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 read the reviewer's comments and reply accordingly.

- We appreciate the editor's efforts on this manuscript. In the revision, we carefully revised the manuscript based on reviewers' comments.

Comments from Reviewer #1:

The authors did not adequately address my comments.

- Thank you for your feedback and for taking the time to review our manuscript. We appreciate your input, and we apologize if it seemed that we did not adequately address your comments in our previous revision. We take your comments seriously and are committed to improving our manuscript to meet your expectations. we have made the necessary revisions and provided a detailed response to each of your comments in this revision. We are dedicated to producing a high-quality manuscript, and your feedback is invaluable in achieving that goal. 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. A separate test dataset, which is excluded from the cross-validation process, is typically employed to validate the mode's performance. However, the test data were no provided by the authors. Ref: https://en.wikipediaorg/wiki/Training,validation,and test data sets#Cross-validation

- Thanks for the comment. We randomly divided a test set (20% of total data) that was not involved in the training and CV hyperparameter selection process, and a separate test dataset has been updated to zendo (https://zenodo.org/records/10283739). Hyperparameter selection and further model training were performed using the training dataset. The hyperparameters selected using only the training data are consistent with the hyperparameters we previously selected using all the data. The model was tested using a combination of meteorological, emission, and $PM_{2.5}$ components features in the test set (Figure S4). The model shows a prediction R^2 of 0.68 and RMSE of 17.26 μg/m^3. We have added the corresponding results in Section 3.2.

Changes in Lines 155-158: "First, 20% of the data (not involved in training) were randomly selected for model evaluation (Figure S4). Training was performed using a combination of $PM_{2.5}$ components, meteorological, and emission features. The model showed a prediction R^2 of 0.68 and RMSE of 17.26 μg/m^3."

2. Poor R2 is not only due to the features also to the algorithm itself and the hyperparameters. How to determine what exactly is the cause? R2 could have been examined after the authors conducted the identical investigation using linear regression, what then is the purpose of LightGBM? Does not this simply because LightGBM produces superior outcomes in comparison to linear regression? And how did the authors ensure the current LightGBM-based mode is better than other models?

- Thanks for the comment. It is indeed difficult to fully distinguish what causes the low R^2. We designed a complementary experiment to measure the impact of the model itself by comparing popular regression models (including Linear Polynomial Regression, Quadratic Polynomial Regression, Random Forest, XGBoost, and LightGBM) with the same features ($PM_{2.5}$ components). The results (Table A6) show that all models show similar performance, e.g., all models show lower R^2 in the winter in the BTH (0.16 - 0.4), and higher R^2 in the SCB region (0.7 - 0.8). This is a side evidence that the low R^2 is more affected by the features than the model itself, as the commonly used regression models can hardly obtain high R^2 with insufficient

explanatory features (e.g., chemical component features in winter in BTH).

LightGBM is used because it can better capture the non-linear relationship between the input features and the target, compensating for the shortcomings of linear models. Linear models can only capture the effects of some linear processes on $PM_{2.5}$ concentrations, e.g. more primary emissions lead to higher $PM_{2.5}$ concentrations when secondary pollution is low. However, the effects of emissions and meteorological factors on $PM_{2.5}$ concentrations are highly non-linear in scenarios with high secondary pollution, for example, high relative humidity increases the total $PM_{2.5}$ concentration by promoting the hygroscopic growth of $PM_{2.5}$ and the production of secondary particulate matter on the one hand, but on the other hand, it promotes the deposition of particulate matter, which reduces the concentration of $PM_{2.5}$. The non-linear models such as LightGBM can better describe the nonlinear process among meteorological-emission-$PM_{2.5}$ concentration and further identify the sources of model bias in CTMs. The model comparison results (Table A6) also show the better prediction ability of the nonlinear models. The LightGBM model shows superior accuracy and robustness than the other models. Therefore, we chose the LightGBM model to identify the source of simulation bias for CTMs. We have added the discussion in Section 3.3.

Changes in Lines 249-258: "I In addition, the main objective of this study was diagnosing the contributors to CMAQ simulation biases using machine learning, rather than for prediction. Since meteorology or emissions can only partially explain the simulation bias, a low R^2 is justified when fitting the model with only meteorology or emissions variables (Figure 4). We designed a complementary experiment to measure the impact of the model itself by comparing popular regression models (including multiple linear regression, polynomial regression (degree:2), Random Forest, XGBoost, and LightGBM) with the same features ($PM_{2.5}$ components). The results (Table A6) show that all models show similar performance, e.g., all models show lower R^2 in the winter in the BTH (0.16 - 0.4), and higher R^2 in the SCB region (0.7 - 0.8). This is a side evidence that the low R^2 is more affected by the features than the model itself, as the commonly used regression models can hardly

obtain high R^2 with insufficient explanatory features (e.g., chemical component features in winter in BTH)."

3. Why not develop distinct models for each chemical component individually?

- Thanks for the comment. The main reason is that the observation data of chemical components are not openly available in China. The observation network of chemical components has been set up in China, but the data are closed-source. Currently, there are open-source machine learning based chemical reanalysis datasets in China, like (Liu et al., 2022; Wei et al., 2023), however, due to their high uncertainty, we cannot use them as the true values to build models. We attempted to crawl the data from the literature, but the quantity and quality of the data was insufficient. We hope to make China's air quality observation data more open source in the future. Using sufficient observed data on chemical composition and combining it with machine learning models, the sources of bias in CTM can be more accurately identified to guide model improvement.

4. Why not separate each chemical components bias into its own model ?

- Thanks for the comment. As mentioned above, the main reason that we do not model each chemical bias separately is that there are no publicly available observations of chemical composition in China. In the future, we hope to strengthen cooperation and promote data sharing to more accurately identify CTMs simulation deviations and guide CTMs' improvement.

5. What effect do feature interactions have?

- Thanks for the comment. For LightGBM, the interaction between features (multicollinearity) does not affect model predictive power. LightGBM uses a Leaf-wise Tree Growth algorithm, a node-splitting strategy that is less affected by covariance (Ke et al., 2017). The most extreme case of multicollinearity is when there are two identical features. When one feature is used, the decision tree will not use another feature because it adds no new valid information. The multicollinearity

between features will affect features' relative importance. If two variables are correlated, the importance of both will slightly decrease. Previous studies (Hou et al., 2022; Ye et al., 2022) have used ML to explain the causes of air pollution and model bias, and although there was multicollinearity between the input features they used, they got reliable conclusions, showing the slight influence of multicollinearity and the reliability of tree-based machine learning methods.

Changes in Lines 244-249: "Although we filtered the features according to their relative importance and priori knowledge, collinearity still exists among the input features. Features collinearity does not affect the performance of tree-based models like Random Forest and LightGBM (Belgiu and Drăguţ, 2016; Chen et al., 2016; Ke et al., 2017), but the contribution of a single feature might be slightly influenced by other features. Previous studies (Hou et al., 2022; Ye et al., 2022) have used ML to explain the causes of air pollution and model bias, and although there was multicollinearity between the input features they used, they got reliable conclusions, showing the slight influence of multicollinearity and the reliability of tree-based machine learning methods."

Comments from Reviewer #2:

General comments: This study uses ML algorithms to determine the source of CTM bias, it is an interesting and innovative study and the approach has the potential to be generalised to similar studies. After revisions, the manuscript has been greatly improved and here are a few minor issues that need to be addressed.

- Thank you for your feedback and for taking the time to review our manuscript. We carefully address the concern and provide a detailed response to each of your comments in the revision.

Specific comments:

1. It is suggested to add the description of observation data, because this study is based

on it, including the number and distribution of stations, and the number of effective observation dramas.

- Thanks for the comment. We have added the description of observation data to give more information.

Changes in Lines 65-67: "A total of about 350,000 valid observations were selected. The distribution of observation sites (about 1200) is shown in Figure S1. The stations are unevenly distributed, with dense stations in eastern populated areas and sparse stations in western Xinjiang and Tibet."

2. In the abstract, please show important results of model performance.

- Thanks for the comment. We have added the model performance results in abstract. The ML model can capture the complex relationship between input variables and simulation bias well (test R^2 = 0.68). Small performance gap between training and testing indicated model's good generalization ability (delta R^2: 0.16 – 0.18).

Changes in Lines 19-21: "The ML model can capture the complex relationship between input variables and simulation bias well (test R^2 = 0.68) with small performance gap between training and validation (delta R^2: 0.16 – 0.18)."

3. L45 Add full name of XGboost.

- Thank you for your comment. We have added the full name of XGboost, as "eXtreme Gradient Boosting", which is an optimised implementation of Gradient Boosting Decision Trees that improves speed and performance.

4. L54: "lightGBM" or LightGBM? Please make it case-sensitive.

- Thanks for the comment. We uniformly modified the expression: 'LightGBM', and scrutinised the whole manuscript

5. L111: Add formula of R2 and RMSE or reference.

- Thanks for the comment. We added the formula of R^2 and RMSE in Section 2.3.

Changes in Lines 111-115: "The dataset was randomly divided into five parts, one was taken in turn as a test and the rest was used for training, which was repeated five times, and then the mean coefficient of determination (R^2) and the root mean square error (RMSE) were calculated (Wei et al., 2020).

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \hat{y})^2} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2} \quad (2)$$

"

6. L153: "an" – a.

- Thanks for the comment. We apologise for our carelessness, we have corrected the expression and carefully checked the entire manuscript for grammatical correctness.

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.

Hou, L. L., Dai, Q. L., Song, C. B., Liu, B. W., Guo, F. Z., Dai, T. J., Li, L. X., Liu, B. S., Bi, X. H., Zhang, Y. F., and Feng, Y. C.: Revealing Drivers of Haze Pollution by Explainable Machine Learning, Environmental Science & Technology Letters, 9, 112-119, 10.1021/acs.estlett.1c00865, 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.

Liu, S., Geng, G., Xiao, Q., Zheng, Y., Liu, X., Cheng, J., and Zhang, Q.: Tracking Daily Concentrations of PM2.5 Chemical Composition in China since 2000, Environ. Sci. Technol., 56, 16517-16527, 10.1021/acs.est.2c06510, 2022.

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.

Wei, J., Li, Z., Chen, X., Li, C., Sun, Y., Wang, J., Lyapustin, A., Brasseur, G. P., Jiang, M., Sun, L., Wang, T., Jung, C. H., Qiu, B., Fang, C., Liu, X., Hao, J., Wang, Y., Zhan, M., Song, X., and Liu, Y.: Separating Daily 1 km PM2.5 Inorganic Chemical Composition in China since 2000 via Deep Learning Integrating

Ground, Satellite, and Model Data, Environ. Sci. Technol., 57, 18282-18295, 10.1021/acs.est.3c00272, 2023.

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.