Fig. 11. Decreasing rate of RMSE, MB and increasing rate of R for ozone adjusting effects of the three machine learning methods over four sub-regions in different seasons.

relationship, while LSTM-RNN is more suitable for capturing non-linear relationship of periodic time series (Breiman 2001; Schmidhuber 2015).

3.2.3. Evaluation in different seasons

Fig. 11 illustrates the variation rate of statistical factors of ozone forecasts by CMAQ alone and the coupled models, and reveals the adjusting effects of the machine learning coupled models over the Sichuan-Chongqing region in different seasons. As shown in the figure, there were the increasing of R and the decreasing of RMSE by using the machine learning methods. Obviously, adjusting results can significantly revise deviations of ozone concentration prediction by using CMAQ in different seasons and over the sub-regions. Moreover, adjusting effects for four seasons and different sub-regions varied among the three adjusting methods. Specifically, in spring and winter of the three basin regions, the decreasing rates of RMSE were about 35–70% after Lasso regression adjusting. The adjusting results were better with random forest and LSTM-RNN, with higher R increasing rate and RMSE decreasing 60–80%. In contrast, adjusting effects were better in summer and autumn. RMSE decreased by above 60% after Lasso regression adjusting. Meanwhile, R increasing rate was significantly higher than Lasso regression, RMSE decreased above 75% by using random forest and LSTM-RNN methods. Additionally, adjusting result of random forest was superior to LSTM-RNN in spring and winter, while they were similar in summer and autumn. Adjusting effect in PWSB was generally better than that in the basin regions, with the higher R increasing rate. The decrease in RMSE of ozone prediction in PWSB was more significant in summer compared to other three seasons. LSTM-RNN model had the best performance in adjusting prediction of summer and autumn ozone. In spring and winter, the results from LSTM-RNN was similar to those from random forest, and better than those from Lasso regression.

4. Summary and conclusion

In this study, three machine learning algorithms (Lasso regression, random forest and LSTM-RNN) were used to adjust hourly ozone concentration prediction of WRF-CMAQ alone in the Sichuan-Chongqing region. Prediction and observation of atmospheric composition and meteorological data from 2018 were collected and processed to construct the basis data set. We divided the Sichuan-Chongqing region into 4 sub-regions, and obtained the training and testing data sets for each sub-region through random selection. The training data set was used to optimize and train the model and the testing data set was used to evaluate the adjusting effects of the three methods. Among them, the Lasso regression and random forest model were adopted to realize the feature optimization in different sub-regions. Results of feature optimization showed that factors having the most significant influence on the hourly ozone prediction varied with the different sub-regions. Machine learning alone and the machine learning coupled with CMAQ model, with various feature optimization schemes were compared. As a result, the coupled model with L-R coupled feature selection scheme had the best performance.

Regional testing results illustrated that adjusting through the three machine learning coupled with CMAQ models can reduce hourly ozone concentration prediction deviations and improve the correlation coefficients of CMAQ forecasts. In the three basin regions of WSB, SSB and NESB, R reached approximately 80% and RMSE decreased nearly 60% when using Lasso adjusting, while R reached approximately 90% and RMSE decreased approximately 70% by using the machine learning methods of random forest and LSTM-RNN. The results from random forest were slightly better than those from LSTM-RNN. However, in the PWSB region with plateau topography, LSTM-RNN had the best performance, with R of 91% and the reduction rate of 80.2% for RMSE. Testing results of different stations were commonly similar to regional testing. In the 18 cities located in Sichuan basin, adjusting of Lasso reduced RMSE by 50–70%. Adjusting of random forest and LSTM-RNN was better than that of Lasso, with the decreasing rate of 70–80% for RMSE. Additionally, random forest had the best performance in basin cities. However, in the other four cities located on the edge of the plateau, LSTM-RNN produced more satisfactory results. Testing results varied for different seasons and sub-regions. In the three basin regions, the adjusting effects of the three methods were more significant in summer and autumn. In PWSB, the adjusting effects were better in summer.

Author contributions

Hua Lu, Conceptualization, Methodology, Validation, Writing – original draft, Min Xie, Methodology, Writing – review & editing, Project administration, Xiaoran Liu, Writing – review & editing, Bojun Liu, Methodology, Minzhi Jiang, Methodology, Yanghua Gao, Project administration, Xiaoli Zhao, Data curation

Funding

This work was supported by the National Key Research and Development Program of China, China (2018YFC0213502), Chongqing Science and Technology Commission key research and development project, China (cstc2019jscx-tjsbX0007), the open research fund of Chongqing Meteorological Bureau, China (KFJJ-201607), the Innovation team project of Chongqing Meteorological Bureau, China (ZHCXTD-202003; ZHCXTD-202023), and Chongqing Science and Technology Commission technology innovation and application demonstration project, China (cstc2018jszx-zdyfxmX0003).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to express our gratitude to China National Environmental Monitoring Centre and China Meteorological Bureau to provide the monitoring data.

References

Appel, K.W., Bhave, P.V., Gilliland, A.B., Sarwar, G., Roselle, S.J., 2008. Evaluation of the community multiscale air quality (CMAQ) model version 4.5: sensitivities impacting model performance; Part II—particulate matter. Atmos. Environ. 42, 6057–6066.

Arhami, M., Kamali, N., Rajabi, M., 2013. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. Environ. Sci. Pollut. Res. Int. 20 (7), 4777–4789.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., Di Carlo, P.J.A.P.R., 2017. Recursive neural network model for analysis and forecast of PM10 and PM2.5, 2017 Atmospheric Pollution Research 652–659.

Binkowski, F., Roselle, S.J., 2003. Models-3 Community Multiscale Air Quality (CMAQ) model aerosol component 1. Model description. J. Geophys. Res. 108 (d6), 4183.

Borrego, C., Monteiro, A., Pay, M., Ribeiro, I., Miranda, A.I., Basart, S., Baldasano, J., 2011. How bias-correction can improve air quality forecasts over Portugal. Atmos. Environ. 45, 6629–6641.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cai, H., Gui, K., Chen, Q., 2018. Changes in haze trends in the sichuan-chongqing region, China, 1980 to 2016. Atmosphere 9 (7), 277.

Djalalova, I., Delle Monache, L., Wilczak, J., 2015. $PM_{2.5}$ analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model. Atmos. Environ. 119 (Oct.), 431–442.

Dudhia, J., 2001. Coupling an advanced land surface–hydrology model with the penn state–NCAR MM5 modeling system. Part I: model implementation and sensitivity. Mon. Weather Rev. 129, 569–585.

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J.J.A.E., 2015. Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. 107, 118–128.

Ferreira, J., Carvalho, A.c., Keizer, J., 2013. On Re-initialization Methods and Spin-Up Periods Effects on WRF Precipitation Diagnostics. PURDUE UNIVERSITY.

Freeman, B., Taylor, G., Gharabaghi, B., Thé, J., 2017. Forecasting air quality time series using deep learning. J. Air Waste Manag. Assoc. 68 (8), 866–886.
Halevy, A., Norvig, P., Pereira, F., 2009. The unreasonable effectiveness of data. IEEE Intell. Syst. 24 (2), 8–12.
Hong, S.-Y., Noh, Y., Dudhia, J., 2006. A new vertical diffusion package with an explicit treatment of entrainment processes. Mon. Weather Rev. 134 (9), 2318–2341.
Huang, J., McQueen, J., Wilczak, J., Djalalova, I., Stajner, I., Shafran, P., Allured, D., Lee, P., Pan, L., Tong, D., Huang, H.-C., DiMego, G., Upadhayay, S., Monache, L., 2016. Improving NOAA NAQFC $PM_{2.5}$ predictions with a bias correction approach. Weather Forecast. 32 (2), 407–421.
Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D., Liu, Y., 2018. Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain. Environ. Pollut. 242 (Nov), 675–683.
Jasaitis, D., Vasiliauskienė, V., Chadyšienė, R., Pečiulienė, M., 2016. Surface ozone concentration and its relationship with UV radiation, meteorological parameters and radon on the eastern coast of the baltic sea. Atmosphere 2 (7), 27.
Jerez, S., López-Romero, J., Turco, M., Lorente-Plazas, R., Gómez-Navarro, J., Jimenez-Guerrero, P., Montávez, J., 2020. On the spin-up period in WRF simulations over europe: trade-offs between length and seasonality. J. Adv. Model. Earth Syst. 4 (12), 1–18.
Jia, R., Liu, Y., Chen, B., Zhijuan, Z., Huang, J., 2015. Source and transportation of summer dust over the Tibetan Plateau. Atmos. Environ. 123 (Dec), 210–219.
Kang, J., Wang, H., Yuan, F., Wang, Z., Huang, J., Qiu, T., 2020. Prediction of precipitation based on recurrent neural networks in jingdezhen, jiangxi province, China. Atmosphere 3 (11), 246.
Kumar, J., Goomer, R., Singh, A.K., 2018. Long short term memory recurrent neural network (LSTM-RNN) based workload forecasting model for cloud datacenters. Procedia Computer Science 125, 676–682.
Lightstone, S., Moshary, F., Gross, B., 2017. Comparing CMAQ forecasts with a neural network forecast model for $PM_{2.5}$ in New York. Atmosphere 8 (9), 161.
Ma, J., Chu, B., Liu, J., Liu, Y., Zhang, H., He, H., 2017. NO promotion of SO_2 conversion to sulfate: an important mechanism for the occurrence of heavy haze during winter in Beijing. Environ. Pollut. 233, 662–669.
Maji, K.J., Ye, W.F., Arora, M., Nagendra, S.M.S., 2019. Ozone pollution in Chinese cities: assessment of seasonal variation, health effects and economic burden. Environ. Pollut. 247, 792–801.
McKeen, S., Grell, G., Peckham, S., Wilczak, J., Djalalova, I., Hsie, E.Y., Frost, G., Peischl, J., Schwarz, J., Spackman, R., Holloway, J., de Gouw, J., warneke, C., Gong, w., Bouchet, V., Gaudreault, S., Racine, J., McHenry, J., McQueen, J., Mathur, R., 2009. An evaluation of real-time air quality forecasts and their urban emissions over Eastern Texas during the summer of 2006 Second Texas Air Quality Study field study. J. Geophys. Res. Atmos. D7 (114), 1–26.
Mlawer, E., Taubman, S., Brown, P., Iacono, M., Clough, S., 1997. Radiative transfer for inhomogeneous atmospheres: RRTM, A validated correlated-k model for the longwave. J. Geophys. Res. 102, 16663–16682.
Mok, K.M., Miranda, A.I., Yuen, K.-V., Hoi, K., Monteiro, A., Ribeiro, I., 2017. Selection of bias correction models for improving the daily PM10 forecasts of WRF-EURAD in Porto, Portugal. Atmospheric Pollution Research 4 (8), 628–639.
Monin, A., Obukhov, S., 1954. Basic laws of turbulent mixing in the surface layer of the atmosphere. Tr. Akad. Nauk. SSSR Geophiz. Inst. 24 (151), 163–187.
Ning, G., Wang, S., Yim, S.H.L., Li, J., Hu, Y., Shang, Z., Wang, J., Wang, J., 2018. Impact of low-pressure systems on winter heavy air pollution in the northwest Sichuan Basin, China. Atmos. Chem. Phys. 18 (18), 13601–13615.
Qiao, X., Guo, H., Wang, P., Tang, Y., Wenye Deng, H.Z.J.A., Research, A.Q., 2019. Fine particulate matter and ozone pollution in the 18 cities of the Sichuan basin in southwestern China: model performance and characteristics. Aerosol and air quality research 10 (19), 2308–2319.
Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Network. 61, 85–117.
Skamarock, W.C., 2008. A Description of the Advanced Research WRF Version 3. *A Description of the Advanced Research WRF Version 3*.
Su, X., Junlin, A., Zhang, Y., Zhu, P., Zhu, B., 2020. Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. Atmospheric Pollution Research 6 (11), 51–60.
Tao, W., Xue, L., Brimblecombe, P., Lam, Y., Li, L., Zhang, L., 2016. Ozone pollution in China: a review of concentrations, meteorological influences, chemical precursors, and effects. Sci. Total Environ. 575, 1582–1596.
Tian, M., Wang, H., Chen, Y., Zhang, L., Shi, G., Liu, Y., Yu, J., Zhai, C., Wang, J., Yang, F., 2016. Highly time-resolved characterization of water-soluble inorganic ions in $PM_{2.5}$ in a humid and acidic mega city in Sichuan Basin, China. Sci. Total Environ. 580, 224–234.
Tibshirani, R., 2011. Regression shrinkage selection via the LASSO. J. Roy. Stat. Soc. B 3 (73), 273–282.
Wang, H., Tian, M., Chen, Y., Shi, G., Liu, Y., Yang, F., Zhang, L., Deng, L., Yu, J., Peng, C., Cao, X., 2017a. Seasonal characteristics, formation mechanisms and geographical origins of $PM_{2.5}$ in two megacities in Sichuan Basin, China. Atmos. Chem. Phys. 2 (17), 865–881.
Wang, K., Gao, J.J., Tian, H.Z., Dan, M., Yue, T., Xue, Y., Zou, P., Wang, C., 2017b. An emission inventory spatial allocatemethod based on POI data. China Environ. Sci. 6 (37), 2377–2382.
Wang, N., Lyu, X., Deng, X., Guo, H., Deng, T., Li, Y., Changqin, Y., Li, F., Wang, S.Q., 2016. Assessment of regional air quality resulting from emission control in the Pearl River Delta region, southern China. Sci. Total Environ. 573, 1554–1565.
Wang, Y., Gao, W., Wang, S., Tao, S., Gong, Z., Ji, D., Wang, L., Liu, Z., Yanfeng, H., Tian, S., Li, J., Mingge, L., Yang, Y., Chu, B., Petäjä, T., Kerminen, V.-M., He, H., Hao, J., Zhang, Y., 2020. Contrasting trends of $PM_{2.5}$ and surface-ozone concentrations in China from 2013 to 2017. National Science Review 8, 1331–1339.
Wong, D., Pleim, J., Mathur, R., Binkowski, F., Otte, T., Gilliam, R., Pouliot, G., Xiu, A., Young, J., Kang, D., 2011. WRF-CMAQ two-way coupled system with aerosol feedback: software development and preliminary results. Geosci. Model Dev. Discuss. (GMDD) 2 (5), 299–312.
Wu, Z., Xie, M., Gao, Y., Lu, H., Zhao, L., Gao, S., 2018. Inversion of SO_2 emissions over chongqing with ensemble square root kalman filter. Research of Environmental Sciences 31, 25–33.
Xie, M., Liao, J., Wang, T., Zhu, K., Zhuang, B., Han, Y., Li, M., Li, S., 2016a. Modeling of the anthropogenic heat flux and its effect on regional meteorology and air quality over the Yangtze River Delta region, China. Atmos. Chem. Phys. 16, 6071–6089.
Xie, M., Shu, L., Wang, T., Liu, Q., Gao, D., Li, S., Zhuang, B., Han, Y., Li, M.-m., Chen, P., 2016b. Natural emissions under future climate condition and their effects on surface ozone in the Yangtze River Delta region, China. Atmos. Environ. 150, 162–180.
Xie, M., Zhu, K., Wang, T., Chen, P., Han, Y., Li, S., Zhuang, B., Shu, L., 2016c. Temporal characterization and regional contribution to O_3 and NOx at an urban and a suburban site in Nanjing, China. Sci. Total Environ. 551–552, 533–545.
Xie, M., Zhu, K., Wang, T., Yang, H., Zhuang, B., Li, S., Li, M., Zhu, X., Ouyang, Y., 2014. Application of photochemical indicators to evaluate ozone nonlinear chemistry and pollution control countermeasure in China. Atmos. Environ. 99, 466–473.
Yarwood, G., Rao, S., Yocke, M., Whitten, G., 2005. Updates to the Carbon Bond Chemical Mechanism: CB05 Final Report to the US EPA. RT-0400675.
Yin, Z., Wang, H., 2017. Role of atmospheric circulations on haze pollution in december 2016. Atmos. Chem. Phys. 18 (17), 11673–11681.
Zamani, M., 2019. $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere 10 (7), 373.
Zhan, C.-c., Xie, M., Fang, D.-x., Wang, T., Wu, Z., Lu, H., Li, M.-m., Chen, P., Zhuang, B.-l., Li, S., Zhang, Z.-q., Gao, D., Reng, J.-y., Zhao, M., 2019. Synoptic weather patterns and their impacts on regional particle pollution in the city cluster of the Sichuan Basin, China. Atmos. Environ. 208, 34–47.
Zhan, C., Xie, M., Huang, C., Liu, J., Wang, T., Xu, M., Ma, C., Yu, J., Jiao, Y., Li, M., Li, S., Zhuang, B., Zhao, M., Nie, D., 2020. Ozone affected by a succession of four landfall typhoons in the Yangtze River Delta, China: major processes and health impacts. Atmos. Chem. Phys. 20, 13781–13799.
Zhang, H., Zhang, S., Wang, P., Qin, Y., Wang, H., 2017. Forecasting of PM 10 time series using wavelet analysis and wavelet-ARMA model in Taiyuan, China. J. Air Waste Manag. Assoc. 67 (7), 776–788.
Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part I: history, techniques, and current status. Atmos. Environ. 60, 632–655.
Zhao, S., Yu, Y., Qin, D., Yin, D., Dong, L., He, J., 2018. Analyses of regional pollution and transportation of $PM_{2.5}$ and ozone in the city clusters of Sichuan Basin, China. Atmospheric Pollution Research 2 (10), 374–385.
Zhao, S., Yu, Y., Yin, D., Qin, D., He, J., Dong, L., 2017. Spatial patterns and temporal variations of six criteria air pollutants during 2015 to 2017 in the city clusters of Sichuan Basin, China. Sci. Total Environ. 624, 540–557.
Zhong, J., Zhang, X., Yunsheng, D., Wang, Y., Liu, C., Wang, J., Zhang, Y., Che, H., 2018. Feedback effects of boundary-layer meteorological factors on cumulative explosive growth of $PM_{2.5}$ during winter heavy pollution episodes in Beijing from 2013 to 2016. Atmos. Chem. Phys. 18, 247–258.