Reviewer 4

This manuscript presents an application of the LightGBM machine learning model to a multi-source dataset to quantify the respective contributions of meteorology and anthropogenic emissions to PM2.5 and PM10 variability in the BTH and YRD regions from 2015 to 2020. Accurately attributing these drivers is essential for formulating effective air-quality policies. However, there are several major concerns regarding the methodology, and the subsequent conclusions, which I believe need to be thoroughly addressed before the manuscript can be considered for publication.

Response: We thank this reviewer for his comments. We will respond to his comments point by point as shown below.

Major Comments：

1. The method used to separate meteorological and emission contributions (Section 2.6), which involves fixing one set of variables to a baseline year (2015) while allowing others to vary, is a central component of the analysis. While this approach could be used in physical based models (e.g., CTMs), its application to a purely data-driven model like LightGBM warrants further discussion. Machine learning models learn non-linear relationships that are specific to the co-varying patterns present in the training data. Creating scenarios with combinations of variables that have not been observed historically (e.g., 2020 meteorology with 2015 emissions) may represent an out-of-distribution task. However, the model evaluation is only based on the sample-based cross validation. The performance and the physical interpretability of its output under such conditions could be uncertain. Furthermore, the prediction is merely based on instantaneous states, excluding the cumulative effects of previous moments. The authors are encouraged to provide further justification for this method's suitability in an ML context, perhaps by citing literature where this technique has been validated for similar models or by conducting a sensitivity analysis to support the attribution results.

Response: Thank you for raising this important concern regarding the use of a machine-learning–based "fixed-emissions / varying-meteorology" counterfactual framework. We agree that such scenarios may combine predictor states that did not co-occur historically, and therefore require careful justification when applied in an ML context. This approach has been widely adopted in previous atmospheric studies that used machine-learning or statistical models to perform meteorological normalization or emission–meteorology attribution, particularly in China. Notably, Vu et al. (2019) employed a method conceptually consistent with our design: they used a machine-learning model to predict PM concentrations under modified meteorological conditions while holding emissions or other drivers fixed, thereby generating counterfactual pollutant levels for attribution purposes. This demonstrates that ML models can produce physically coherent responses under perturbed input scenarios and provides methodological support for the counterfactual strategy adopted in our study. In addition, our Figures 6, 7, and 9 show that the temporal evolution of key variables is strongly consistent with their SHAP contributions ($|R| \geq 0.89$), indicating that the model responds coherently to temporal changes in emissions and meteorological drivers. This empirical consistency provides further confidence that the model behaves stably when input conditions are altered in counterfactual experiments.

Overall, while we acknowledge that this method provides empirical attribution rather than strict causal inference, existing literature and our internal consistency checks support its suitability for trend attribution within a machine-learning framework.

References cited:

Vu, T. V., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., and Harrison, R. M.: Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique, Atmos. Chem. Phys., 19, 11303–11314, https://doi.org/10.5194/acp-19-11303-2019, 2019.

2. Several aspects of the machine learning implementation is confusing. The ML model includes PM10 (when predicting PM2.5) and vice versa among the feature set. Please discuss potential information leakage and quantify how much predictive skill derives from cross-pollutant auto-correlation versus direct meteorology/emission inputs. A sensitivity experiment retraining the model without inter-pollutant inputs would clarify the true drivers. The feature importance ranking presented in Fig. 3 indicates that the date is the most significant input variable. However, the methodology for constructing this feature is not detailed in either the manuscript or the supplement. Upon examining your code base at https://zenodo.org/records/16346573, I noted that treating temporal features, specifically using the date (e.g., YYYYMMDD) as a continuous numerical input, may present a significant methodological limitation. Tree-based models cannot inherently interpret the cyclical nature of time from a simple int/float representation. Consequently, this input feature could inadvertently act as an identifier for different days, potentially causing the model's predictions to over-rely on training data from the same day. Given the absence of a temporal-based split in the model evaluation, the current performance metrics might be overestimated.

Response: Thank you for raising this important concern regarding potential information leakage and the use of temporal identifiers in the machine-learning model. We fully agree that including cross-pollutant variables (e.g., PM_{10} when predicting $PM_{2.5}$) or using raw date encodings could bias the model and inflate its apparent skill. In the revised manuscript, we have substantially updated the modeling framework to directly address these issues. All pollutant concentrations ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3) and all explicit date-based numerical identifiers have been entirely removed from the input feature set. The model now relies exclusively on meteorological variables, aggregated emission variables, and cyclic temporal descriptors (month_sin, month_cos, season), which prevents both cross-pollutant leakage and unintended use of date identifiers as quasi-indices. To further ensure robust temporal generalization, we replaced the previous random 5-fold validation with a strict leave-one-year-out (LOGO) cross-validation scheme. Under LOGO, each full year is held out for testing while the model is trained on all remaining years, ensuring that no overlapping months appear in both training and test sets. The resulting predictive performance remains strong and is more conservative than in the earlier version. Because the revised feature space no longer contains any pollutant-concentration inputs or date-identifier variables, the specific sources of leakage highlighted in the review have been eliminated by design. Consequently, a separate sensitivity experiment (e.g., removing

cross-pollutant predictors) is no longer applicable within the updated model structure. The revised feature engineering and model training steps are now clearly described in the Methods section (see Lines 209–265).

3. The reported magnitudes for the PM concentration trends seem unexpectedly low. For instance, a reported annual decline of −0.07±0.03 μg m−3 for the BTH region appears to be several orders of magnitude smaller than what would be derived from the absolute concentration changes observed over the study period in public reports. The authors are kindly requested to verify these calculations and confirm the units. Additionally, it is suggested that the time series line charts, including the fitted lines calculated from Section 2.5, be presented in the supplement and clarify whether the trends in these regions remained stable throughout the 2015-2020 period, or if there were notable shifts at any point.

Response: Thank you for pointing out this issue. We carefully re-examined all trend calculations following your comment and confirmed that the previously reported annual rates were affected by an implementation error in the earlier version of the analysis. We have now fully corrected the trend computations for all cities and both regions.

The updated results show substantially larger and physically reasonable declines in $PM_{2.5}$ and PM_{10} concentrations over 2015–2022, fully consistent with both the observed time-series patterns and previously reported regional reductions in China. To improve transparency, we have added the complete monthly time-series plots together with their OLS trend lines as Fig. S3 in the Supplement. These figures clearly illustrate stable downward trajectories across both BTH and YRD.For clarity, we now explicitly state that:All trend estimates follow the regression framework described in Section 3.3Units have been verified (μg m^{-3} yr^{-1}), and the updated values align with the magnitude of observed multi-year declines. No abrupt structural shifts were detected within 2015–2022; the decreasing trends remain largely monotonic, with year-to-year variability superimposed on a consistent downward baseline.We appreciate the reviewer's careful attention to this point, which helped ensure the accuracy and robustness of the trend analysis.

Other Comments：

Line 131-134: Please specify the temporal and spatial resolution of the CEDS emissions dataset. Also, explain why only NO was chosen rather than the full NOx.

Response: Thank you for your comment. The temporal and spatial resolution of the CEDS emissions dataset has now been clearly stated in the revised manuscript: it provides monthly mean fluxes at a 0.5° × 0.5° global grid resolution.Regarding the second point, the "NO" field used in the earlier version was mislabeled. The variable corresponds to NO_x emissions, and we have corrected the terminology throughout the manuscript to reflect this. No methodological change is involved—the analysis has always been based on NO_x.

Line 144: Random Forest is not a gradient-boosting method; it belongs to the Bagging family.

Response: Thank you for pointing this out. Random Forest has now been removed from the section discussing gradient-boosting methods.

Line 158-162: Pearson's R measures linear association and is not sufficient alone for nonlinear models like LightGBM. The coefficient of determination (R2) would more appropriately assess the model's explanatory power in this context.

Response: Thank you for your comment. We agree that Pearson's *R* alone is not sufficient for evaluating a nonlinear model such as LightGBM. In the revised manuscript, we now report both *R* and R^2 to more appropriately characterize model performance.

Specifically, the model achieves R = 0.82, R^2 = 0.67 for $PM_{2.5}$ and R = 0.81, R^2 = 0.65 for PM_{10}. These additions ensure that the evaluation metrics fully reflect the model's explanatory capability.

By the way, the provided source code reveals a critical error in calculating R2. The code r2_value = r2_score(test, y) reverses the required (y_true, y_pred) argument order for the sklearn.metrics.r2_score function (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html). The formula for R2 $(1-SSres/SStot)$ is dependent on the total sum of squares of the true values. Reversing the arguments changes this denominator to the total sum of squares of the predicted values, which is mathematically incorrect and yields a metric that is not R2.

Response: Thanks for your comments. We have corrected this issue. The argument order in the r2_score function has been fixed to r2_score (y_true, y_pred) to ensure mathematically valid R^2 calculations. The revised manuscript now reflects the corrected values.

Line 268: The x-axis of Figure 2 needs a clear unit label.

Response: Thanks for your comments. The unit label on the x-axis of Figure 2 has now been added.