Review for manuscript: *Diagnosing drivers of PM2.5 simulation biases in China from meteorology, chemical composition, and emission sources using an efficient machine learning method*

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:

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

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

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.

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.

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.

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

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

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

L170. Remove the extra period before the citation.

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

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.