A Novel Bagging Ensemble Approach for Predicting Summer Time Ground Level Ozone Concentration

Journal:	*Journal of the Air & Waste Management Association*
Manuscript ID	UAWM-2018-0120
Manuscript Type:	Technical Paper—Air Pollution
Date Submitted by the Author:	18-Jun-2018
Complete List of Authors:	S, Mohan; Indian Institute of Technology Madras, Department of Civil Engineering Packiam, Saranya; Indian Institute of Technology Madras, Department of Civil Engineering
Keywords:	ensemble machine learning, Ground level Ozone, bagging trees, base learners, random forest
Abstract:	Abstract Ground level ozone is a criteria air pollutant having fatal effect on human health and surrounding environment. Formation of ground level ozone is a complex photochemical phenomenon and involves numerous intricate factors most of which are interrelated with each other. Machine learning techniques can be adopted to predict the ground level ozone. The main objective of the present study is to develop the state-of-the-art ensemble bagging approach to model the summer time ground level ozone in an industrial area comprising a hazardous waste management facility. Factors such as NO, NO2 and meteorological parameters were taken into account while modeling the ground level ozone. Multilayer perceptron, RTree, REPTree and Random forest were employed as the base learners. The error measures used for checking the performance of each model includes CC, RMSE, MAE, RRSE, R2 and IA. The model results were validated against an independent test data set. Bagged random forest predicted the ground level ozone better with higher coefficient of determination 0.9432 and with lower error rates of RRSE = 6.357; MAE =6.5774; RAE= 0.2289. This study scaffolded the current research gap in big data analysis identified with air pollutant prediction.

Implications: The main focus of this paper is to model the summer time ground level O_3 concentration in an Industrial area comprising of hazardous waste management facility. Comparison study was made between the base classifiers and the ensemble classifiers. Most of the conventional models can well predict the average concentrations. In this case the peak concentrations are of importance as it has serious effect on human health and environment. The models developed should also be homoscedastic.

A Novel Bagging Ensemble Approach for Predicting Summer Time Ground Level Ozone Concentration

Author 1

- Mohan S, PhD
- Professor, Environmental and Water Resources Engineering Division, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India- 600 036. Telephone: +91-044-2257-4261. e-mail: smohan@iitm.ac.in.

Author 2 †

- Saranya Packiam, M.Tech
- PhD Scholar, Environmental and Water Resources Engineering Division, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India- 600 036. e-mail: saranyap28@gmail.com.

- †Corresponding Author, PhD Scholar, Environmental and Water Resources Engineering Division, Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India- 600 036. e-mail: saranyap28@gmail.com.

Introduction

In the recent years, ground level ozone (O_3) has become a serious concern in many countries across the world. Ground level O_3 is a secondary pollutant that is formed when the primary pollutants Nitrogen Oxides (NO_X) and Volatile Organic Compounds (VOCs) react in the presence of sunlight. NOx and Non Methane VOCs are considered to be the major precursors of surface level ozone and other precursors include carbon monoxide, methane and Sulphur di oxide. Since O_3 has a long life time varying from hours to a day depending on humidity, temperature and air movement, it acts as a transboundary pollutant. O_3 is toxic beyond certain level as it is a strong oxidant. It can damage the lung tissue and weaken the immune system of human beings (Jakab et al., 1995). It can cause impairment of rubber goods and surface coating of materials. O_3 causes membrane damage on leaves especially in plants like bean, tobacco, birch etc (Chaudhary and Agrawal, 2015; Lee et al., 2017).

The formation of O_3 is considered to be a complex reaction since, the production of O_3 is altered by the influence of solar intensity, meteorological conditions, NO_X and VOC ratio and type of hydrocarbon (Finlayson-pitts and Pitts, 2000; Jenkin and Clemitshaw, 2002). Several researches have focused on identifying the factors causing O_3 formation and its transport (Jana et al., 2014; Souza and Kova, 2016; Tony and Sexauer, 2015). Besides, recent studies have reported statistically strong relationship between peak O_3 levels and meteorological parameters, NO_X and VOC (Knezovic et al., 2018; Thi et al., 2017).

Modeling of ground level O_3 has been one of the notable topics during the last decade in the air pollution community. Numerous approaches for predicting ground level O_3 have been reported in literature (Lu and Wang, 2014). These approaches can be categorized as follows: traditional statistical approach, deterministic approach (chemistry transport models) and machine learning. Traditional statistical approach includes: multiple linear regression (Abdul-Wahab et al., 2005; Özbay et al., 2011) multiple linear regression combined with principal component analysis (Abdul-Wahab et al., 2005; Pavón-Domínguez et al., 2014; Rajab et al., 2013; Tan et al., 2016) and DAUMOD-GRS models (Pineda Rojas et al., 2016). Deterministic models include WRF CHEM (Hoshyaripour et al., 2016) and WRF CMAQ (Astitha et al., 2017; Hogrefe et al., 2015; Sharma et al., 2016; Sharma and Khare, 2017), CHIMERE (Boynard et al., 2011). With the development of data mining tools, machine learning techniques have gained much interest, for example, multilayer perceptron (Fontes et al., 2014; Kumar et al., 2017; Lu and Wang, 2014; Mishra and Goyal, 2016), support vector machine (Gong and Ordieres-Meré, 2016; Lu and Wang, 2014), Ensemble approach (Bagging) (Al Abri et al., 2015).

The complexity of O_3 formation combined with uncertainty in the measurement of parameters involved makes the modeling process intricate. Non linear relationship between O_3 and its

contributing factors makes the linear models unfit (Cannon et al., 2011). Neural networks proved to be strong nonlinear estimators except the limitation of over fitting (Singh et al., 2013).

In the recent years, researchers have focused on advanced models like ensemble models which showed better performance than standard single machine learning classifiers (Gong and Ordieres-Meré., 2016; Hu et al. 2018). Framing and building of ensemble models is difficult and hence consumes more time for training the data set, combination of base classifiers and tuning of parameters associated with each classifier (Lu and Wang, 2014). Ensemble classifiers proved to perform well when compared with single base classifiers in the sectors like banking (Erdal and Karahanoğlu, 2016), medical applications (Zheng et al., 2018) and, industries (Hu et al., 2018). Three types of ensemble methods include: bagging, boosting and stacked generalization. Bagging (Boot Strap Aggregating) technique was developed by Breiman (Hacer et al., 2015). Bagging decreases the residual error between the observed and predicted values by creating bootstrapped replica data sets (Friedman, 2002). Cannon et al., (2011) adopted ensemble neural network approach for predicting the summer season O_3 levels. Ensemble neural network models showed 7% increase in the variance compared to multiple linear regression models. Singh et al., 2013 used ensemble trees to predict the air quality utilizing meteorological parameters as estimators. Performance was checked in terms of classification as well as regression. They found that both bagging and boosting trees performed better than single SVM classifier. These studies were constrained in a way that bagging was employed with REPTree. REPTree is the default classifier associated with bagging.

In this paper, the supremacy and viability of ensemble bagging classifiers over base classifiers such as multi layer perceptron, RTree, REPTree and Random forest for predicting summer time ground level O_3 prediction has been presented. A comparative study was carried out to assess the performance of machine learning techniques using the WEKA tool kit (WEKA 3.8.2). The main objectives of the present study are (i) to develop ensemble bagging model to predict the ground level O_3 concentration (dependent variable) utilizing air quality (NO, NO_2) and meteorological parameters (temperature, solar irradiance, relative humidity, wind speed and wind direction) as the independent variables and (ii) to evaluate the capability of each modeling method. A comparative analysis is provided to assess the performance of ensemble bagging classifiers and single classifier in predicting O_3 concentration.

Ensemble Method

Bagging

For a regression problem bagging works as follows (Hacer et al., 2015):

A training set D consists of data $\{(X_i, Y_i)_{i=1,2\ldots,n}\}$ where X_i is a realization of a multi-dimensional predictor variable and Y_i is a realization of a real-valued variable. A predictor $(Y|X = x) = f(x)$ is denoted by

$$C_n(x) = h_n(D_1, \ldots D_n)(x) \quad (1)$$

Bagging is explained as follows:

Step 1: Create a bootstrapped sample

$$D_i^* = (Y_i^*, X_i^*) \quad (2)$$

According to the empirical distribution of the pairs $D_i = (X_i, Y_i)$ where (i=1,2,…,n).

$$C_n^*(x) = h_n(D_1^*, \ldots, D_n^*)(x) \quad (3)$$

Step 2: Assess the bootstrapped predictor by the plug-in-principle, where $C_n(x)=h_n(D_1\ldots,D_n)(x)$

Finally the bagged predictor is

$$C_{n:B}(x) = E|D_n^*(x)| \quad (4)$$

The procedure for the model development of ensemble bagging trees is shown in Figure 1.

Figure 1 here

Data Collection and Data Preprocessing

Data Collection

An industrial site at Gummidipoondi town with geographical position of 13.4069°N and 80.1103°E and located 45 km north of Chennai city, Tamilnadu was chosen for ozone measurements. The industrial complex in which the study area is located comprises of steel, chemicals, Auto ancillary, auto components and plastic industries. In particular, the monitoring location is nearby a hazardous waste management facility which includes an incinerator and a landfill. In addition, Asian Highway 45 (AH 45) is at a distance of 500 m from the study location and is illustrated in Figure 2. Hazardous waste landfill facility has a storage facility wherein hazardous wastes such as solvents, flammables, explosives etc. were stored. These wastes acted as key source of VOCs in addition to the VOCs emanating from landfill. NO_X is produced mainly from the trucks moving inside the industrial area and other vehicles plying on AH 45 road. Hence it would result in large concentration of O_3 at the study area.

Figure 2 here

O_3 42 M analyzer was used to measure the ozone concentration at the site. O_3 analyzer was installed in an existing building at the site. The inlet tube of the O_3 analyzer was open to the atmosphere outside the building at a height of 3m. NO_X measurements were done using 32M NO_X analyzer, which was kept besides the O_3 analyzer. NO_X analyzer uses the principle of chemiluminiscence. The gas measurements were recorded every 15 minutes to depict the O_3

variation. Furthermore, meteorological parameters were also measured by using Spectrum Watchdog 2000 series weather station model 2900ET.

Continuous measurements of all parameters namely O_3, NO, NO_2, meteorological parameters were made during the period between Jan 2016 and Dec 2016 except during the maintenance of the analyzers. There were a total of 8278 instances recorded during the entire study period. The ensemble classifiers were trained for predicting the summer time data (March to May) of 2130 instances. Trials were conducted to evaluate the performance and suitability of different classifiers for the dataset. Summary statistics of the data set are shown in Table 1. Further, the box plot of NO, NO_2, O_3 is shown in Figure 3 (a) and hourly variation of meteorological parameters across the time period is shown in Figure 3 (b).

Figure 3 here

In the present study, prediction of surface ozone concentrations was carried out using data mining algorithms. The process was instigated using machine learning open-source software WEKA 3.8.2. Two types of data analysis exist: classification and prediction. Models are built based on the information obtained from important attribute classes. Meta - Bagging classifier is used for the present study. Both explorer and experimenter environments in WEKA were used for the data analysis.

Bagging algorithm, which is considered to a prevalent ensemble learning method was used to predict the O_3 concentration. Instance based machine learning classifiers such as decision stump, Rtree and Reduced error pruning tree, random forest, MLP and SMOreg were employed as the base learner.

Table 1 here

Data preprocessing

Data preprocessing is an essential stage in obtaining the final data set which can be used for further data mining (García et al., 2015). There were missing values and outliers in the data set, which were recorded during the malfunctioning of analyzers, during high temperatures and clogging of filter paper. In order to remove these outliers, data preprocessing techniques such as "interquartile range for attributes" and "remove with values for instances" were adopted. Wind direction is considered as one of the attribute for predicting the O_3 concentration and was measured with reference to 360° on the compass (true North) in a clockwise direction. The data possess 0° and 360° which will be considered as different values by the algorithm. Thus to avoid such mistake, wind speed and wind direction were combined together. Sine and cosine functions of the wind direction were calculated and replaced with wind direction. Attribute selection was done using filters such as CfssubsetEval (Best first and Greedystepwise). Principal components

and ReleifFattributeEval. It was observed that there was no improvement in the prediction accuracy. Rather the values of error measures deteriorated through the application of filters as it resulted in the loss of input information.

Performance indices

The indices used for analyzing the performance of analyzers are as follows: Correlation Coefficient (CC), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE), R^2 and Index of Agreement (IA).

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(O_j - P_{ij})^2}{n}} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|P_{ij} - O_j\right| = \frac{1}{n}\sum_{j=1}^{n}|e_i| \quad (6)$$

$$CC = \frac{n\sum O_j P_{ij} - (\sum O_j)(\sum P_{ij})}{\sqrt{n(\sum O_j^2) - (\sum O_j)^2}\sqrt{n\left(\sum P_{ij}^2\right) - (\sum P_{ij})^2}} \quad (7)$$

$$RAE = \frac{\sum_{j=1}^{n}|P_{(ij)} - O_{(j)}|}{\sum_{j=1}^{n}|O_j - \mu_o|} \quad (8)$$

$$RRSE = \frac{\sum_{j=1}^{n}(P_{(ij)} - O_{(j)})^2}{\sum_{j=1}^{n}(O_j - \mu_o)^2} \quad (9)$$

$$IA = 1 - \left[\frac{\sum_{j=1}^{n}(P_{ij} - O_j)^2}{\sum_{j=1}^{n}\left(|P_{ij} - \mu_o| + |O_j - \mu_o|\right)^2}\right] \quad (10)$$

Where, $P_{(ij)}$ is the predicted value, O_j is the observed value, μ_o is the mean of the observed values, n= number of pairs of data, e_i is the absolute error.

Model Building

Initially, data set was tested using conventional multiple regression techniques combined with PCA. These methods are highly data dependent. For the present data set, the prediction was relatively poor using the multiple linear regression. Since the error measures obtained from the above method are RAE =69.21% and RRSE= 73.44%, the prediction is relatively poor as is evident from the Figure 4. In the next phase, machine learning techniques were used to improve the prediction of peak O_3 concentration. Supervised classification was used and classifiers employed for the study include: Function (linear regression, Multilayer Perceptron, RBF-Network, SMOreg); Lazy (IBK, Kstar, LWL); Meta (additive regression, bagging, Random-Subspace); Rules (M5Rules); Trees (REPTrees, Random forest, Decision-Stump and M5P). The

parameters for each classifier were optimized and over fitting was avoided by using *k* fold cross validation. Ten fold cross validation was used to create the models. This involves splitting the dataset into ten equal subsets and using nine subsets as training data and one as test data. The procedure was iterated 10 times so that every subset was utilized as test data once. The final model was then the average of the 10 iterations.

Figure 4 here

Most of the above mentioned classifiers were able to predict the lower concentration values well but the prediction for high concentrations were relatively poor. The main objective of the present study is to predict the peak values of O_3, as it can pose health risk to the workers in the surrounding areas and not the time series analysis of O_3. High levels of O_3 were recorded during the summer season (March 2016- May 2016). Back trajectory analysis using HYSPLIT was carried out (shown in Figure 5) and it proved that during the summer months, the air mass had marine origin and was relatively clean. Hence the high concentration of O_3 observed was assumed to be mainly due to the photochemical reaction of the precursors present in-situ and local transportation. The time dependent analysis was not helpful in improving the prediction accuracy.

In order to quantify the performance of the ensemble models in predicting peak O_3 concentrations, the error measures for the values greater than 180 $\mu g/m^3$ of observed O_3 concentration were also considered. The peak O_3 levels were further evaluated manually by calculating the error in the predicted peak values. The performance of the classifier was estimated based on the number of peaks predicted within certain error ranges. The percentage of errors considered in this study was 5%, 10%, 15 % and 20%.

Figure 5 here

Results and Discussion

The formation of ensemble consists of two steps, namely, creation of individual ensemble members and appropriately consolidates the output from the individual ensemble learners. In the present study, ensemble learning approach, in particular, boot strap aggregating (bagging) was utilized. The base learners selected for the present study are MLP, RTree, REPTree and random forest. To reduce the impact of the variability of the training set, the experiments were repeated hundred times. Each time, all algorithms were trained on the same portion of the training data and evaluated on the same test data. Initially, runs were performed with default parameters and optimized results were not obtained. Hence, selection of optimal parameters becomes important for better prediction of O_3 using both bagging and base learners. The parameters associated with bagging technique are number of iterations and bag size percent. In the present study, the number

of iterations was chosen as 100, 200, 300, 400 and 500 and the size of bag as 60%, 70%, 80%, 90% and 100%. By default, the number of iterations is 10 and bag size is 100%. For all the models, the number of seed was taken as 1.

R^2 value was used to check the goodness of the fit and the relationship present in the data. Also, this error measure outlines the degree of descriptiveness of the regression model. For each model, improvement in R^2 values is shown in Figures 6 (a) – (d). Optimum parameters for bagging were selected as the one which had R^2 value close to unity and other error measures close to zero.

Figure 6 here

By simultaneously tuning the parameters in bagging as well as base learner, the improvement in R^2 values were 30.6%, 7.38% and 13.08% for RTree, REPTree and random forest, respectively. The performance of the optimized models were analyzed using CC, MAE, RRSE, RAE and IA evaluation criteria as shown in Figures 7 (a) – (e). The distributions of the observed and the predicted concentrations are shown as box plot in Figure 8. In Figure 8, the square box indicates the median, the horizontal line at the top and bottom indicate the maximum and minimum value and 'x' depict the 1^{st} and 99^{th} percentiles. The skewness of the actual and the predicted data were analyzed for all the methods and were found to be positively skewed.

Figure 7 here

Figure 8 here

To check the statistical significance among the prediction models, Wilcoxon signed-rank test (two- tailed test at 95% confidence level) was carried out. It is to be noted that the Wilcoxon signed rank test was not impacted by the outlier data points. The results of the Wilcoxon signed-rank are shown in Table 2. The test results shown in Table 2 indicated that RTree, REPTree and bagged RTree models follow similar trend. Bagged random forest model predicted both peak and low values accurately.

Table 2 here

The scatter plots of the observed and the predicted O_3 concentrations are shown in Figures 9 (a) – (h). From Figures 9 (a) – (h), it can be observed that the bagged random forest model performed better compared to other models. Figures 9 (a) - (h) confirms that the predicted peak value points remained fairly constant between the base classifiers except for random forest. There was a big shift in values close to the 45° line in the bagged REPTree and bagged random forest classifiers, but the majority of the peak values remain unchanged in ensemble MLP and RTrees classifiers. Also, the variance around the regression line is same for all the data points in case of bagged random forest (Figure 9 (h)), hence the model is homoscedastic for entire range of data. When compared with the other models, the ensemble random forest performed well with R^2 value of

0.9432. From Figure 9 (g), it can be observed that Bagged REP tree predicted the peak values well but the model is heteroscedastic, i.e., it performs well only in a particular range of data. Hence, bagged REP tree model was checked with specified range of data. It was found once again to be heteroscedastic and hence, not suitable for prediction of O_3.

Figure 9 here

The ensemble random forest model produced marginally better performance across the entire data set and much better prediction of peak O_3 concentration (Table 3). Considering the entire data set, bagged random forest model experienced up to 70.66% reduction in RMSE value among the ensemble models. RTree performed better than the bagged RTree with lower error rate and higher coefficient of determination.

Table 3 here

Peak flow performance was further confirmed by manually determining the error rates of each peak within the dataset (shown in Table 3). Peak values (greater than 180 $\mu g/m^3$) were determined manually producing results which confirmed the performance of each classifier. The increased performance of ensemble models was further confirmed by manual peak calculations (Table 4). Sixteen peaks with less than 10 % error in the bagged random forest model indicate very high performance, and this result is the major difference between the bagged random forest and the bagged REPTree results.

Table 4 here

To check the performance of the bagged random forest classifier, independent data set which was included in training phase was used. The data set included 10 days hourly data (240 instances) measured during the summer season. Initially, the test data set was subjected to preprocessing techniques. The prediction of ground level O_3 by ensemble models is shown in Figure 10. It is evident from Table 4 that the bagged random forest predicted better with increased correlation coefficient and decreased values of other error measures such as CC, MAE, RRSE, RAE, R^2 and IA. Similar results were reported by Erdal and Karahanoğlu, 2016 and Nawahda, 2016 . From Figure 10, it is evident that there were less predictions of under estimations and over estimations. Also, the peak values which are the key focus of this study are predicted well within 5% error range. Hence bagged random forest model proved to be effective and reliable method for predicting the ground level O_3 concentration. It is to be noted, that the execution of these models is subject to the data set that it is applied to.

The O_3 measurements tend to have lot of noise and the noises disrupted the training data for building the model. Bagging helped to reduce such noises present in the data. Bagging trains a large number of strong learners in parallel and finally merges the output of all the strong learners

together so as to smooth out their predictions. Due to the above procedure, it endeavors to avoid the problem of over fitting linked with classifiers such as MLP.

Figure 10 here

Conclusion

The objective of the present study was to accurately predict the summer time ground level O_3 concentration in an industrial area in Chennai, Tamilnadu. The data set included hourly averages of O_3, and NO, NO_2 and meteorological parameters such as relative humidity, temperature, solar irradiance, wind speed and wind direction. Single base classifiers and ensemble classifiers were employed for the prediction of O_3 concentration observed during the summer season. In addition, the performance was checked against multiple regression model combined with PCA. It was found that most of the models developed were heteroscedastic. The ensemble classifiers yielded better results than the base classifiers and also had better accuracy in predicting both low as well as high concentration of O_3. Bagged random forest performed better than the other methods such MLP, RTree and REPTree. The developed bagged random forest model was homoscedastic and showed lower values of RAE, MAE, RRSE and higher values of CC and IA compared to the traditional methods such as MLP. These models can be used as tools while framing ozone control strategies and setting O_3 standards. Ensemble approach reduces the bias by effectively using the training data set. Also, it lowers the variance by combining the outputs multiple times from the same learning method.

Acknowledgements

The authors thank the China Section of the Air & Waste Management Association for the generous scholarship they received to cover the cost of page charges, and make the publication of this paper possible. The authors gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for granting the permission to use their HYSPLIT transport and dispersion model and/or READY website (http://www.arl.noaa.gov/ready.html) used in this publication.

References

Abdul-Wahab, S.A., Bakheit, C.S., Al-Alawi, S.M., 2005. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ. Model. Softw. 20, 1263–1271. https://doi.org/10.1016/j.envsoft.2004.09.001

Al Abri, E.S., Edirisinghe, E.A., Nawadha, A., Kingdom, U., 2015. Modelling Ground-Level Ozone Concentration using Ensemble Learning Algorithms. Int. Conf. Data Min. (DMIN). Steer. Comm. World Congr. Comput. Sci. Comput. Eng. Appl. Comput. 148–154.

Astitha, M., Luo, H., Rao, S.T., Hogrefe, C., Mathur, R., Kumar, N., 2017. Dynamic evaluation of two decades of WRF-CMAQ ozone simulations over the contiguous United States. Atmos. Environ. 164, 102–116. https://doi.org/10.1016/j.atmosenv.2017.05.020

Boynard, A., Beekmann, M., Foret, G., Ung, A., Szopa, S., Schmechtig, C., Coman, A., 2011. An ensemble assessment of regional ozone model uncertainty with an explicit error representation. Atmos. Environ. 45, 784–793. https://doi.org/10.1016/j.atmosenv.2010.08.006

Cannon, A.J., Lord, E.R., Cannon, A.J., Lord, E.R., 2011. Forecasting Summertime Surface-Level Ozone Concentrations in the Lower Fraser Valley of British Columbia : An Ensemble Neural Network Approach PAPER Forecasting Summertime Surface-Level Ozone Concentrations in the Lower Fraser Valley of British Columbia : 2247. https://doi.org/10.1080/10473289.2000.10464024

Chaudhary, N., Agrawal, S.B., 2015. The role of elevated ozone on growth, yield and seed quality amongst six cultivars of mung bean. Ecotoxicol. Environ. Saf. 111, 286–294. https://doi.org/10.1016/j.ecoenv.2014.09.018

Erdal, H., Karahanoğlu, İ., 2016. Bagging ensemble models for bank profitability: An emprical research on Turkish development and investment banks. Appl. Soft Comput. J. 49, 861–867. https://doi.org/10.1016/j.asoc.2016.09.010

Fontes, T., Silva, L.M., Silva, M.P., Barros, N., Carvalho, A.C., 2014. Can artificial neural networks be used to predict the origin of ozone episodes? Sci. Total Environ. 488–489, 197–207. https://doi.org/10.1016/j.scitotenv.2014.04.077

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

García, S., Luengo, J., Herrera, F., 2015. Data Preprocessing in Data Mining. Intell. Syst. Ref. Libr. 72. https://doi.org/10.1007/978-3-319-10247-4

Gong, B., Ordieres-Meré, J., 2016. Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: Case study of Hong Kong. Environ. Model. Softw. 84, 290–303. https://doi.org/10.1016/j.envsoft.2016.06.020

Hacer, Y.A. mu, Aykut, E., Halil, E., Hamit, E., 2015. Optimizing the monthly crude oil price forecasting accuracy via bagging ensemble models. J. Econ. Int. Financ. 7, 127–136. https://doi.org/10.5897/JEIF2014.0629

Hogrefe, C., Pouliot, G., Wong, D., Torian, A., Roselle, S., Pleim, J., Mathur, R., 2015. Annual application and evaluation of the online coupled WRF-CMAQ system over North America under AQMEII phase 2. Atmos. Environ. 115, 683–694. https://doi.org/10.1016/j.atmosenv.2014.12.034

Hoshyaripour, G., Brasseur, G., Andrade, M.F., Gavidia-Calderón, M., Bouarar, I., Ynoue, R.Y., 2016. Prediction of ground-level ozone concentration in São Paulo, Brazil: Deterministic versus statistic models. Atmos. Environ. 145, 365–375. https://doi.org/10.1016/j.atmosenv.2016.09.061

Hu, G., Li, H., Xia, Y., Luo, L., 2018. Computers in Industry A deep Boltzmann machine and multi-grained scanning forest ensemble collaborative method and its application to industrial fault diagnosis. Comput. Ind. 100, 287–296. https://doi.org/10.1016/j.compind.2018.04.002

Jakab, G.J., Spannhake, E.W., Canning, B.J., Kleeberger, S.R., Gilmour, M.I., 1995. The effects of ozone on immune function. Environ. Health Perspect. 103 Suppl, 77–89.

Jana, P.K., Bhattacharyya, S., Banerjee, A., 2014. Effect of some climatic parameters on tropospheric and total ozone column over Alipore (22 . 52 ◦ N , 88 . 33 ◦ E), India 1653–1669.

Knezovic, M., Pintaric, S., Mornar, M., Kes, V.B., Nesek, V., 2018. The role of weather conditions and normal level of air pollution in appearance of stroke in the region of Southeast Europe. Acta Neurol. Belg. 118, 267–275. https://doi.org/10.1007/s13760-018-0885-0

Kumar, N., Middey, A., Rao, P.S., 2017. Prediction and examination of seasonal variation of ozone with meteorological parameter through artificial neural network at NEERI, Nagpur,

India. Urban Clim. 20, 148–167. https://doi.org/10.1016/j.uclim.2017.04.003

Lee, J.Y., Lee, S.H., Hong, S.C., Kim, H., 2017. Projecting future summer mortality due to ambient ozone concentration and temperature changes. Atmos. Environ. 156, 88–94. https://doi.org/10.1016/j.atmosenv.2017.02.034

Lu, W.Z., Wang, D., 2014. Learning machines: Rationale and application in ground-level ozone prediction. Appl. Soft Comput. J. 24, 135–141. https://doi.org/10.1016/j.asoc.2014.07.008

Mishra, D., Goyal, P., 2016. Neuro-Fuzzy approach to forecasting Ozone Episodes over the urban area of Delhi, India. Environ. Technol. Innov. 5, 83–94. https://doi.org/10.1016/j.eti.2016.01.001

Nawahda, A., 2016. An assessment of adding value of traffic information and other attributes as part of its classifiers in a data mining tool set for predicting surface ozone levels. Process Saf. Environ. Prot. 99, 149–158. https://doi.org/10.1016/j.psep.2015.11.004

Özbay, B., Keskin, G.A., Doğruparmak, Ş.Ç., Ayberk, S., 2011. Multivariate methods for ground-level ozone modeling. Atmos. Res. 102, 57–65. https://doi.org/10.1016/j.atmosres.2011.06.005

Pavón-Domínguez, P., Jiménez-Hornero, F.J., Gutiérrez de Ravé, E., 2014. Proposal for estimating ground-level ozone concentrations at urban areas based on multivariate statistical methods. Atmos. Environ. 90, 59–70. https://doi.org/10.1016/j.atmosenv.2014.03.032

Pineda Rojas, A.L., Venegas, L.E., Mazzeo, N.A., 2016. Uncertainty of modelled urban peak O3concentrations and its sensitivity to input data perturbations based on the Monte Carlo analysis. Atmos. Environ. 141, 422–429. https://doi.org/10.1016/j.atmosenv.2016.07.020

Rajab, J.M., MatJafri, M.Z., Lim, H.S., 2013. Combining multiple regression and principal component analysis for accurate predictions for column ozone in Peninsular Malaysia. Atmos. Environ. 71, 36–43. https://doi.org/10.1016/j.atmosenv.2013.01.019

Sharma, A., Sharma, S.K., Rohtash, Mandal, T.K., 2016. Influence of ozone precursors and particulate matter on the variation of surface ozone at an urban site of Delhi, India. Sustain. Environ. Res. 26, 76–83. https://doi.org/10.1016/j.serj.2015.10.001

Sharma, S., Khare, M., 2017. Simulating ozone concentrations using precursor emission inventories in Delhi – National Capital Region of India. Atmos. Environ. 151, 117–132. https://doi.org/10.1016/j.atmosenv.2016.12.009

Singh, K.P., Gupta, S., Rai, P., 2013. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos. Environ. 80, 426–437. https://doi.org/10.1016/j.atmosenv.2013.08.023

Souza, A. De, Kova, E., 2016. Assessment of Ozone Variations and Meteorological Influences in West Center of Brazil , from 2004 to 2010. https://doi.org/10.1007/s11270-016-3002-0

Tan, K.C., San Lim, H., Jafri, M.Z.M., 2016. Prediction of column ozone concentrations using multiple regression analysis and principal component analysis techniques: A case study in peninsular Malaysia. Atmos. Pollut. Res. 7, 533–546. https://doi.org/10.1016/j.apr.2016.01.002

Thi, H., Kwon, E.E., Kim, K., Kumar, S., Chambers, S., Kumar, P., Kang, C., Cho, S., Oh, J., Brown, R.J.C., 2017. Factors regulating the distribution of O 3 and NO x at two mountainous sites in Seoul , Korea. Atmos. Pollut. Res. 8, 328–337. https://doi.org/10.1016/j.apr.2016.10.003

Tony, R., Sexauer, M., 2015. Science of the Total Environment Identi fi cation of sources contributing to PM 2 . 5 and ozone at elevated sites in the western U . S . by receptor analysis : Lassen Volcanic National Park , California , and Great Basin National Park , Nevada. Sci. Total Environ. 530–531, 505–518. https://doi.org/10.1016/j.scitotenv.2015.03.091

Zheng, C., Mathew, K., Chen, C., Chen, Y., Tang, H., Dozier, A., Kas, J.J., Vila, F.D., Rehr, J.J., Piper, L.F.J., Persson, K.A., Ong, S.P., 2018. Automated generation and ensemble-learned matching of X-ray absorption spectra. npj Comput. Mater. 2018, 1–9. https://doi.org/10.1038/s41524-018-0067-x

Table 1. Statistic summary of the attributes for the prediction of ground level O_3 concentration during the study period

Attribute	Statistics Summary (Training Set) (1890 records)				Statistics Summary (Test Set) (240 records)			
	Min	Max	Mean	Std. dev	Min	Max	Mean	Std. dev
O_3 ($\mu g/m^3$)	0.016	497.232	25.3383	35.8386	0.024	398.928	23.713	27.963
Relative humidity (%)	24.53	99.07	82.128	17.78	21.5	99.1	77.7	15.8
Temperature (^oC)	22.99	40.59	33.019	2.60	21.5	40.3	30.2	2.8
Wind speed (mph)	0.21	4.04	1.36	0.621	0.15	4.5	1.9	0.7
Wind direction (degrees)	0	358	142.16	101.32	0	355	169	83
Solar radiation (W/m^2)	0	933.5	196.69	265.5	0	924.9	173.83	172.6
NO ($\mu g/m^3$)	11.56	335.48	69.75	37.36	8.47	261.8	18.8	35.9
NO_2 ($\mu g/m^3$)	1.97	452.52	72.48	42.36	1.42	378.2	20.53	44.74

Table 2. Wilcoxon signed rank test for base classifiers and bagged classifiers

	MLP	Rtree	REPTree	Random forest	B-MLP	B-Rtree	B-REPTree	B-Random forest
MLP								
Rtree	-13.036*							
REPTree	-11.5*	-7.967						
Random forest	-10.396*	-9.394*	-2.946*					
B-MLP	-12.141*	-2.553*	-8.498	-11.095				
B-Rtree	-10.731*	-7.071*	-0.073	-2.552*	-8.173			
B-REPTree	-10.596*	-7.397*	-1.389	-1.489	-11.67	0.99*		
B-Random forest	-9.579*	-8.994*	-4.038*	-4.519*	-12.662	-4.04*	-3.319*	

*significant at 0.05 level

Table 3. Performance of models for entire data set and peak values (>180 µg/m^3) in terms of RRSE and R^2

	Entire Dataset			Peak Values (>180 µg/m^3)	
	RRSE	R^2	prediction with error <10%	RRSE	R^2
MLP	23.4372	0.6159	155	8.804	0.4385
RTree	13.8421	0.6351	366	8.633	0.5608
REPTree	10.4818	0.7794	432	8.521	0.5285
Random forest	10.437	0.7896	427	7.47	0.44
B-MLP	10.3526	0.7872	355	7.815	0.5331
B-RTree	10.8496	0.7761	398	7.646	0.5696
B-REPTree	8.3353	0.8371	421	6.354	0.5342
B-Random forest	6.3571	0.9432	588	1.8196	0.5991

Table 4. Error measures for peak values of O_3 (>180 μg/m^3)

	Single				Ensemble			
	<1%	<5%	<10%	<20%	<1%	<5%	<10%	<20%
MLP	0	3	10	15	0	7	12	22
RTree	4	11	16	25	2	8	18	24
REPTree	1	6	12	18	3	4	10	18
Random forest	1	11	16	22	2	15	19	29

Table 5. Performance of the Bagging classifier for predicting ground level O_3 concentration using a test dataset

Classifier	Training data						Test Data					
	CC	MAE	RRSE	RAE	R^2	IA	CC	MAE	RRSE	RAE	R^2	IA
B-MLP	0.8873	10.3418	10.3526	0.36	0.7872	0.94	0.7948	13.4157	10.4372	0.857	0.6419	0.75
B-RTree	0.881	10.8396	10.8496	0.3773	0.7761	0.93	0.8969	13.83	13.8421	0.4813	0.5987	0.84
B-REPTree	0.9142	23.4158	8.3353	0.3645	0.8371	0.95	0.8729	10.472	10.4818	0.3645	0.7959	0.85
B-Random forest	0.9711	6.5774	6.3576	0.2286	0.9432	0.98	0.9422	10.427	10.437	0.3629	0.8321	0.89

Caption for Figures

Figure 1. Model development procedure of ensemble bagging tree

Figure 2. Map of the study location

Figure 3. (a) Box plot variation of NO,NO_2 and O_3 across the study period (b) Hourly variation of relative humidity, temperature and wind speed

Figure 4. Scatter plots of observed vs. predicted ozone in case of multiple linear regression combined with principal component analysis

Figure 5. seven days air mass back trajectory using HYSPLIT reaching the monitoring location

Figure 6. Improvement of R^2 statistic for various iteration and cluster size in case of (a) MLP (b) RTree (c) REPTree (d) random forest

Figure 7. Performance comparison of single models and ensemble models in terms of (a) CC (b) RRSE (c) MAE (d) RAE (e) IA

Figure 8. Box plot representation of observed vs. predicted concentration for different models

Figure 9. Scatter plots of observed vs. predicted ozone for (a) MLP (b) bagged MLP (c) RTree (d) bagged RTree (e) REPTree (f) bagged REPTree (g) random forest (h) bagged random forest

Figure 10. Prediction of hourly O_3 concentration using single classifier and ensemble classifier for independent data set. (Plot is shown for 1 day for clear understanding)

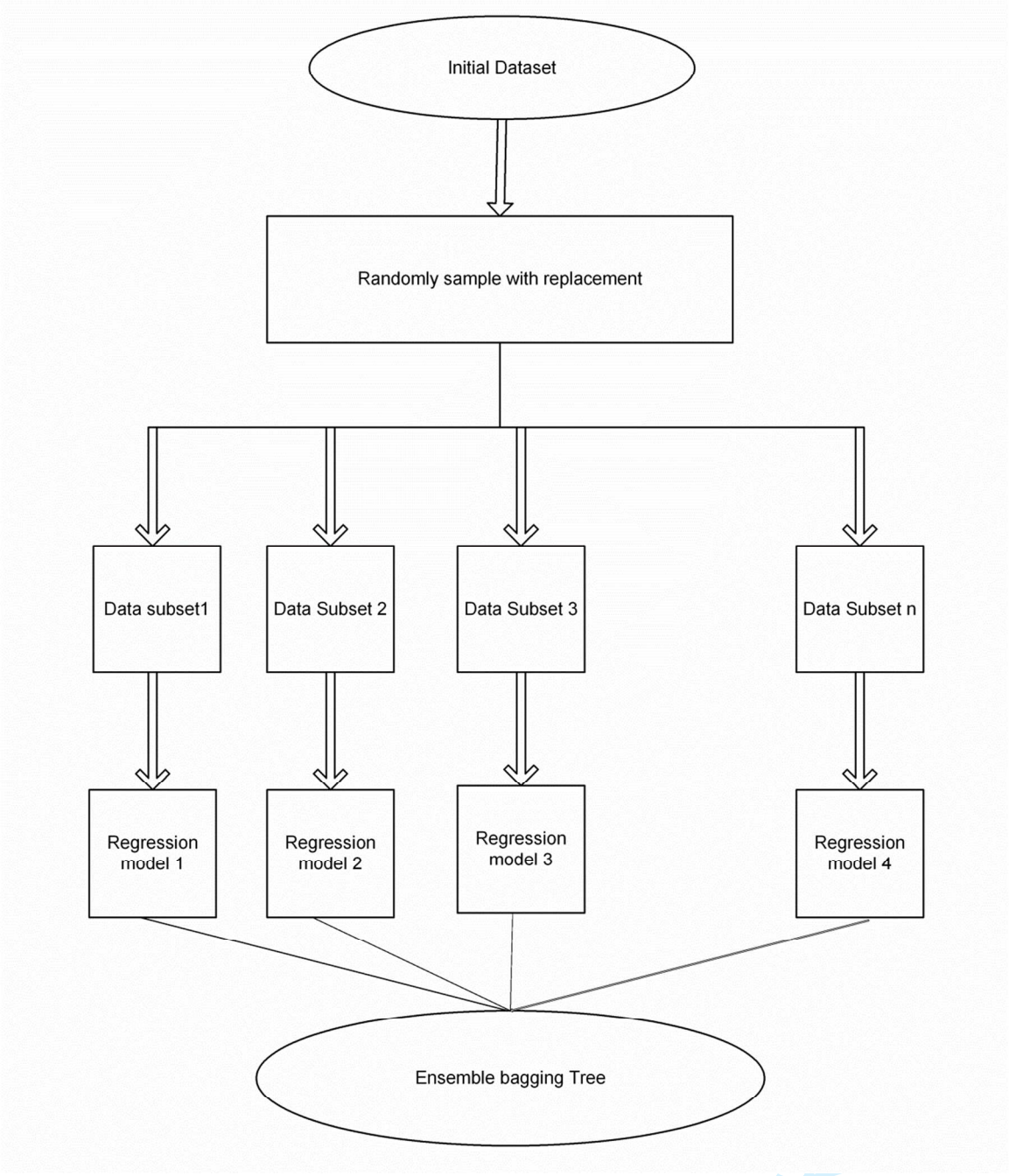
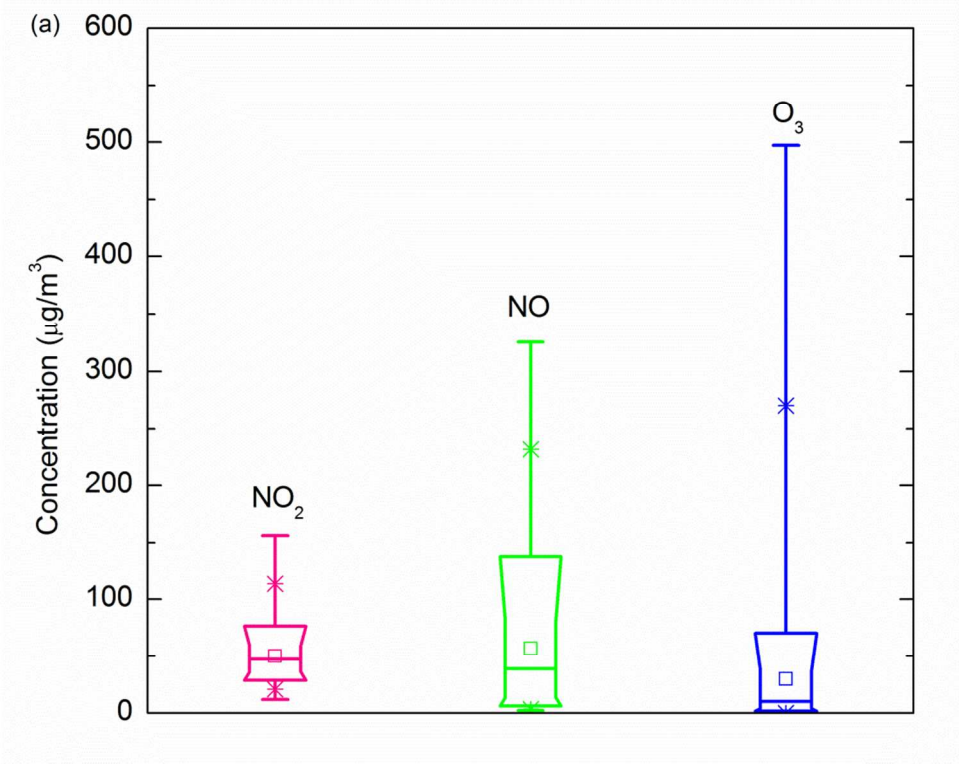


Figure 1. Model development procedure of ensemble bagging tree

Figure 2. Map of the study location



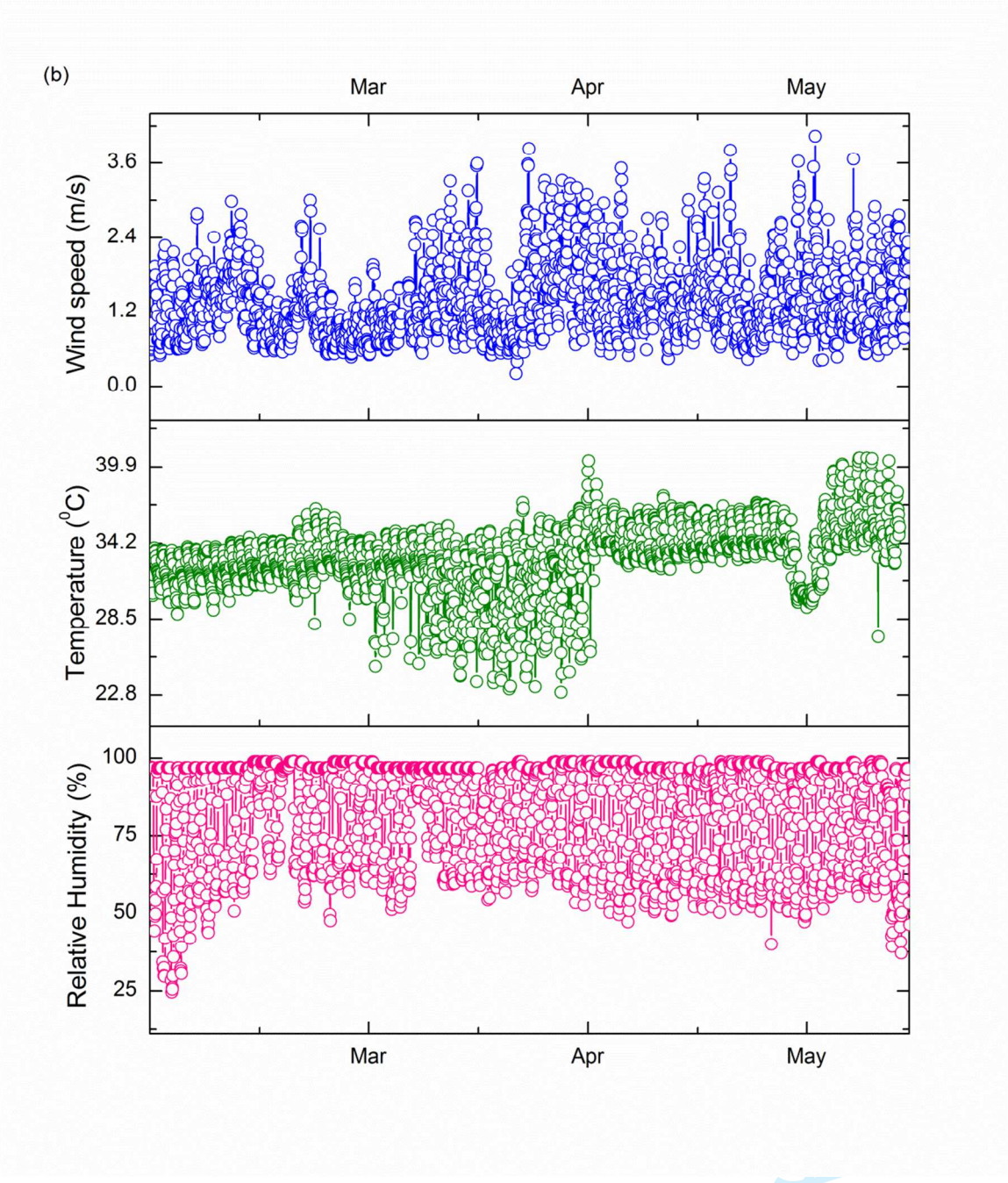


Figure 3. (a) Box plot variation of NO,NO_2 and O_3 across the study period (b) Hourly variation of relative humidity, temperature and wind speed

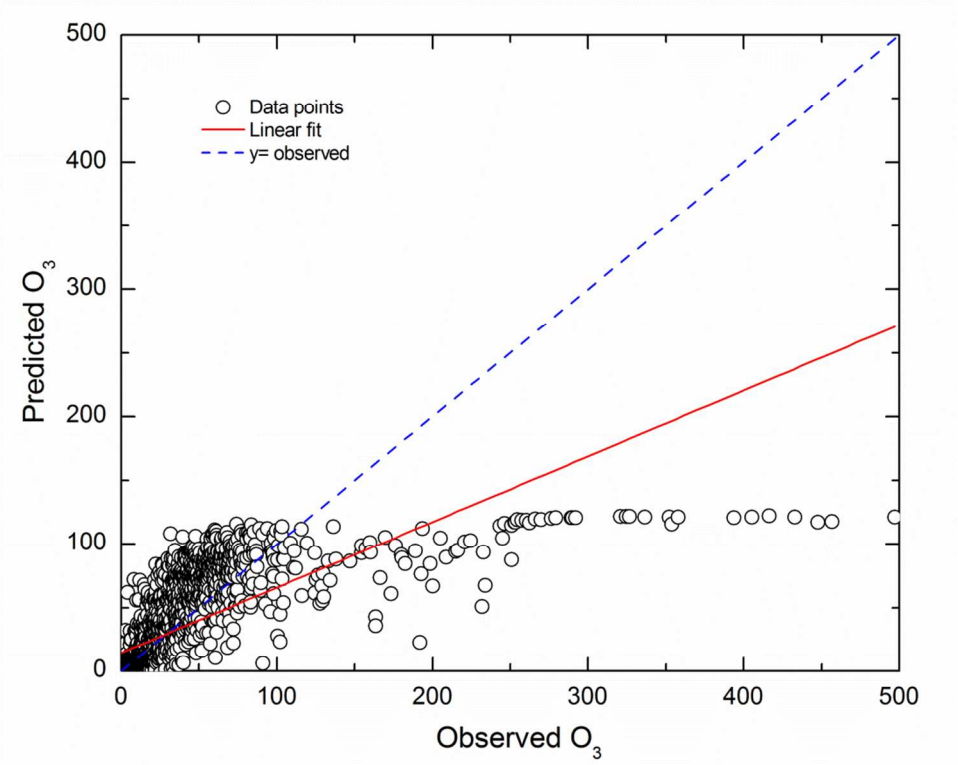


Figure 4. Scatter plots of observed vs. predicted ozone in case of multiple linear regression combined with principal component analysis

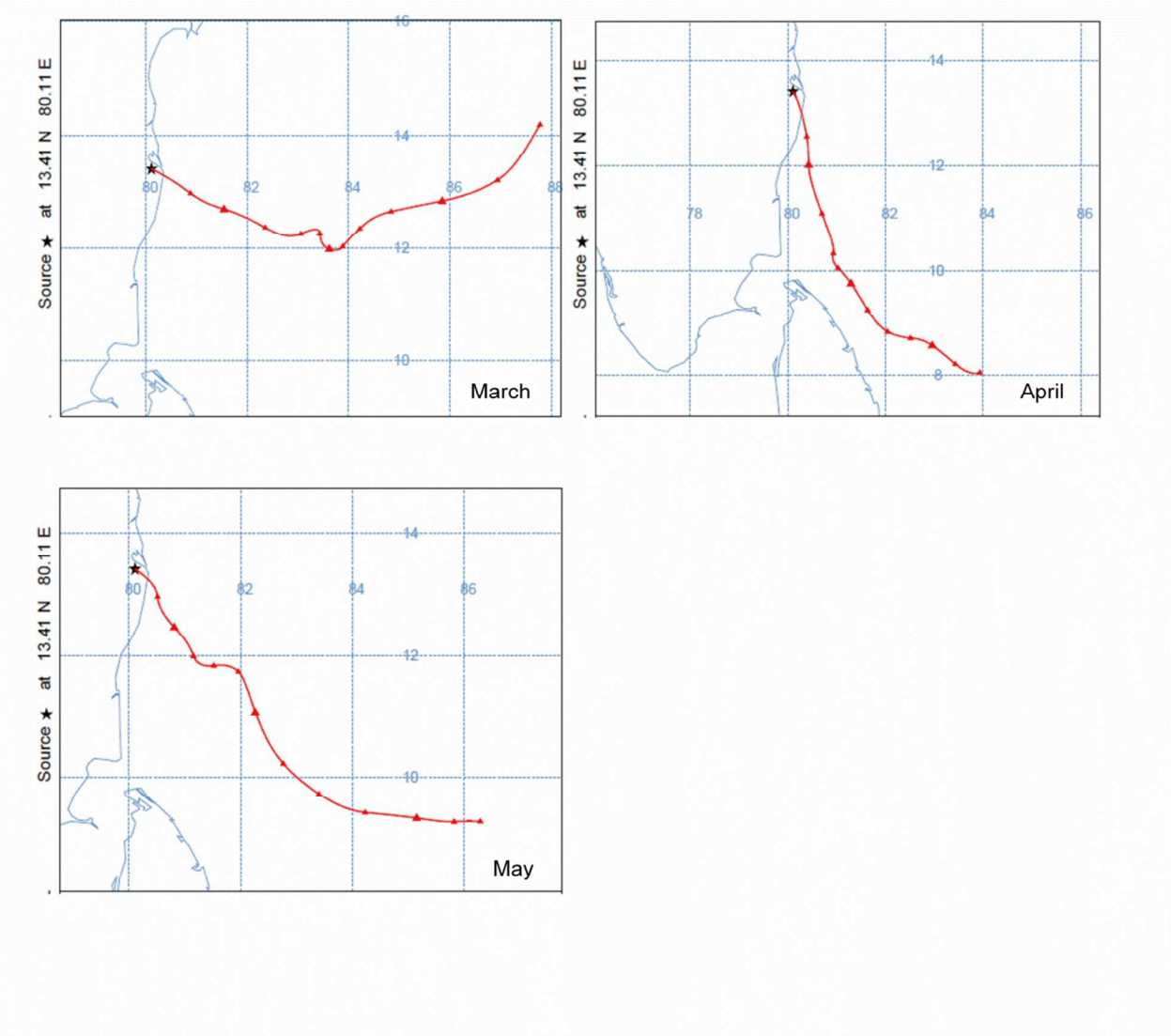
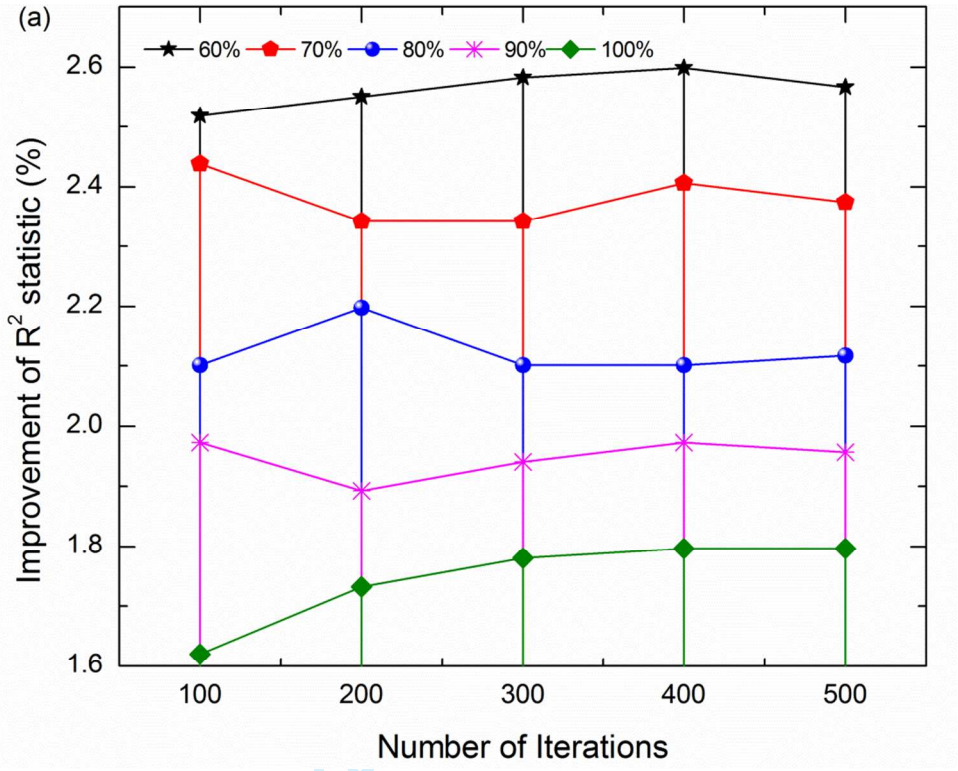
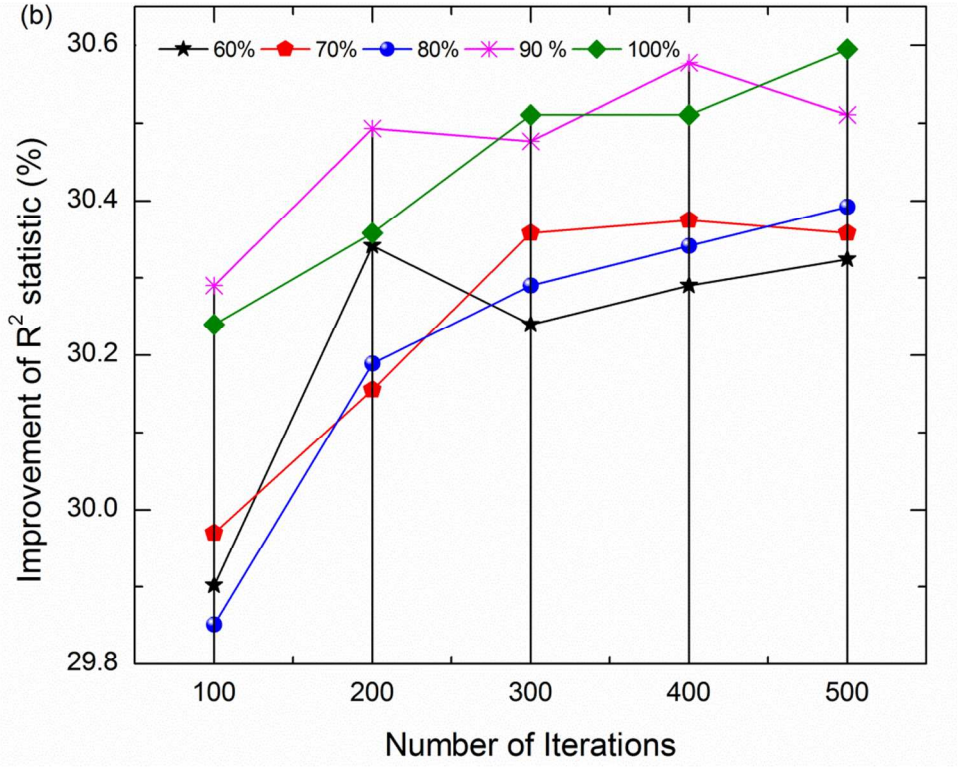


Figure 5. seven days air mass back trajectory using HYSPLIT reaching the monitoring location


(a)
60%
70%
80%
90%
100%
Improvement of R2 statistic (%)
2.6
2.4
2.2
2.0
1.8
1.6
100
200
300
400
500
Number of Iterations


(b)
60%
70%
80%
90 %
100%
Improvement of R2 statistic (%)
30.6
30.4
30.2
30.0
29.8
100
200
300
400
500
Number of Iterations

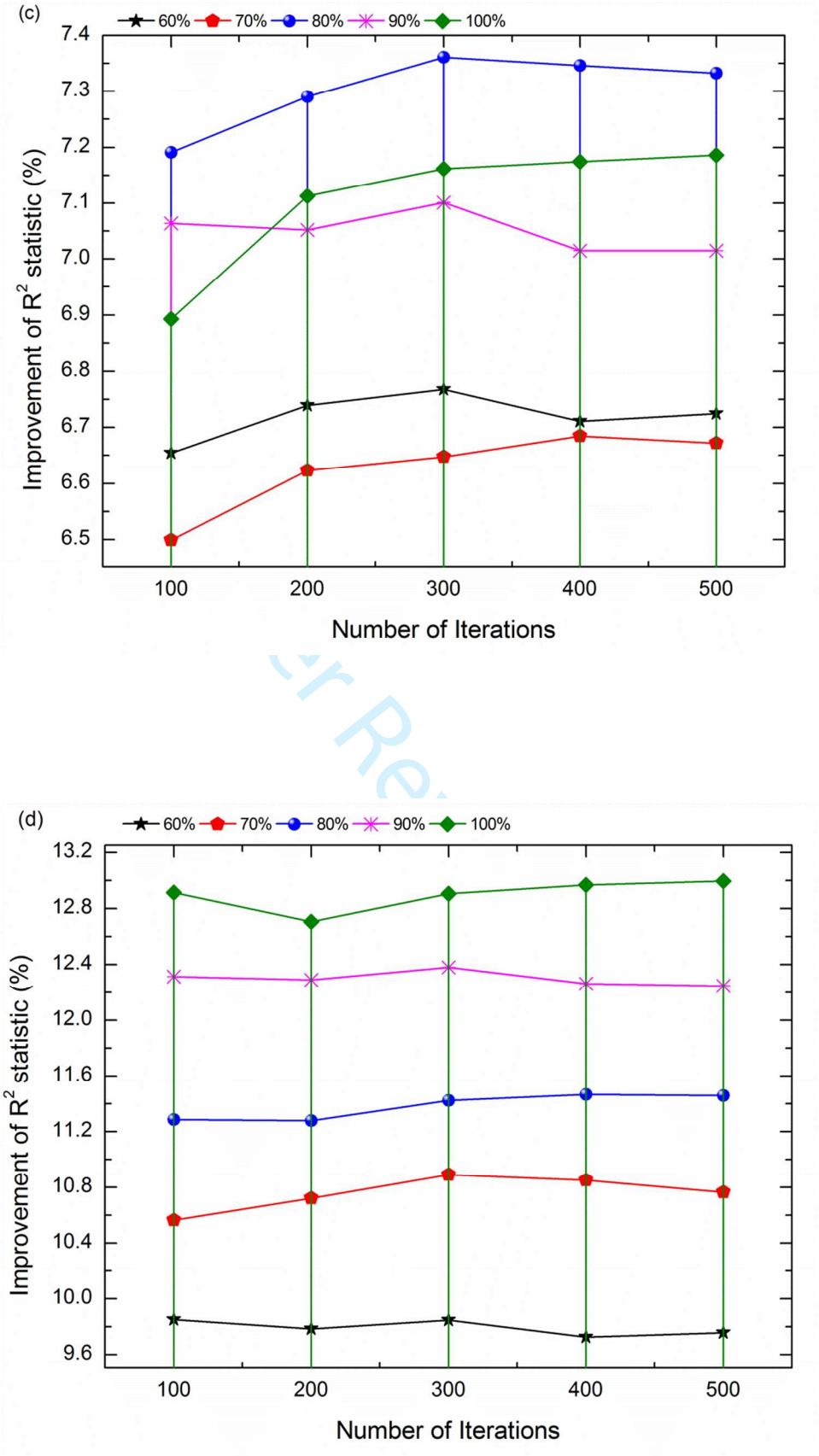
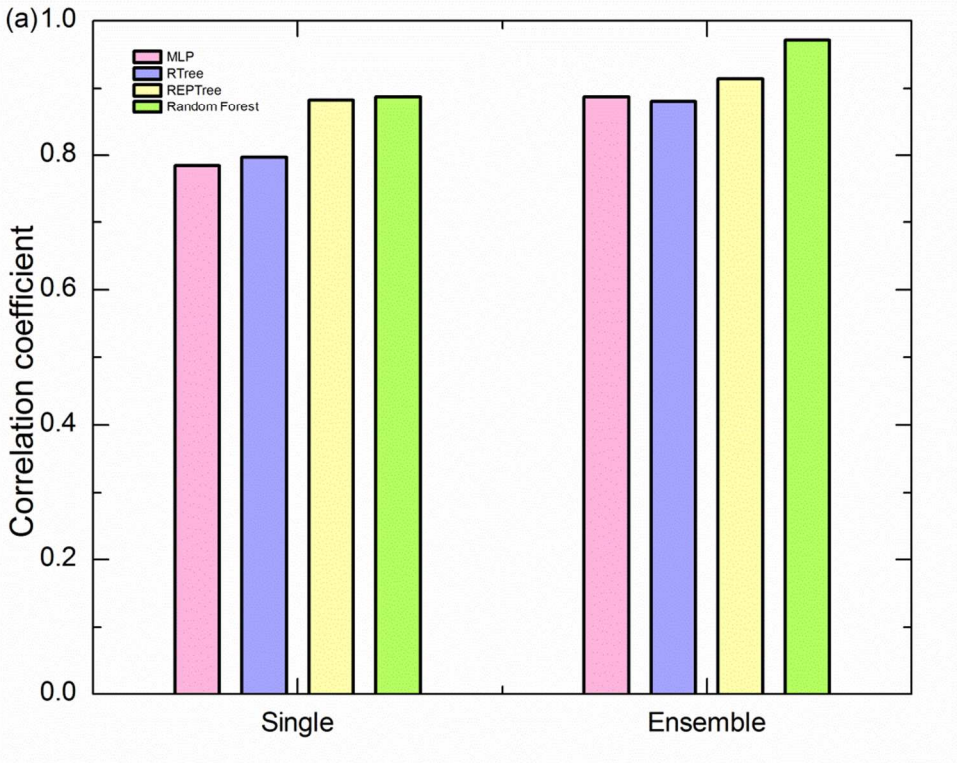
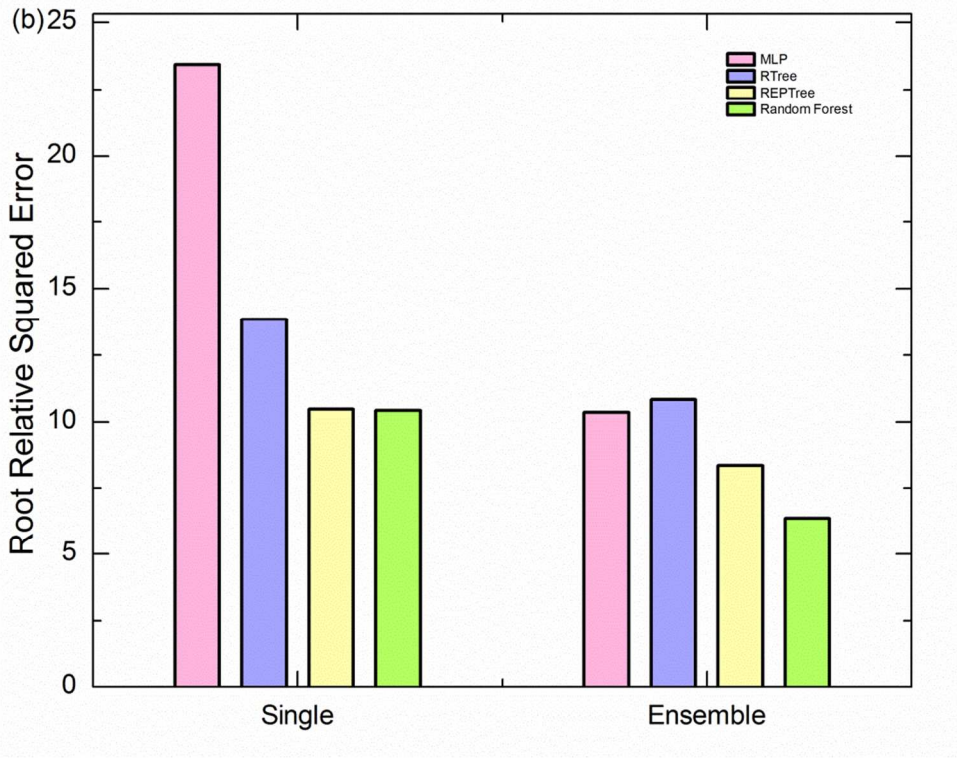
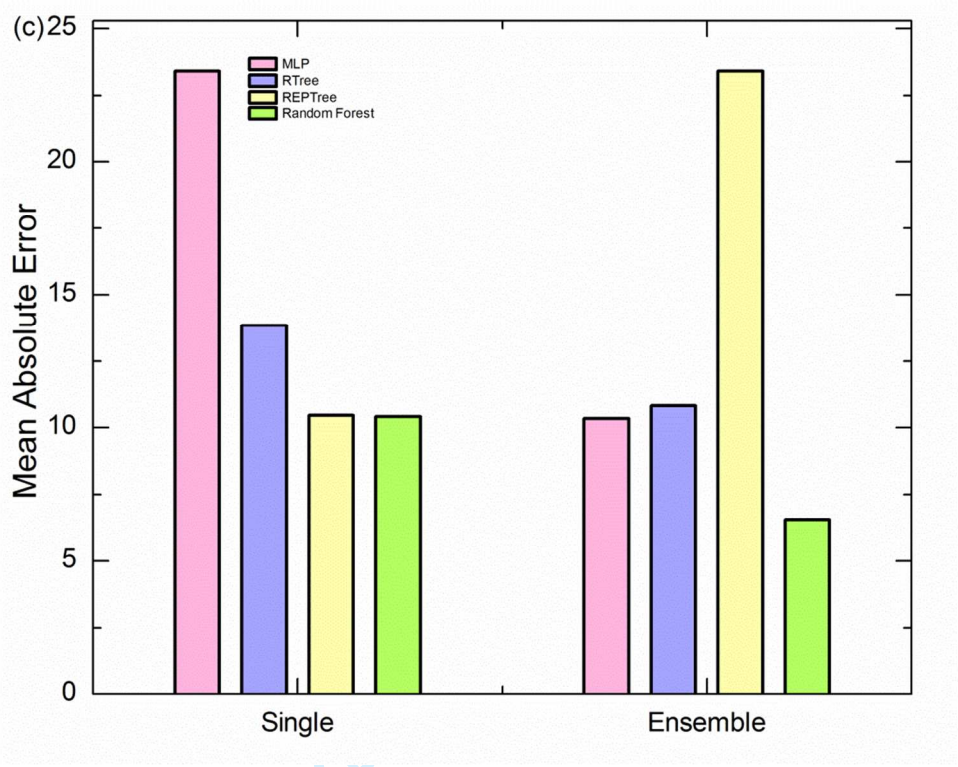
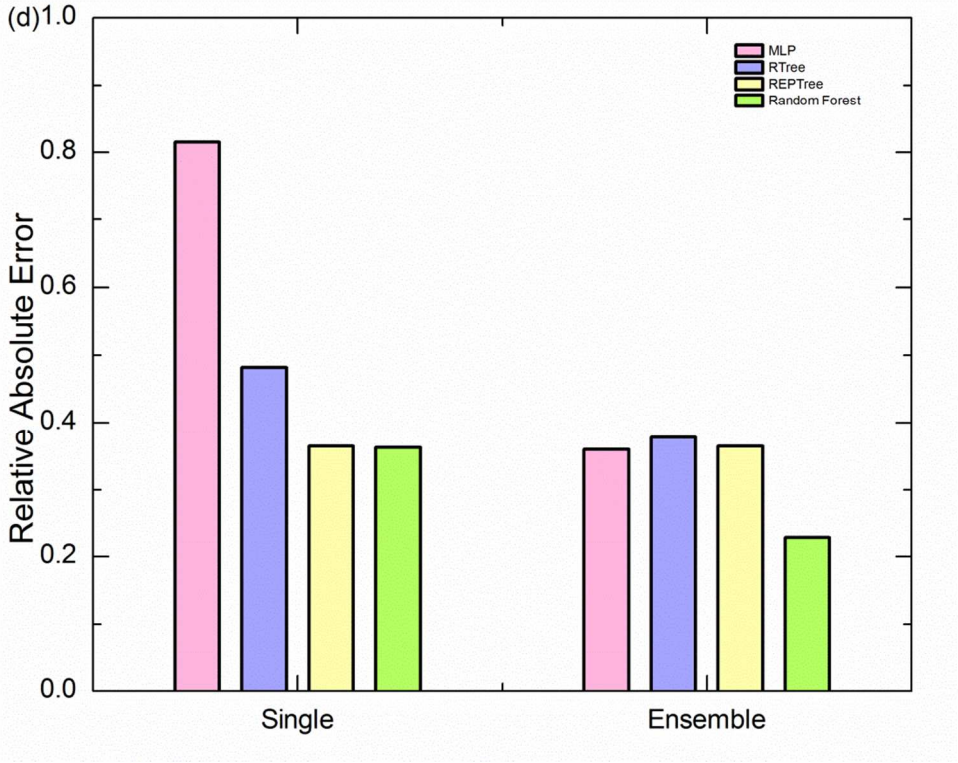


Figure 6. Improvement of R^2 statistic for various iteration and cluster size in case of (a) MLP (b) RTree (c) REPTree (d) random forest


(a)
1.0
0.8
0.6
0.4
0.2
0.0
Correlation coefficient
MLP
RTree
REPTree
Random Forest
Single
Ensemble


(b)
25
20
15
10
5
0
Root Relative Squared Error
MLP
RTree
REPTree
Random Forest
Single
Ensemble


(c)
Mean Absolute Error
0
5
10
15
20
25
MLP
RTree
REPTree
Random Forest
Single
Ensemble


(d)
Relative Absolute Error
0.0
0.2
0.4
0.6
0.8
1.0
MLP
RTree
REPTree
Random Forest
Single
Ensemble

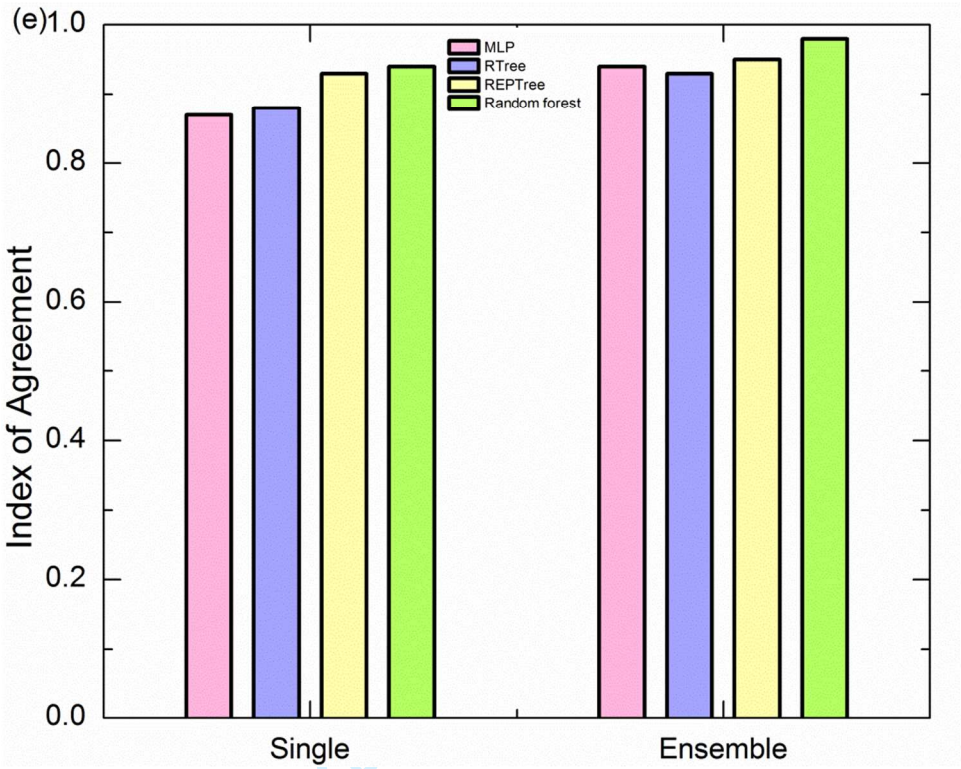


Figure 7. Performance comparison of single models and ensemble models in terms of (a) CC (b) RRSE (c) MAE (d) RAE (e) IA

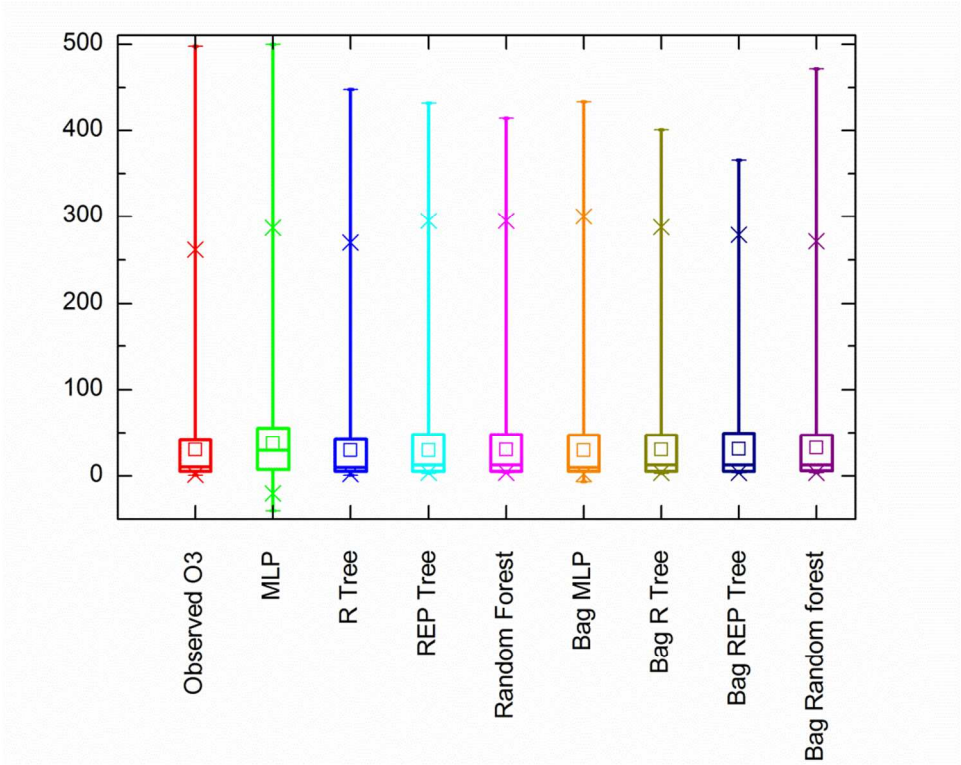
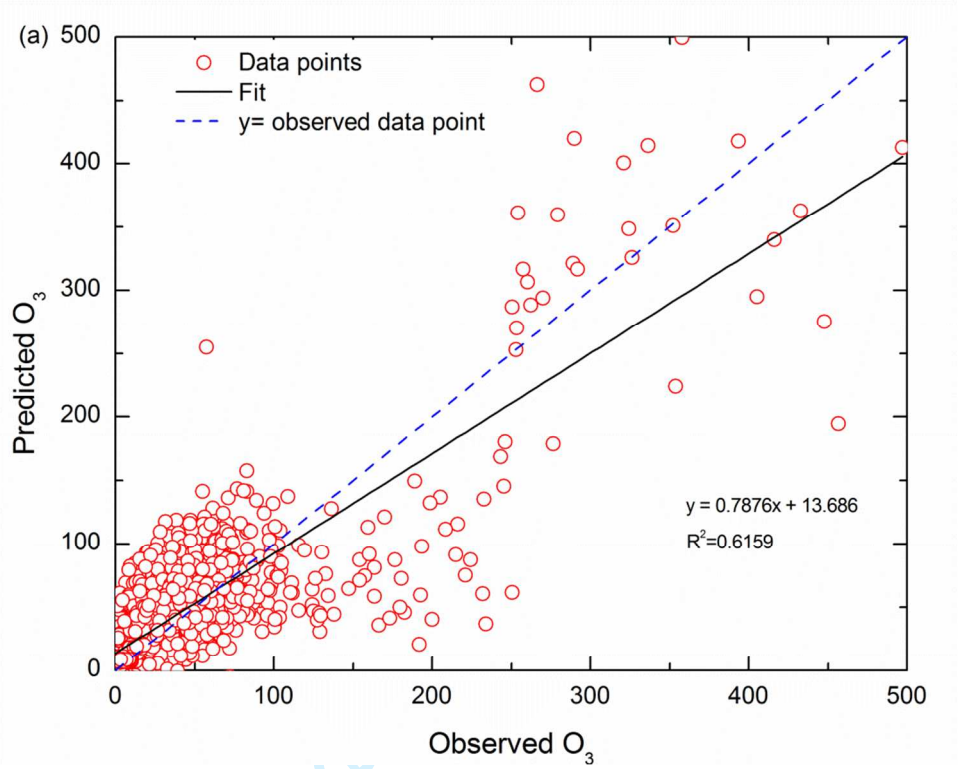
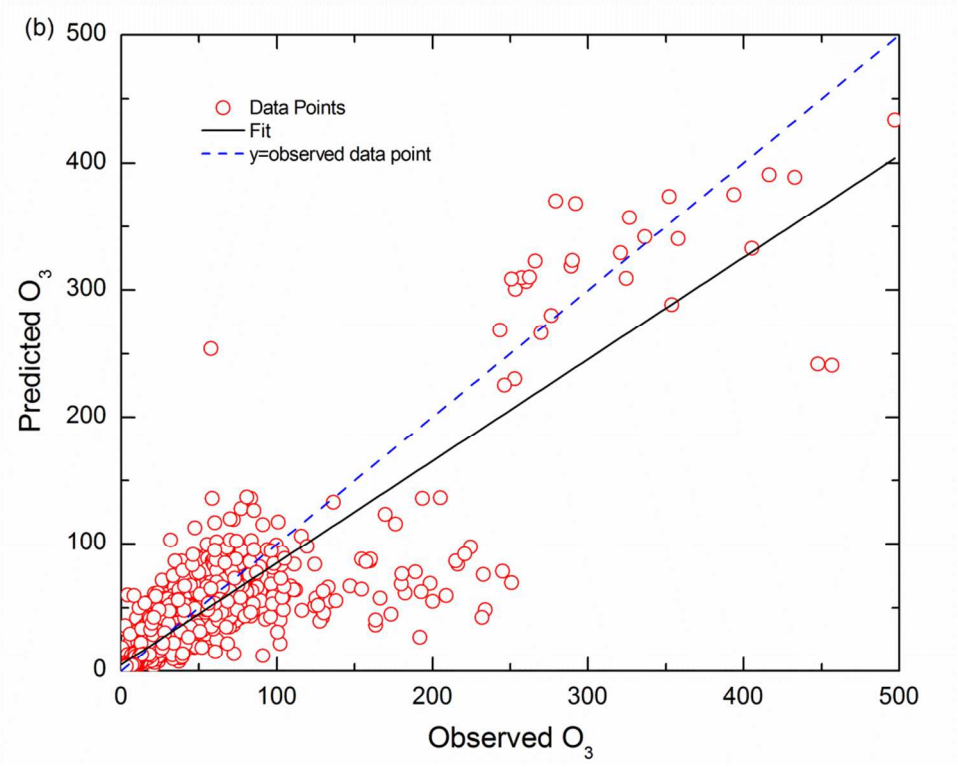
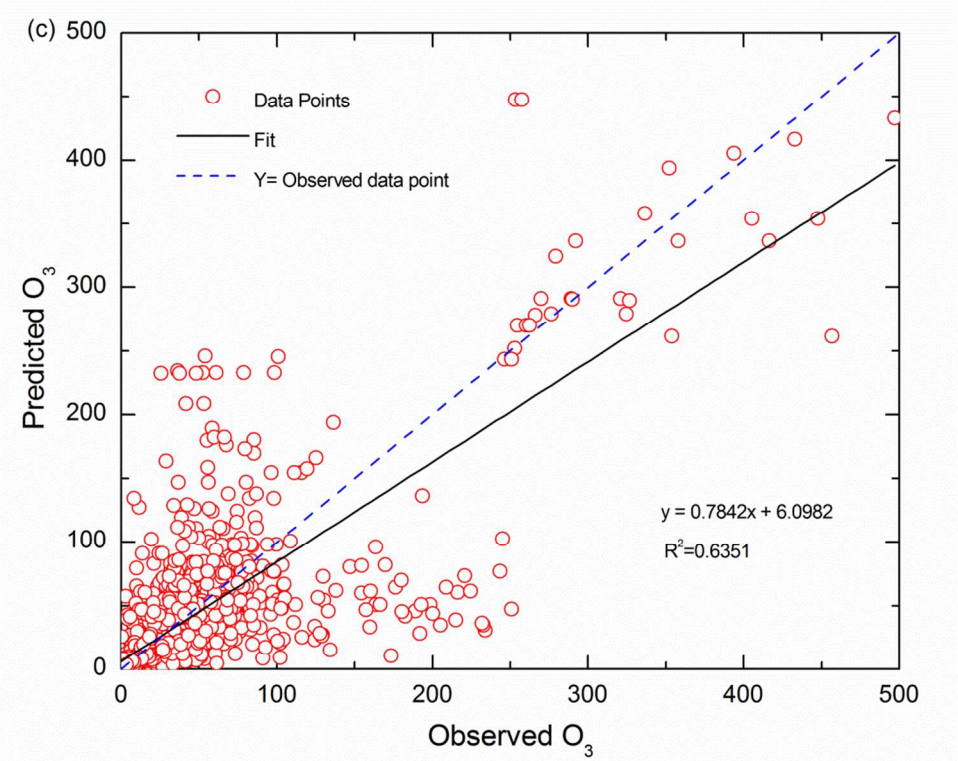
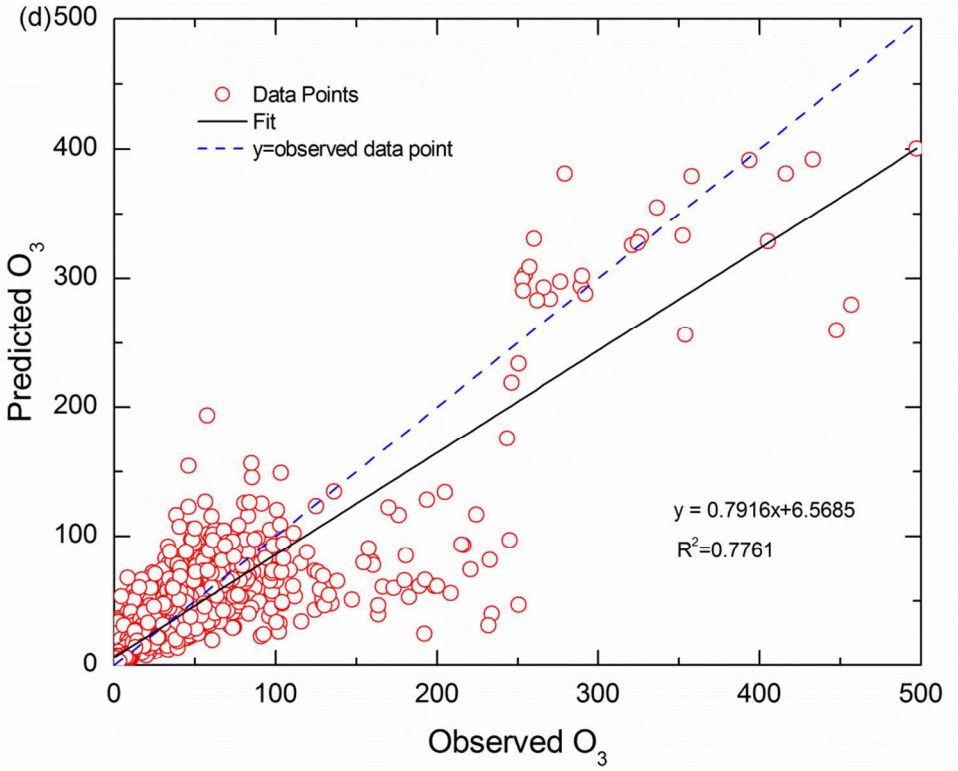


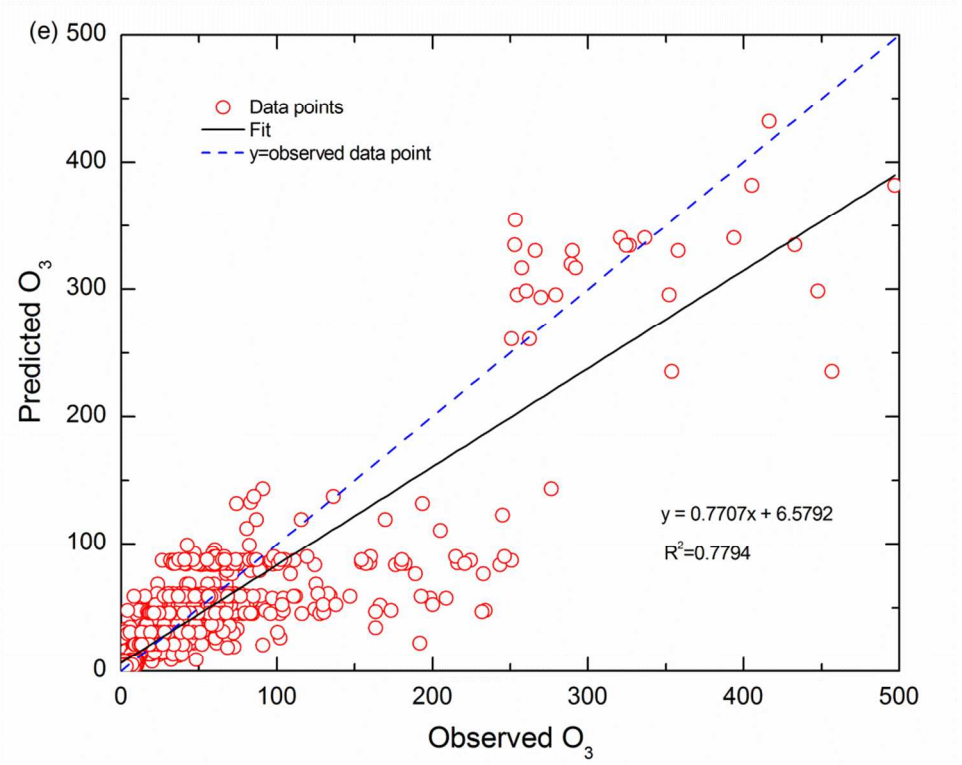
Figure 8. Box plot representation of observed vs. predicted concentration for different models

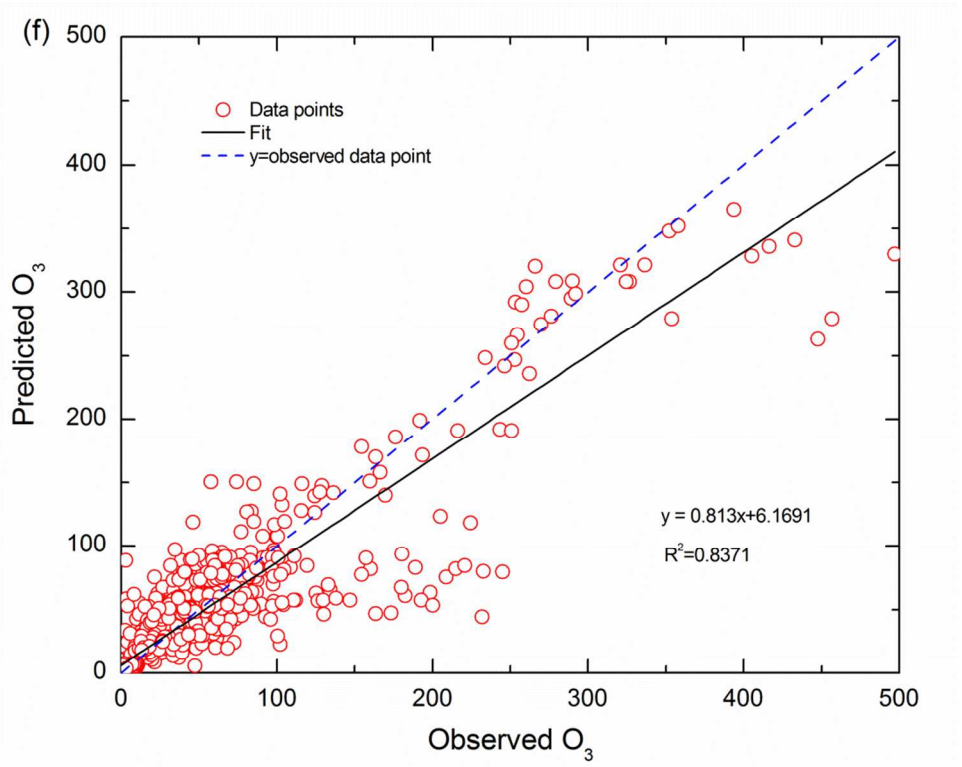

(a)
Data points
Fit
y= observed data point
y = 0.7876x + 13.686
R2=0.6159
Predicted O3
Observed O3
0
100
200
300
400
500


(b)
Data Points
Fit
y=observed data point
Predicted O3
Observed O3
0
100
200
300
400
500


(c)
Data Points
Fit
Y= Observed data point
y = 0.7842x + 6.0982
R2=0.6351
Predicted O3
Observed O3
0
100
200
300
400
500


(d)
Data Points
Fit
y=observed data point
y = 0.7916x+6.5685
R2=0.7761
Predicted O3
Observed O3
0
100
200
300
400
500


(e)
Data points
Fit
y=observed data point
y = 0.7707x + 6.5792
R2=0.7794
Predicted O3
Observed O3
0
100
200
300
400
500


(f)
Data points
Fit
y=observed data point
y = 0.813x+6.1691
R2=0.8371
Predicted O3
Observed O3
0
100
200
300
400
500

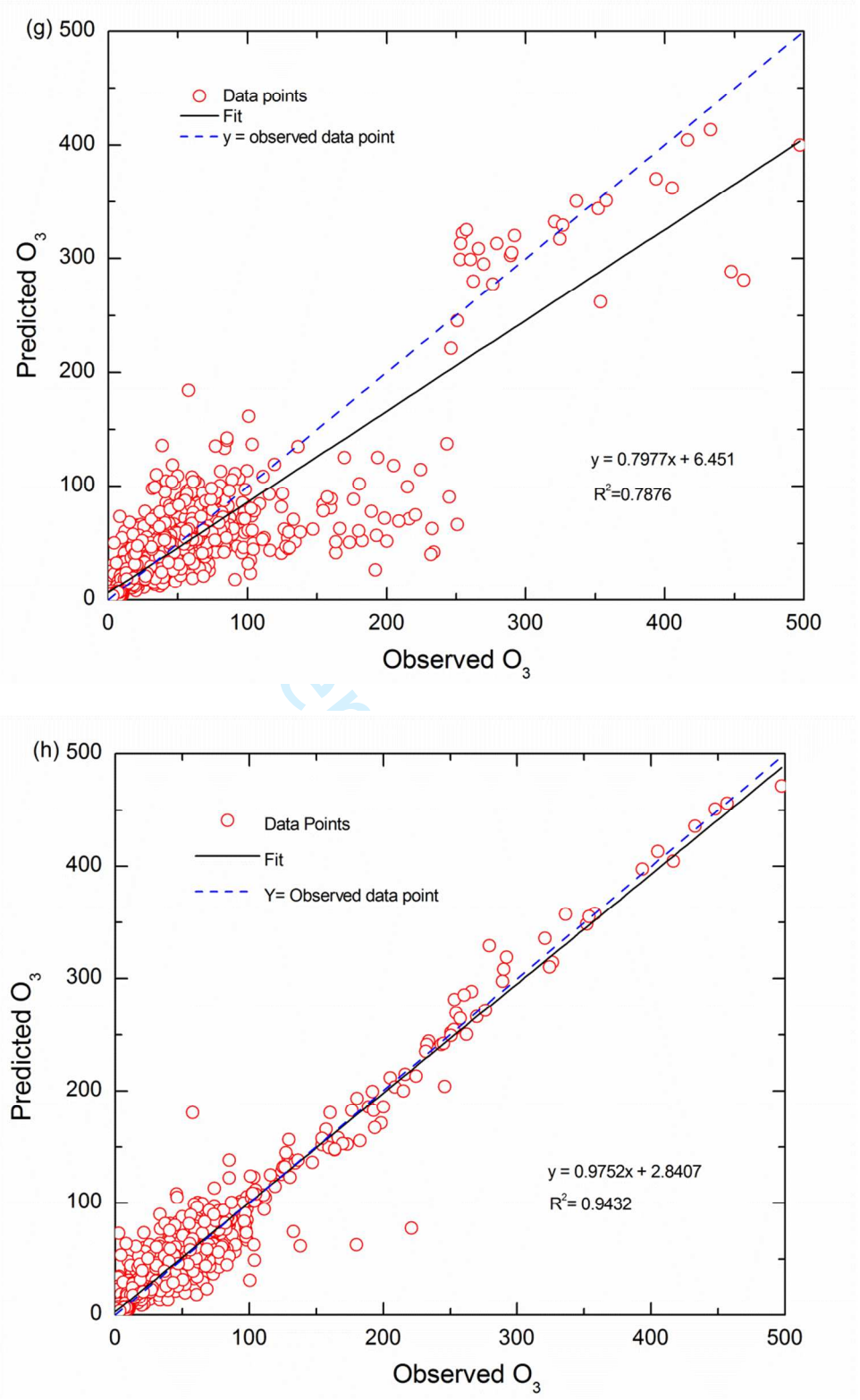


Figure 9. Scatter plots of observed vs. predicted ozone for (a) MLP (b) bagged MLP (c) RTree (d) bagged RTree (e) REPTree (f) bagged REPTree (g) random forest (h) bagged random forest

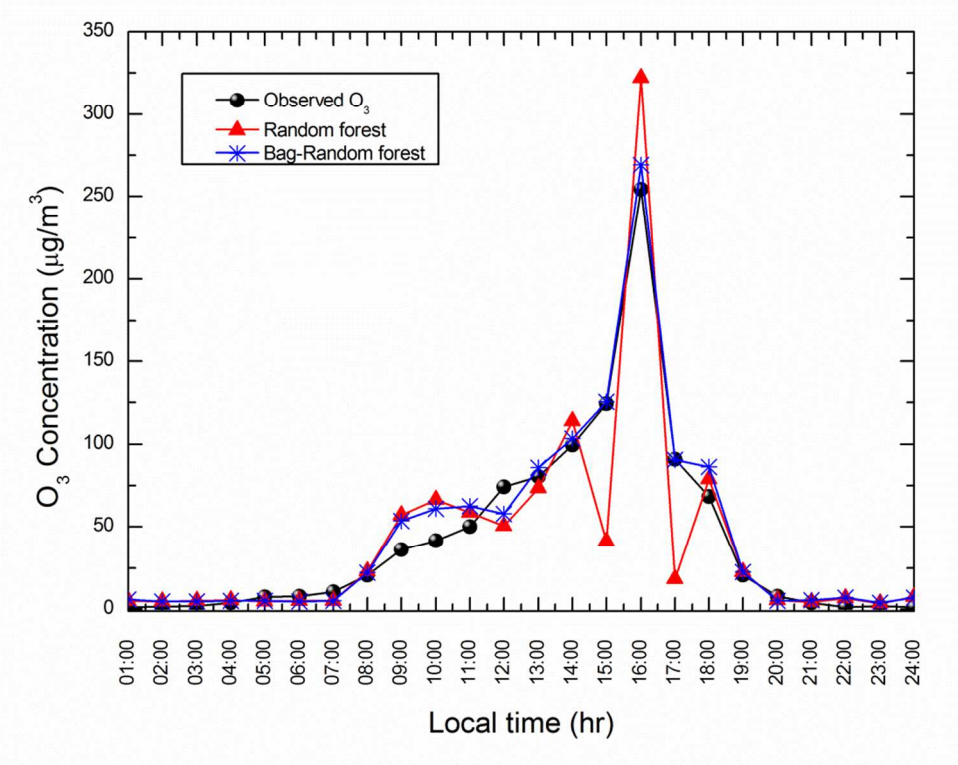


Figure 10. Prediction of hourly O_3 concentration using single classifier and ensemble classifier for independent data set. (Plot is shown for 1 day for clear understanding)