Point-by-point responses to the comments & suggestions from the editor

Journal: Geoscientific Model Development

Manuscript ID: EGUSPHERE-2023-1531

Title: "Diagnosing drivers of $PM_{2.5}$ simulation biases in China from meteorology, chemical composition, and emission sources using an efficient machine learning method"

Comments from the editor:

Please reply to the reviewers' comments carefully.

- We appreciate the editor's efforts on this manuscript. In the revision, we carefully revised the manuscript based on reviewers' comments and made detailed responses. We hope to meet the requirements of the journal. Please contact us if there are any problems.

Comments from Reviewer #1:

Summary: In this manuscript, the author introduces LightGBM, a tree-based regression method, as a powerful tool for evaluating the performance of the Community Multiscale Air Quality (CMAQ) model. The primary focus is on diagnosing the CMAQ's effectiveness in pinpointing the predominant contributing factors responsible for prediction bias, particularly in relation to the prediction of PM 2.5 concentration. To comprehensively assess potential biases associated with each source, LightGBM is employed to conduct separate time series regressions for features grouped into three major sources

Major comments:

After reading the manuscript, I think some major comments from Anonymous Reviewer #1 of last round are still not adequately addressed. In my opinion, the author should clearly address the following aspects:

- Thank you for your feedback and time to review our manuscript. We apologize for not adequately addressing the comments in the last revision. We have made careful revisions and have provided detailed responses to each of your comments in this revision. Thank you again for your time and consideration. We look forward to

working closely with you to address your concerns and make the necessary improvements.

1. Dataset setting:

a. Provide a clear description of the 350,000 valid observations, specifying whether it represents the sum of all time series data points across multiple monitoring stations. Clearly state the methodology for training and testing data separation.

- Thanks for the comment. The observation data is the sum of all time series data points across multiple monitoring stations. We have added the specific description in Section 2.1. The training and testing data were randomly separated in an 8:2 ratio for all stations in the region of interest. We have clarified it in Section 2.3.

Changes in Lines 65-66: "A total of about 350,000 observations meeting quality control criteria were selected from all time series data points across multiple monitoring stations."

Changes in Lines 118-119: "The separate test sets (not involved in the training and CV hyperparameter selection process) were divided by randomly sampling 20% of the data from all stations in the region of interest."

b. Clarify how random samples of observations are selected. Specify whether the 20% random sampling is performed at the station level or across all stations in the region of interest.

- Thanks for the comment. The 20% random sampling was performed across all stations in the region of interest. We have clarified it in Section 2.3.

Changes in Lines 118-119: "The separate test sets (not involved in the training and CV hyperparameter selection process) were divided by randomly sampling 20% of the data from all stations in the region of interest."

c. Time series data usually cannot be directly learned through tree-based model without additional pre-processing/feature engineering. Discuss the absence of data preparation and feature engineering before feeding data into tree-based models. If temporal structure is considered negligible, provide justification; otherwise, explain the approach taken to handle temporal aspects.

- Thanks for the comment. We combined time series data from multiple stations, eliminated the extreme values of 0.1%, and did not specifically preprocess the temporal structure. In previous studies of pollution prediction based on tree models, time-related features are usually added to characterize the temporal pattern of pollutant changes to further improve the prediction ability, e.g., Wei et al. (2021a) improved the model performance by adding temporal features of day of year and Unix timestamps. However, the goal of this study is to identify the contributors to the simulation bias based on feature importance rather than prediction, and the inclusion of temporal features cannot provide any useful information for us instead it is difficult to attribute them to meteorological or emissions contributions. For example, the simulated bias shows a clear temporal pattern, being larger in the winter and smaller in the summer, so temporal features would show a high contribution to the simulation bias, but provide no valid information. Therefore, we did not include temporal features in our model.

Changes in Lines 268-272: "Previous pollution prediction studies based on tree models usually added the time-related features to describe the temporal pattern of pollutant changes to further improve the prediction ability, e.g., Wei et al. (2021a) improved the model performance by adding temporal features of day of year and Unix timestamps. However, the inclusion of temporal features cannot provide any useful information about contributors of simulation biases instead it is difficult to attribute them to meteorological or emissions contributions.Therefore, temporal features were not included in our model."

2. Tree-based model justification

a. In the section (L95-100), provide examples of similar applications in terms of dataset, model, and research area. Demonstrate why tree-based methods are suitable for the specific dataset. Justify the selection of tree-based models beyond considerations of memory and computation time.

- Thanks for the comment. We have provided examples of similar applications. Tree-based ML models typically outperform deep learning approaches in tabular data due to limited data number and relative sample structure (compared to image, video, and natural language). The LightGBM model has shown accurate performance in

many fields, with fast speeds, less overfitting, and independence from collinearity. We have justified the selection of tree-based models in Section 2.3.

Changes in Lines 95-109: "Tree-based ML models typically outperform deep learning approaches in tabular data (e.g., air pollutant observation datasets), and thus have been widely developed (Grinsztajn et al., 2022). Wei et al. (2021a) compared several models when reconstructing $PM_{2.5}$ data records in China and found that the tree model showed superior performance. The LightGBM model is an optimized Gradient Boosting Decision Tree (GBDT) (Ke et al., 2017), and has shown accurate performance in many fields (Wei et al., 2021b; Yan et al., 2021; Sun et al., 2020; Liang et al., 2020). Compared to XGBoost, a widely used GBDT, LightGBM uses Histogram's decision tree algorithm along with Gradient-based One-Side Sampling (GOSS), which saves memory and computation time (Ke et al., 2017). Three tree-based models, Random Forest, XGBoost, and LightGBM, were compared in our previous study (Wang et al., 2023). Using the same input data and hyperparameters, LightGBM is as accurate as XGBoost, but faster and less overfitting (the difference in accuracy between training and testing). Besides, Multiple colinearities between features such as pollutant concentrations and meteorological factors can greatly affect the performance of traditional linear models. When independent variables are correlated, changes in one variable are associated with changes in the other, making it difficult for the model to independently estimate the relationship between each independent and dependent variable. However these collinearities does not affect the performance of tree-based models like Random Forest and LightGBM, because they do not require the assumption of feature independence (Belgiu and Drăguţ, 2016; Chen et al., 2016; Ke et al., 2017). So the lightGBM model was used to diagnose $PM_{2.5}$ simulation biases in this study."

b. Introduce a discussion on multicollinearity in the methodology section.

- Thanks for the comment. We have added the discussion of multicollinearity in Section 2.3.

Changes in Lines 103-109: "Besides, Multiple colinearities between features

such as pollutant concentrations and meteorological factors can greatly affect the performance of traditional linear models. When independent variables are correlated, changes in one variable are associated with changes in the other, making it difficult for the model to independently estimate the relationship between each independent and dependent variable. However, these collinearities do not affect the performance of tree-based models like Random Forest and LightGBM, because they do not require the assumption of feature independence (Belgiu and Drăguţ, 2016; Chen et al., 2016; Ke et al., 2017)."

3. Cross validation

a. Clearly explain how cross-validation is performed and provide a statement on how the two metrics (R^2 and RMSE) influence model selection decisions. Clearly articulate the criteria for jointly considering model performance using these two metrics.

- Thanks for the comment. Cross-validation (5-fold) combined with RMSE was used to select hyperparameters. Two metrics (R^2 and RMSE) were used to evaluate the model performance in the separate test sets. Higher R^2 and lower RMSE represent better model performance. We have clarified cross-validation and model evaluation in Section 2.3.

Changes in Lines 114-119: "Cross-validation (5-fold) combined with RMSE was used to select hyperparameters. The dataset was randomly divided into five parts, one was taken in turn as the test set and the rest was used for training, which was repeated five times and the average test RMSE was calculated. Looping to increase model complexity, ending the loop and returning to the hyperparameters when the model test RMSE does not decrease significantly (< 0.01) or the gap between training and test RMSE increases significantly (< 0.05). The separate test sets (not involved in the training and CV hyperparameter selection process) were divided by randomly sampling 20% of the data from all stations in the region of interest."

Changes in Lines 109-111: "Two metrics were calculated to evaluate model performance, including the coefficient of determination (R^2) and the root mean square error (RMSE) (Wei et al., 2020)."

Minor comments:

L15. Clarify the term "efficient" to provide a precise understanding within the context of this study.

- Thanks for the comment. We defined the “efficient” as “fast speed and low requirement of computational resources”, and have reorganized the corresponding expression.

Changes in Lines 14-18: “Accurate diagnosis of simulation biases is critical for improvement of models, interpretation of results, and management of air quality, especially for the simulation of fine particulate matter ($PM_{2.5}$). In this study, an efficient method with fast speed and low requirement of computational resources based on tree-based machine learning (ML) method, the Light Gradient Boosting Machine (LightGBM), was designed to diagnose CTMs simulation biases.”

L16. Instead of broadly referring to "machine learning," explicitly specify that LightGBM is a tree-based method. Additionally, consider breaking the sentence into two for enhanced readability.

- Thanks for the comment. We changed the "machine learning" to “tree-based machine learning (ML) method, the Light Gradient Boosting Machine (LightGBM)”. We split the sentence in two and reorganized the expression accordingly.

Changes in Lines 16-20: “In this study, an efficient method with fast speed and low requirement of computational resources based on tree-based machine learning (ML) method, the Light Gradient Boosting Machine (LightGBM), was designed to diagnose CTMs simulation biases. The drivers of the Community Multiscale Air Quality (CMAQ) model biases compared to observations in simulating $PM_{2.5}$ concentrations from three perspectives of meteorology, chemical composition, and emission sources.”

L20. Reevaluate the assertion that an R^2 value of 0.68 constitutes good performance. Provide references from existing literature to substantiate this claim. Additionally, the relative performance gap of 0.16 is about 23.5% of 0.68, which might not be

compelling enough; its significance in the context of overfitting and the ability to be applied to other fields is weak.

- Thanks for the comment. We have reevaluated the model. The ML models were separately trained by meteorology, $PM_{2.5}$ components, and source sectors for different regions and seasons, and the test sets were used to measure the model performance (Figure 4). The meteorology, $PM_{2.5}$ components, and source sectors can partially explain the simulation bias, with mean test R^2 of 0.64, 0.52, and 0.50, respectively, and the RMSE was 17.41, 19.82, and 19.56 $\mu g/m^3$, respectively. We removed inappropriate content and have changed the description of the overfitting issue.

Changes in Lines 167-172: "The ML models were separately trained by meteorology, $PM_{2.5}$ components, and source sectors for different regions and seasons, and separate test sets were used to measure the model performance (Figure 4). All three feature combinations can partially explain the simulation bias. The mean test R^2 for meteorology, $PM_{2.5}$ components, and source sectors were 0.64, 0.52, and 0.50, respectively, and the RMSE were 17.41, 19.82, and 19.56 $\mu g/m^3$, respectively. The model performed better in summer than in winter. This may be attributed to the high simulation biases in winter due to severe $PM_{2.5}$ pollution and complex causes, while $PM_{2.5}$ pollution in summer is lighter with lower CMAQ simulation bias."

Changes in Lines 266-267: "Besides, LightGBM shows comparable accuracy to XGBoost but is faster and shows smaller accuracy gaps between training and testing with less overfitting."

L65. Revise the description of "valid" observations to explicitly convey that these observations adhere to the quality control criteria outlined in L62-64. Reorganize the sentences for better coherence.

- Thanks for the comment. We have revised the description of "valid" to "meeting quality control criteria", and reorganized the sentences for better coherence.

Changes in Lines 65-66: "A total of about 350,000 observations meeting quality control criteria were selected from all time series data points across multiple monitoring stations."

L76. Consider either elaborating on the model's enhancements or removing the sentence for conciseness.

- Thanks for the suggestion. We took your advice and removed the sentences for conciseness.

L80-82. Clearly indicate that CMAQ is employed for simulating PM 2.5 components when introducing the model. Adjust the sequence of information to improve logical flow.

- Thanks for the comment. We have adjust the sequence of the CMAQ introduction for better consistency.

Changes in Lines 72-73: "The CMAQ simulation (36 km×36 km) was carried out to simulate $PM_{2.5}$ components in mainland China and surrounding regions in 2019. The list of $PM_{2.5}$ components simulated by CMAQ is shown in Table A1."

L86. Enhance the fluency by adding a connecting word at the beginning of the sentence.

- Thanks for the comment. We have reorganized the presentation to make the sentences more coherent and checked the coherence of the entire manuscript.

Changes in Lines 85-86: "PPM from different source sectors are tracked by non-reactive tracers (10^{-5} of the PPM emission rates), and source-specific PPM concentrations are then calculated by multiplying the tracer with 10^5."

L119-120. Define "success" in quantitative or qualitative terms to provide a clearer understanding of the criteria for evaluating success.

- Thanks for the comment. We eliminated the vague expression "success" and quantified it with specific statistics.

Changes in Lines 132-134: "Observed $PM_{2.5}$ concentrations were highest in BTH (51.172 $\mu g/m^3$) and lowest in PRD (28.273 $\mu g/m^3$). The CMAQ model underestimates $PM_{2.5}$ concentrations of -8.59 $\mu g/m^3$, -2.66 $\mu g/m^3$, -6.21 $\mu g/m^3$, and -19.25 $\mu g/m^3$ in the BTH, YRD, PRD, and NWCHN, respectively (Figure 1b)."

L170. Remove the extra period before the citation.

- Thanks for the comment. We have removed the extra period before the citation, and we apologize for our carelessness and have carefully examined the entire manuscript.

L245. Replace "Features collinearity" with "Multicollinearity among features."

- Thanks for the comment. We have replaced "Features collinearity" with "Multicollinearity among features."

Changes in Lines 252-253: "Multicollinearity among features does not affect the performance of tree-based models like Random Forest and LightGBM"

Table A6. Use bold font to highlight the best metric performance. Additionally, if XGB and LGB exhibit similar performance, with XGB slightly superior, consider including computational time as an additional metric to justify the preference for LGB over XGB.

- Thanks for the comment. We have highlighted the best metric performance and include computational time as an additional metric in Table A6.

Changes in Lines 266-267: "Besides, LightGBM shows comparable accuracy to XGBoost, but is faster and shows smaller accuracy gaps between training and testing with less overfitting."

Comments from Reviewer #2:

This manuscript requires additional revisions.

1. Cross-validation is mainly used for hyperparameter tuning, to select the best hyperparameter combination and prevent overfitting to the training data. After selecting the final model, we still need to evaluate performance on an independent test set to check the model's ability to generalize to real data. However, I noticed the evaluation of model performance in this paper is still based on 5-fold cross-validation (e.g., Lines 110-111 and Lines 158-159). The authors should rewrite Section 3.2 by using the independent testing data for model evaluation instead of cross-validation.

- Thanks for the comment. We have rewritten the first part of Section 3.2 by using the independent testing data for model evaluation. Specifically, The ML models were separately trained by meteorology, $PM_{2.5}$ components, and source sectors for different regions and seasons, and separate test sets were used to measure the model

performance (Figure 4). The 5-fold cross-validation combined with RMSE was used to select hyperparameters.

Changes in Lines 167-172: "The ML models were separately trained by meteorology, $PM_{2.5}$ components, and source sectors for different regions and seasons, and separate test sets were used to measure the model performance (Figure 4). All three feature combinations can partially explain the simulation bias. The mean test R^2 for meteorology, $PM_{2.5}$ components, and source sectors were 0.64, 0.52, and 0.50, respectively, and the RMSE was 17.41, 19.82, and 19.56 $\mu g/m^3$, respectively. The model performed better in summer than in winter. This may be attributed to the high simulation biases in winter due to severe $PM_{2.5}$ pollution and complex causes, while $PM_{2.5}$ pollution in summer is lighter with lower CMAQ simulation bias."

Changes in Lines 114-119: "Cross-validation (5-fold) combine with RMSE to select hyperparameters. The dataset was randomly divided into five parts, one was taken in turn as the test set and the rest was used for training, which was repeated five times and the average test RMSE was calculated. Looping to increase model complexity, ending the loop and returning to the hyperparameters when the model test RMSE does not decrease significantly (< 0.01) or the gap between training and test RMSE increases significantly (< 0.05). The separate test sets (not involved in the training and CV hyperparameter selection process) were divided by randomly sampling 20% of the data from all stations in the region of interest."

2. LightGBM has two types of feature importance, namely "split" and "gain." Could you please clarify which feature importance type was used for analysis in this study? I think clearly mentioning the type used would strengthen the analysis.

- Thanks for the comment. We used "gain" to measure feature importance, which is the size of the gain resulting from splitting through a certain feature. The type of "split" is the number of splits using a particular feature. For some highly indicative categorical features that may split only once during tree growth, but have high importance, at which time the split method may be inaccurate. So we used "gain" to measure feature importance.

Changes in Lines 120-122: "The target variable was set to be the difference between observed and simulated daily $PM_{2.5}$ concentrations, and the key contributors

to the simulation bias were then determined by the relative importance (calculated by gain) of the input features (Ye et al., 2022; Loyola-González et al., 2023)"

Reference

Belgiu, M. and Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions, ISPRS-J. Photogramm. Remote Sens., 114, 24-31, 2016.

Chen, T. Q., Guestrin, C., and Assoc Comp, M.: XGBoost: A Scalable Tree Boosting System, 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, Aug 13-17, WOS:000485529800092, 785-794, 10.1145/2939672.2939785, 2016.

Grinsztajn, L., Oyallon, E., and Varoquaux, G.: Why do tree-based models still outperform deep learning on tabular data?, arXiv preprint arXiv:2207.08815, 2022.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems, 30, 2017.

Liang, W., Luo, S., Zhao, G., and Wu, H.: Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms, Mathematics, 8, 765, 2020.

Loyola-González, O., Ramírez-Sáyago, E., and Medina-Pérez, M. A.: Towards improving decision tree induction by combining split evaluation measures, Knowledge-Based Systems, 277, 110832, https://doi.org/10.1016/j.knosys.2023.110832, 2023.

Sun, X., Liu, M., and Sima, Z.: A novel cryptocurrency price trend forecasting model based on LightGBM, Finance Research Letters, 32, 101084, 2020.

Wang, S., Wang, P., Zhang, R., Meng, X., Kan, H., and Zhang, H.: Estimating particulate matter concentrations and meteorological contributions in China during 2000–2020, Chemosphere, 330, 138742, https://doi.org/10.1016/j.chemosphere.2023.138742, 2023.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., and Cribb, M.: Reconstructing 1-km-resolution high-quality PM2. 5 data records from 2000 to 2018 in China: spatiotemporal variations and policy implications, Remote Sensing of Environment, 252, 112136, 2021a.

Wei, J., Li, Z., Pinker, R. T., Wang, J., Sun, L., Xue, W., Li, R., and Cribb, M.: Himawari-8-derived diurnal variations in ground-level PM2.5 pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM), Atmos. Chem. Phys., 21, 7863-7880, 10.5194/acp-21-7863-2021, 2021b.

Wei, J., Li, Z. Q., Cribb, M., Huang, W., Xue, W. H., Sun, L., Guo, J. P., Peng, Y. R., Li, J., Lyapustin, A., Liu, L., Wu, H., and Song, Y. M.: Improved 1 km resolution PM2.5 estimates across China using enhanced space-time extremely randomized trees, Atmos. Chem. Phys., 20, 3273-3289, 10.5194/acp-20-3273-2020, 2020.

Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., Ma, C., Yan, J., and Wang, X.: LightGBM: accelerated genomically designed crop breeding through ensemble learning, Genome Biology, 22, 1-24, 2021.

Ye, X., Wang, X., and Zhang, L.: Diagnosing the Model Bias in Simulating Daily Surface Ozone Variability Using a Machine Learning Method: The Effects of Dry Deposition and Cloud Optical Depth, Environ. Sci. Technol., 56, 16665-16675, 10.1021/acs.est.2c05712, 2022.