Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-753-SC1, 2019

Atmospheric Chemistry and Physics
Discussions

# *Interactive comment on* "Improving the prediction of an atmospheric chemistry transport model using gradient boosted regression trees" *by* Peter D. Ivatt and Mathew J. Evans

**Hyun Soo Kim**

hskim98@gist.ac.kr

Received and published: 5 November 2019

Ivatt and Evans developed a gradient-boosted decision tree model (XGBoost) to correct the bias associated with GEOS-Chem-predicted O3 concentrations. They found that the bias-corrected model estimates ozone concentrations more accurately than the uncorrected model. The use of machine learning algorithms for air quality applications is a hot topic. I suggest that the following points need to be addressed to improve the quality of the manuscript.

1. This paper utilized the XGBoost model for bias correction of GEOS-Chem-predicted O3 concentrations. In the model training, the ground (EMEP, EPA, and GAW) and

ozone-sonde (WOURDC) observations were used. During the pre-processing of the training data set, the data comprising O3 concentrations above 100 ppbv were excluded. In general, a decrease in the maximum value of the target tends to increase the accuracy indicator of the machine learning model. However, the accuracy expressed in numerical values cannot always indicate the optimization level of the trained model. It is more logical to exclude the effects of stratospheric ozone using altitude information obtained from ozonesonde and ATom observations.

2. In the model development, the author presented only two important hyperparameters (depth and tree number) of the XGBoost model. One of the most important aspects of this study is the optimization of bias correction via logical development. Therefore, the results of sensitivity test conducted to determine important structural parameters (e.g., depth, number of trees, learning rate, tree boosting algorithm, and sub-sampling rate) should be provided.

3. In the model training, the K-fold algorithm was applied. The training and validation data set used observations from 2010 to 2015, which were divided into five blocks. The hyperparameters of the XGBoost model were determined by conducting the five independent model trainings. However, because the validation data sets were utilized K-1 times to optimize weight and bias matrix, the K-fold algorithm may be associated with the risk of training data leakage. Therefore, this issue should be addressed in the manuscript.

4. During the description of independent variables, the author only listed 81 variables as input features. In general, the accuracy of machine learning model is mainly attributed to the combination of input variables. In addition, imbalanced data set plays an important role in the performance of machine learning model. Therefore, it is necessary to use integrated analysis to identify the characteristics of the independent variables. Further, the close correlation between input features can distort the regression coefficients. In order to minimize the errors associated with munticorrelinearity, the author should provide a detailed analysis or rationale for the determination of input variables.

5. There is no detailed description of the observations. If the amount of the ground-based observations is much larger than that of the ozonesonde observations, it can affect the overall results of bias corrections. Therefore, the evaluation accuracy with ATom is clearly lower than that with the ground observations.

6. Data pre-processing has not been explained in detail. There are several ways involving data normalization. Several studies have suggested normalization methods to improve the performance of the predictive models. Optimization of data pre-processing methods should be addressed in the manuscript.

7. The grid resolution of GEOS-Chem was 4 ÌŁ x 5 ÌŁ. In order to prepare training and validation data sets, the observations were preprocessed to match with the grid of the GEOS-Chem. However, the limits on spatial representation of the observations are likely to cause sub-grid problems (i.e. the current grid size of GEOC-Chem is too large). The comparative results involving the Japan case clearly indicate the possibility of this problem. Therefore, it is necessary to develop a bias correction model with a higher resolution.

8. In this study, the importance of input variables was estimated based on gain. The importance of input features can be analyzed using various criteria (e.g. gain, weight and cover). Since these analyses can provide new scientific insights, a more comprehensive analysis of features based on various criteria is required.

---

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-753, 2019.