Application of the XGBoost Machine Learning Method in PM_{2.5} Prediction: A Case Study of Shanghai

Jinghui Ma^{1,2,3}, Zhongqi Yu^{2,3*}, Yuanhao Qu^{2,3}, Jianming Xu^{2,3,4}, Yu Cao^{2,3}

¹ Fudan University, Shanghai 200433, China

² Shanghai Typhoon Institute, Shanghai Meteorological Service, Shanghai 200030, China

³ Shanghai Key Laboratory of Meteorology and Health, Shanghai Meteorological Service, Shanghai 200030, China

⁴ Anhui Province Key Laboratory of Atmospheric Science and Satellite Remotes Sensing, Hefei, 230000, China

ABSTRACT

Air quality forecasting is crucial to reducing air pollution in China, which has detrimental effects on human health. Atmospheric chemical-transport models can provide air pollutant forecasts with high temporal and spatial resolution and are widely used for routine air quality predictions (e.g., 1–3 days in advance). However, the model's performance is limited by uncertainties in the emission inventory and biases in the initial and boundary conditions, as well as deficiencies in the current chemical and physical schemes. As a result, experimentation with several new methods, such as machine learning, is occurring in the field of air quality forecasting. This study combined hourly PM_{2.5} mass concentration forecasts from an operational air quality numerical prediction system (WRF-Chem) at the Shanghai Meteorological Service (SMS) with comprehensive near-surface measurements of air pollutants and meteorological conditions to develop a machine learning model that estimates the daily PM_{2.5} mass concentration in Shanghai, China. With correlation coefficients that are higher by 50–100% and a standard deviation that is lower by 14–24 $\mu\text{g m}^{-3}$, the machine learning model provides significantly better daily forecasting of PM_{2.5} than the WRF-Chem model. Thus, this research offers a new technique for enhancing air quality forecasting in China.

Keywords: XGBoost algorithm; PM_{2.5}; WRF-Chem; Machine learning.

INTRODUCTION

Accurate air quality forecasting is important for both severe air pollution response and self-protection of human health (Bedoui *et al.*, 2016). However, air quality forecasting is rather complicated and dominated by meteorological conditions, and emission inventory. Thus large uncertainties still exist in the current ambient air quality forecasting which does not meet the requirements of current air pollution mitigation in China.

There are several approaches commonly used to predict ambient air quality: the numerical model forecasting method and the statistical forecasting method. In addition, numerical forecast modeling requires detailed emissions, and users need a deep understanding of the transformation mechanisms of various air pollutants to enable the selection of suitable physical and chemical schemes that are used in the model's configuration (Yumimoto and Uno, 2006). However, it is

difficult to accurately describe the spatial and temporal variations of urban pollutant emissions and to completely quantify them within the model. To improve the simulation accuracy of air quality models, Xu *et al.* (2008) found that the application of air pollutant measurements can effectively reduce the bias of the emission data and developed new method for estimating air pollution emissions based on a Newtonian relaxation and nudging technique. Just *et al.* (2018) demonstrated how machine learning technique with quality control and spatial features substantially improves satellite-derived AOD for air pollution modeling. Current numerical model predictions still have considerable deviations when making predictions on specific regions. The main reasons include: prediction deviation of numerical model on synoptic system, the model cannot describe real-time pollution emissions and errors in the parameterization scheme of the numerical model itself.

The statistical forecasting method is relatively simple, economical, and easy to be implemented. However, the forecast effect is related to the quantity of variables and available data, and the statistical correlation between predictions and predictors varies with respect to predictors' change. Nevertheless, non-linear regression prediction performance of the machine learning-based statistical prediction method

* Corresponding author.

Tel.: (86)021-54896603

E-mail address: yuzhongqi0429@163.com

is superior to that of traditional statistical methods (Chang *et al.*, 2008). Machine learning makes a few assumptions about data, and the results are checked by cross-validation. It removes the classical statistical process, which consists of hypothesis distribution, a mathematical model fitting, hypothesis testing and determination of the P-value. The model prediction based on machine learning algorithms or programs performs well, and the results of cross-validation are readily understood by applicators.

In this respect, the Extreme Gradient Boosting algorithm (XGBoost) was experimented in air quality forecasting. This method is an integrated learning model introduced by Chen *et al.* (2016) from the University of Washington in 2016 (Friedman *et al.*, 2001), and has been widely used in the fields of finance (Wang *et al.*, 2018; Yao *et al.*, 2018), industry (Sun *et al.*, 2018), energy (Li *et al.*, 2018; Torres *et al.*, 2018; Zhang *et al.*, 2018), medicine (Torlay *et al.*, 2017; Hong *et al.*, 2018; Shimoda *et al.*, 2018; Taylor *et al.*, 2018; Turki, 2018; Zhong *et al.*, 2018), traffic (Lin *et al.*, 2018) and the internet (Verma *et al.*, 2018; Zhang *et al.*, 2018). Pan (2018) has applied the XGBoost algorithm to predict hourly PM_{2.5} concentrations in China and compared it with the results from the random forest, the support vector machine, linear regression and decision tree regression, and demonstrated the best performance of the XGBoost algorithm in air quality forecasting.

Shanghai is a mega city in eastern China with a large and highly dense population. It is extremely important to conduct accurate PM_{2.5} forecasting in Shanghai. The current WRF-Chem operational model of the Shanghai Meteorological Service is a mesoscale atmospheric dynamical-chemical coupled online model that was developed by the National Center for Atmospheric Research, the United States Pacific Northwest National Laboratory, the United States National Oceanic and Atmospheric Administration, and other departments. It has been widely used to conduct air quality prediction in China. However, there are still large uncertainties in model predictions. Great efforts have been made by scientists to improve model capacity, including data assimilation, and emission inventory adjustment (Bedoui *et*

al., 2016). In this study, a new model for PM_{2.5} prediction was established using the machine learning XGBoost algorithm and the Lasso linear regression technique (to reduce model over-fitting) based on WRF-Chem outputs and air pollutants and meteorological observations. The new model used two algorithms (XGBoost and Lasso) and was named “the modified XGBoost model” in this study. The modified XGBoost model prediction performance was also compared with the predictions of the Lasso model and the WRF-Chem model.

METHODS

Introduction of the WRF-Chem Numerical Model System

The Regional Atmospheric Environmental Modeling System (RAEMS) for eastern China of SMS was centered at 31.5°N, 118°E with a horizontal resolution of 6 km (Fig. 1). The RADM2 mechanism was used for gas-phase chemistry and the ISORROPIA dynamic equilibrium inorganic aerosol mechanism and the SORGAM organic aerosol mechanism were used for aerosol chemistry. The physical scheme applications were presented in Table 1. The National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) data were used for the initial and boundary meteorological conditions for WRF-Chem model. The previous 24 h prediction was used as the initial chemical condition. Gaseous chemical boundary conditions were based on monthly averages from global chemical transport model MOZART (Jordan *et al.*, 1997). The MEGAN2 model was implemented to calculate the biogenic emissions. The integrated time step was set as 30 s for meteorology, and 60 s for chemistry. The Emission Inventory for China (MEIC) with 0.25° resolution of 2010 from Tsinghua University was applied. Depending on the monitoring results of each industry in Shanghai, emissions were also allocated hourly with the diurnal profile.

Data Introduction

Modeling and evaluation data used in this work covered the period of January 1, 2015, to December 31, 2018. The

Fig. 1. Coverage area of atmospheric environment numerical forecast system in eastern China (Inside the black box), the red dot is the location of Shanghai.

Table 1. Physical scheme options for WRF-Chem.

Physics	Physical scheme
Micro-physics	WSM 6
Cumulus parameterization	Not used
Long wave radiation	RRTM
Short wave radiation	Dudhia
Surface layer	Monlin_Obukhov
Land surface	Unified Noah
Boundary layer	YSU

air pollutant measurements were taken from the National Urban Air Quality Real-Time Release Platform (<http://113.108.142.147:20035/emcpublish/>). The meteorological measurements, including hourly atmospheric pressure (P), air temperature (T), relative humidity (RH), precipitation (Prs), wind direction (Wind_D) and wind speed (Wind_S), were obtained from the meteorological bureau's national ground-based observational stations. Both meteorological and chemical forecasted variables were collected from the RAEMS outputs. WRF high altitude weather forecast data included meteorological variables at five standard layers (500 hPa, 700 hPa, 850 hPa, 925 hPa and 1000 hPa), among which about 5% of data were missing. In this study, only validated observed data and forecast data were used for evaluation.

Data Pre-processing

Because the geographical locations of air quality observational stations and meteorological observatories did not co-located, the air quality observational stations were accompanied with the nearest meteorological observatories. The input for the training set of the modified XGBoost model was the meteorological observational data of the current day, the air quality observational data for the previous day, the 24 h meteorological factors and the PM_{2.5} forecast data outputs from the WRF-Chem model. The final outputs of the modified XGBoost model was PM_{2.5} predictive value.

XGBoost Model Introduction

The machine learning algorithm used in this study was the GBDT (Gradient Boosting Decision Tree), which was an iterative decision tree algorithm composed of a plurality of decision trees (Friedman *et al.*, 2001), namely by iterating multiple trees together to make final decisions. Compared with the logistic regression, which can only be used for linear regression, the GBDT can be used widely for almost all regression problems (linear or non-linear). It can also apply to binary classification problems. XGBoost represents an efficient GBDT algorithm enabling gradient boosting “on steroids” (it is called “Extreme Gradient Boosting” for a reason). It combines software and hardware optimization techniques perfectly and yield superior results and use fewer computing resources than other methods (Chen *et al.*, 2016).

Parallelization

XGBoost approaches the process of sequential tree building using parallelized implementation. Therefore, to increase run time, the loop order is interchanged by employing

initialization through a global scan of all instances and sorting using parallel threads.

Tree Pruning

XGBoost uses the “max_depth” parameter as specified (instead of criterion first) and starts to prune trees backward. This “depth-first” approach significantly improves its computational performance.

Hardware Optimization

The algorithm has been designed to make efficient use of hardware resources. This is achieved by caching awareness via allocating internal buffers in each thread to store gradient statistics. Additional enhancements such as “out-of-core” computing enable optimization of available disk space while handling massive data frames that do not match into the memory.

In addition, XGBoost contains algorithmic enhancements as follows.

Regularization

It penalizes more complex models through both LASSO (L1) and Ridge (L2) regularization to avoid over-fitting.

Sparsity Awareness

XGBoost naturally admits sparse features for input by automatically “learning” best missing value depending on training loss, and it handles different types of sparsity patterns in the data more efficiently.

Weighted Quantile Sketch

XGBoost employs the distributed weighted quantile sketch algorithm to effectively find the optimal split points among weighted datasets.

Cross-validation

The algorithm comes with a built-in cross-validation method at each iteration, which negates the need to explicitly program this search and to indicate the exact number of boosting iterations required in a single run.

With respect to machine learning, it is not sufficient to select the appropriate algorithm for use, and it is also necessary to choose the correct configuration of the algorithm for a dataset by tuning the hyper-parameters. There are likewise several other factors to consider when selecting a winning algorithm, such as computational complexity, applicability and ease of implementation.

Construction of the Model

In order to reduce the risk of model over-fitting, Lasso regression model (Tibshirani, 1996) was used to analyze the importance of forecasting factors by retaining 36 most important factors, as showed in Table 2. Parameter details for model feature selection were shown in Table 3. The XGBoost model was trained and historical data observation and WRF-Chem prediction factors were treated as the inputs of the model. The modeling process was illustrated in Fig. 2.

The significant features were extracted from the forecasting results of the pollutant concentrations and meteorological

Table 2. Predictive factors screened by Lasso regression model.

1. PM25(WRF_chem)	2. SO2(WRF_chem)	3. O3(WRF_chem)
4. PM25(Pre_1h)	5. Td_850(Pre_72h)	6. Td_850(WRF_chem)
7. So2(Pre_1h)	8. NO2(Pre_1h)	9. PM25(Pre_5h)
10. Rh_1000(Pre_24h)	11. PM25(Pre_96h)	12. CO(Pre_4h)
13. SO2(Pre_4h)	14. Tc_500	15. NO2(WRF_chem)
16. PM10(Pre_1h)	17. PM25(Pre_48h)	18. Rh_700(WRF_chem)
19. Rhu(WRF_chem)	20. Z_1000(WRF_chem)	21. Prs(Pre_1h)
22. Tem(Pre_5h)	23. Tc_1000(Pre_48h)	24. Rh_850(WRF_chem)
25. CO(Pre_3h)	26. Rh_925(Pre_72h)	27. O3(Pre_5h)
28. Rh_500(Pre_24h)	29. PM10(Pre_48h)	30. Td_1000(WRF_chem)
31. Rhu(Pre_3h)	32. Rhu(Pre_5h)	33. O3(Pre_48h)
34. Tc_700(Pre_24h)	35. Td_700-Td_500 (WRF_chem)	36. Tc_925-Tc_850(WRF_chem)

Note: Pre is the predictive factor before the model start time.

Table 3. XGBoost model parameters

Parameter name	Parameter values
Learning rate	0.05
Sample subsampling rate	0.85
Characteristic subsampling rate	0.875
Maximum tree depth	7
L1 Regular Coefficient	0
L2 Regular Coefficient	1

factors (such as T, RH, P and WIND_S) at different standard altitude layers from WRF-Chem. First, the WRF-Chem forecast data were directly treated as the basic forecasting factors, which were all independent at different layers. The distribution of each factor between different layers could reflect the stable state of the atmosphere, which directly affected the vertical diffusion of PM_{2.5}. Therefore, the differences between different layers of the same factor were used as derived factors to represent the variation of the features in the vertical direction. These factors were used as input variables for the XGBoost model.

Selection of Prediction Factor

In order to implement multi-step prediction using XGBoost, the 24 h data needed to be input into the model as one sample, which would lead to the significant reduction in the number of training samples. If the structure of the model was excessively complicated, then over-fitting easily occurred, which resulted in insufficient generalization of the model. Therefore, it was necessary to screen the above features to retain more important factors. The Lasso regression analysis method (Tibshirani, 1996) was prepared on the basis of the principle of the least square method. Its core idea is to regularize the parameter items while minimizing the sum of residual squares and to control the sum of absolute values of each parameter within an acceptable range by using the L1 norm. Formulas for calculating linear regression equations are described in detail in Eq. (1):

$$y = X\beta + \varepsilon \quad (1)$$

where $y = (y_1, y_2, \dots, y_n)^T$, $X = (x_1, x_2, \dots, x_d)^T$, $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$, $j = 1, 2, d$. β and ε are the parameters to be estimated

and the residual of the model, respectively. The Lasso model parameters are available from Eq. (2):

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \|y - X\beta\|^2 + \lambda \|\beta\|_1 \quad (2)$$

where the term $\lambda \|\beta\|_1$ is the L1 regularization term, of which the function is to limit the range of parameters and reduce the possibility of model over-fitting. Thus, we fitted the output of WRF-Chem by using the Lasso regression model first. λ is the regularization coefficient, which directly affected the complexity of the model. If λ is too low, it would cause excessive coefficients to become zero, thereby resulting in under-fitting of the model. If λ is too large, it would result in less influence of the regularization term, thereby resulting in over-fitting of the model, so it is necessary to select a reasonable value for λ . In this study, 30% of the samples from January 2015 to October 2017 composed the verification set by random sampling. The relationship between the prediction errors of the Lasso model and λ was analyzed (Fig. 3). With the increase in λ , the model prediction error first decreased and then increased, and the number of non-zero model coefficients decreased. When λ equaled 0.0001, the prediction error of the model was the smallest and the number of non-zero coefficients of the model was 36. Therefore, 0.0001 was selected as the regularization coefficient of the Lasso regression model. Thus, only 36 factors of non-zero coefficients were retained as the prediction factors (Table 2).

Model Training

70% of the data were randomly selected out of the samples from January 2015 to October 2017, which composed the training set, and 30% of the remaining data was used as the verification set for the parameter adjustment of the prediction model. Meanwhile, the data from November 1, 2017, to December 31, 2018, were used as the verification set to evaluate the final prediction of the model. There were many critical parameters that were required to be adjusted in the XGBoost model. The line search optimization of each parameter was conducted based on the accuracy of the model in the verification set, and the decisive parameters determined were given in Table 3.

Fig. 2. Modeling flow chart.

Fig. 3. The relationship between the characteristic number of the coefficient non-zero, the model prediction error and the regularization term coefficient curve.

Data Evaluation Methods

In order to quantitatively evaluate the prediction accuracy of the model, mean bias (MB), mean error (ME), root mean square error (RMSE) and correlation coefficient (R) were calculated and the calculations of MB, ME, RMSE and R are shown in Eqs. (3)–(6):

$$MB = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y}_i) \quad (3)$$

$$MB = \frac{1}{N} \sum_{i=1}^N |y_i - \bar{y}_i| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y}_i)^2} \quad (5)$$

$$R = \frac{N \sum y_i \bar{y}_i - \sum y_i \sum \bar{y}_i}{\sqrt{N \sum y_i^2 - (\sum y_i)^2} \sqrt{N \sum \bar{y}_i^2 - (\sum \bar{y}_i)^2}} \quad (6)$$

where N is the number of samples and y_i and $\bar{y}_i$ are the observed and predicted values of i samples, respectively.

The prediction result of the modified XGBoost model and the residual of the true value is shown in Eq. (7):

$$error = y_{fore} - y_{obs} \quad (7)$$

RESULTS AND DISCUSSION

Analysis of PM_{2.5} Concentration Forecast Results

Based on the constructed modified XGBoost model, pollution observation data at different environmental observation stations, meteorological data at the corresponding

meteorological observation stations and WRF-Chem output data were used as the sample data for model training for different environmental observation stations. In order to quantitatively evaluate the prediction accuracy of the model, the 25th percentile, 75th percentile, median, mean, MB, ME, RMSE and R were calculated.

For analyzing the PM_{2.5} prediction effect of the modified XGBoost model, the modified XGBoost model performance was compared with those of the Lasso model and WRF-Chem model. As showed in Fig. 4, the WRF-Chem prediction had large fluctuation and the peak and valley values of the WRF-Chem prediction were in imperfect agreement with the observed values. Compared with the WRF-Chem and Lasso models, the PM_{2.5} concentration prediction results of the modified XGBoost model had better consistency with the observations.

From Fig. 4, it could be seen that the predicted values of PM_{2.5} concentration of the Lasso, modified XGBoost and WRF-Chem models were consistent with the observed

values in the forecast time series. The modified XGBoost model better reflected the variations of the observations over time and avoided the false peaks and valleys of the WRF-Chem model prediction to a certain extent.

Scatter plots, which can directly reflect the linear relationship between simulated and observed values, were used to compare the consistency between the prediction results of the three different models and the observation data. The scatter distribution of the observation and forecasted points was illustrated in Fig. 5. Compared with the WRF-Chem model, scatter distribution of the Lasso and modified XGBoost models was concentrated diagonally, which showed that the two models had a better corrective effect on the WRF-Chem model prediction.

From the Taylor plot (Fig. 6), the R values of the three models were 0.51 (WRF-Chem), 0.73 (Lasso) and 0.77 (modified XGBoost), and the standard deviations were $6.0 \mu\text{g m}^{-3}$, $5.6 \mu\text{g m}^{-3}$ and $5.0 \mu\text{g m}^{-3}$, respectively. When the observed PM_{2.5} concentration was greater than $50 \mu\text{g m}^{-3}$,

Fig. 4. Comparison between three model predictions and observational data (a) hourly (b) daily average.

Fig. 5. Plots of predicted and observational data (a) WRF-Chem model (b) Lasso model (c) modified XGBoost model.

the R values of the three models were 0.40 (WRF-Chem), 0.50 (Lasso) and 0.60 (modified XGBoost), and the standard deviations were $6.7 \mu\text{g m}^{-3}$, $5.3 \mu\text{g m}^{-3}$ and $5.1 \mu\text{g m}^{-3}$, respectively. When the observed $\text{PM}_{2.5}$ concentration was greater than $75 \mu\text{g m}^{-3}$, the R values of the three models were 0.30 (WRF-Chem), 0.40 (Lasso) and 0.60 (modified XGBoost), and the standard deviations were $7.1 \mu\text{g m}^{-3}$, $6.1 \mu\text{g m}^{-3}$ and $5.1 \mu\text{g m}^{-3}$, respectively. Therefore, in different $\text{PM}_{2.5}$ concentration ranges, the prediction effect of the modified XGBoost model was preferable to those of the Lasso and WRF-Chem models. The modified XGBoost model increased the R values of the WRF-Chem model predictions and actual values by 51.0%, 50.0% and 100.0% in three concentration ranges (full concentration range, greater than $50 \mu\text{g m}^{-3}$ and greater than $75 \mu\text{g m}^{-3}$), and reduced the standard deviations by $16.7 \mu\text{g m}^{-3}$, $23.9 \mu\text{g m}^{-3}$ and $14.1 \mu\text{g m}^{-3}$,

respectively. For the range of high $\text{PM}_{2.5}$ concentrations, the modified XGBoost model had a stronger predictive correction ability.

Analysis of Error Sources of $\text{PM}_{2.5}$ Concentration Forecast

To analyze the error source of the modified XGBoost model, the RMSE (Fig. 7) was calculated for the hourly forecasting results in the future 24 h. From the view of the RMSE changing over time, as showed in Fig. 7, the RMSE of the modified XGBoost model was less than those of the WRF-Chem and Lasso regression models at any time. The WRF-Chem model had a large RMSE for 24 h predictions, and often indicated false peak and valley values. The modified XGBoost model could significantly reduce the RMSE of the 24 h prediction and the false peak and valley errors.

Fig. 6. Taylor plot of Shanghai PM_{2.5} 24 h forecast by WRF-Chem model, Lasso model and modified XGBoost model (r pass 95% confidence test).

Fig. 7. Variation of forecast RMSE over time.

Correlations between the modified XGBoost model residual and the actual pollution and meteorological factors were derived. The calculation results were presented in Table 4. The prediction error of the modified XGBoost model for PM_{2.5} had a negative correlation with the exact values of PM_{2.5} in an R value of -0.65 . The second largest R value was made for PM₁₀, which was -0.35 . R values between the model prediction residual and meteorological factors were clearly smaller than those of the pollutants.

Fig. 8 showed the variation curves of the true value of PM_{2.5}, the predicted value of the modified XGBoost model and the prediction error from March 20 to March 30, 2018. The modified XGBoost model successfully predicted the

peak PM_{2.5} concentration on March 23, 2018. Compare with the observation, the variations of the prediction results of the modified XGBoost model were smaller than those of the observations. The observed PM_{2.5} concentration first increased and then declined in this period. The modified XGBoost model had a good prediction ability for the whole trend, but the turning point of concentration was not forecast accurately. On one hand, the turning point of the concentration might have been due to the change in the actual pollution emissions, and it was difficult for the model to find perfect regularity in the existing data. On the other hand, for part of the data used in the modified XGBoost model was from WRF-Chem predictions, such as PM₂₅(WRF_chem),

Table 4. Correlation coefficient between the modified XGBoost model error and related factors.

	PM _{2.5}	PM ₁₀	CO	NO ₂	SO ₂	O ₃	T	Prs	P	Rh	Wind D	Wind S
r	−0.65*	−0.35*	−0.34*	−0.21	−0.10	−0.02	−0.15	0.12	−0.032	−0.06	−0.021	−0.10

Note: r marked by * have passed the 95% confidence test.

Fig. 8. The variation of modified XGBoost model prediction value, observation value and prediction error with time.

SO₂(WRF_chem), O₃(WRF_chem), Td₈₅₀(WRF_chem), Rhu(WRF_chem), Z₁₀₀₀(WRF_chem), etc., the prediction error of WRF-Chem also led to the decline in modified XGBoost forecasting accuracy.

PM_{2.5} Concentration Prediction Evaluation

In order to quantitatively analyze the prediction effects of the three models, evaluation indexes of the prediction results of different prediction models were estimated, and the consequences for observations greater than 50 $\mu\text{g m}^{-3}$ were showed in Table 5. The RMSE of the modified XGBoost model was 26.1 $\mu\text{g m}^{-3}$, which was about 41% lower than that of the WRF-Chem model. The *R* value between the modified XGBoost results and the observations reached 0.6, which was approximately 50% higher than that of the WRF-Chem model. Among three models, mean value difference between the Lasso model results and observations was the largest. For the 75th percentile, the modified XGBoost model was the closest to the observation, while the difference between the Lasso model and the observations was the largest, which indicated that the modified XGBoost model was more suitable than the other two models at high PM_{2.5} concentrations. From the view of ME, all three models overestimated the PM_{2.5} concentration. ME of the modified XGBoost model was the smallest (20.4 $\mu\text{g m}^{-3}$), ME of the WRF-Chem model was the largest (34.5 $\mu\text{g m}^{-3}$). MB of the modified XGBoost model was the smallest (3.6 $\mu\text{g m}^{-3}$), while that of the Lasso model was the largest. In addition, for the mean value and median predicted by the three models, the modified XGBoost model forecast was the closest to the observational values. The XGBoost model also had the smallest values for median deviation, 25% quantile deviation, 75% quantile deviation, mean deviation/observed

and ME/observed. In summary, under the condition that the observed value was greater than 50 $\mu\text{g m}^{-3}$, the modified XGBoost model performed the best among the three models.

To test the monthly forecasting performance of the modified XGBoost model, the monthly averaged forecast results from January 1 to December 31, 2018, were compared with WRF-Chem forecast results (Fig. 9). The forecast results were selected from 24–48 h, and the average value was regarded as the average daily concentration. The difference between the PM_{2.5} concentrations predicted by the modified XGBoost model and the observation were between −4.9 and 2.9 $\mu\text{g m}^{-3}$, which was lower than the difference between the WRF-Chem prediction and observations (−19.3 to 10.7 $\mu\text{g m}^{-3}$). The monthly average concentrations predicted by the XGBoost model and the WRF-Chem model were both consistent with the peak and valley of the observations. However, the predicted values of the two models in February, May, June, September and November were greater than observations, and the forecast values for the remaining months were smaller than the observations, especially in January and December (the WRF-Chem model and modified XGBoost forecast deviations were −14.8 $\mu\text{g m}^{-3}$, −4.9 $\mu\text{g m}^{-3}$, −9.77 $\mu\text{g m}^{-3}$ and −3.7 $\mu\text{g m}^{-3}$, respectively). Nevertheless, it was evident that the modified XGBoost forecasting model has a correcting effect on the WRF-Chem model forecast in all seasons, especially in winter.

CONCLUSIONS

We developed a modified XGBoost model that incorporated WRF-Chem forecasting data on pollutant concentrations and meteorological conditions (the important factors was shown in Table 2, which could represent the spatiotemporal

Table 5. PM_{2.5} model prediction results evaluation (Observational daily concentration greater than 50 $\mu\text{g m}^{-3}$).

	Observation	WRF-Chem	Lasso	XGBoost
Mean value	79.1	75.4	65.1	77.5
Median	73.0	66.2	66.3	66.5
25% percentile	59.0	43.4	47.9	47.6
75% percentile	91.0	95.4	80.2	87.7
Median deviation		6.8	6.7	5.9
25% quantile deviation		15.6	11.1	10.4
75% quantile deviation		−4.4	12.8	3.3
Mean bias		3.7	15.4	3.6
Average error		34.5	20.9	20.4
Mean deviation/Observed		0.3	0.2	0.1
Mean Error/Observed		0.4	0.3	0.3
Mean Square Error (RMSE)		44.6	28.1	26.1
Correlation coefficient		0.4	0.5	0.6

Fig. 9. Monthly comparison of monthly mean PM_{2.5} concentration and observed values by XGBoost and WRF-Chem model.

characteristics of pollution and meteorology) with observed variations in these two factors, thereby significantly improving the accuracy of PM_{2.5} forecasting in Shanghai, China.

All of the comprehensive evaluation indicators, including the *R* and RMSE values, confirmed that the modified XGBoost model provided more accurate predictions of high PM_{2.5} concentrations (exceeding the standard of 75 $\mu\text{g m}^{-3}$) than the WRF-Chem model. The modified model also improved on the monthly forecasts of the WRF-Chem model in every season, especially during heavy winter pollution.

Since our study was restricted to Shanghai, China, the representativeness of the modified XGBoost model and the reliability of our conclusions are limited. It will be necessary to expand the model's scope of application in future research. Furthermore, applying different machine learning algorithms for PM_{2.5} prediction in Shanghai and conducting a multi-model ensemble prediction would be useful.

ACKNOWLEDGMENTS

The research was supported by the National Key R&D Program (2016YFC0201900), Shanghai Natural Resources Fund (19ZR1462100), National Natural Resources Fund

(41475040) and Shanghai Science and Technology Commission (16DZ120460). We sincerely thank the CMA for providing access to hourly surface data. The authors are grateful to the valuable comments and suggestions of the editor and the two anonymous reviewers, which have helped us improve the paper quality.

REFERENCES

- Bedoui, S., Gomri, S., Samet, H. and Kachouri, A. (2016). A prediction distribution of atmospheric pollutants using support vector machines, discriminant analysis and mapping tools (case study: tunisia). *J. Phys. Chem. C* 114: 15516–15521.
- Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA — August 13–17, 2016, pp. 785–794.
- Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29: 1189–1232.
- Hong, W.S., Haimovich, A.D. and Taylor, R.A. (2018). Predicting hospital admission at emergency department

- trriage using machine learning. *PLoS One* 13: e0201016.
- Jordan, M.I. (1997). Serial order: A parallel distributed processing approach. *Adv. Psychol.* 121: 471–495.
- Just, C.A., De Carli, M.M., Shtein, A., Dorman, M., Lyapustin, A. and Kloog, I. (2018). Correcting measurement error in satellite aerosol optical depth with machine learning for modeling PM_{2.5} in the Northeastern USA. *Remote Sens.* 10: 803.
- Li, P. and Zhang, J.S. (2018). A new hybrid method for China's energy supply security forecasting based on ARIMA and XGBoost. *Energies* 11: 1687.
- Lin, F., Jiang, J., Fan, J. and Wang, S. (2018). A stacking model for variation prediction of public bicycle traffic flow. *Intell. Data Anal.* 22: 911–933.
- Pan, B.Y. (2018). Application of XGBoost algorithm in hourly PM_{2.5} concentration prediction. *IOP Conf. Ser.: Earth Environ. Sci.* 113: 012127.
- Shimoda, A., Ichikawa, D. and Oyama, H. (2018). Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. *Comput. Methods Programs Biomed.* 163: 39–46.
- Sun, B., Lam, D., Yang, D., Grantham, K., Zhang, T., Mutic, S. and Zhao, T. (2018). A machine learning approach to the accurate prediction of monitor units for a compact proton machine. *Med. Phys.* 45: 2243–2251.
- Taylor, R.A., Moore, C.L., Cheung, K.H. and Brandt, C. (2018). Predicting urinary tract infections in the emergency department with machine learning. *PLoS One* 13: e0194085.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B* 58: 267–288.
- Torlay, L., Perrone-Bertolotti, M., Thomas, E. and Baciú, M. (2017). Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. *Brain Inform.* 4: 159–169.
- Torres-Barrán, A., Alonso, Á. and Dorronsoro, J.R. (2019). Regression tree ensembles for wind energy and solar radiation prediction. *Neurocomputing* 326: 151–160.
- Turki, T. (2018). An empirical study of machine learning algorithms for cancer identification. 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, China.
- Verma, P., Anwar, S., Khan, S. and Mane, S.B. (2018). Network intrusion detection using clustering and gradient boosting. 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bangalore, India.
- Wang, M., Yu, J. and Ji, Z. In Shi, Z., Mercier-Laurent, E. and Li, J. (2018). Personal Credit Risk Assessment Based on Stacking Ensemble Model. *Intelligent Information Processing IX*, 2018, Springer International Publishing, Cham, pp. 328–333.
- Xu, X., Xie, L., Cheng, X., Xu, J., Zhou, X. and Ding, G. (2008). Application of an adaptive nudging scheme in air quality forecasting in China. *J. Appl. Meteorol. Climatol.* 47: 2105–2114.
- Yao, J.R., Zhang, J. and Wang, L. (2018). A financial statement fraud detection model based on hybrid data mining methods. 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, pp. 57–61.
- Yumimoto, K. and Uno, I. (2006). Adjoint inverse modeling of CO emissions over Eastern Asia using four-dimensional variational data assimilation. *Atmos. Environ.* 40: 6836–6845.
- Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B. and Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and Xgboost. *IEEE Access* 6: 21020–21031.
- Zhang, B., Yu, Y. and Li, J. (2018). Network intrusion detection based on stacked sparse autoencoder and binary tree ensemble method. 2018 IEEE International Conference on Communications Workshops (ICC Workshops), Kansas City, MO, USA.
- Zhong, R., Wu, Y., Cai, Y., Wang, R., Zheng, J., Lin, D., Wu, H. and Li, Y. (2018). Forecasting hand, foot, and mouth disease in Shenzhen based on daily level clinical data and multiple environmental factors. *Biosci Trends* 12: 450–455.

Received for review, August 23, 2019

Revised, November 18, 2019

Accepted, November 28, 2019