Reviewer 3

This manuscript builds a LightGBM framework that combines ground-monitor observations, reanalysis meteorology, and an emissions inventory (CEDS) to attribute 2015–2020 PM2.5/PM10 trends in the BTH and YRD regions to meteorology versus anthropogenic emissions. The topic is good. However, the current attribution design suffers from endogeneity, potential train–test leakage in validation, and limited uncertainty quantification; in addition, key claims rely on variable-importance metrics and trend magnitudes that need correction/clarification. I recommend major revision.

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

Main comments

1.Attribution mix-up (endogeneity). Authors keep co-pollutants (CO, NO_2, SO_2, PM) as predictors while only "freezing" emissions. Those pollutant levels already reflect emissions, so they leak emission info into the "meteorology-only" case and bias the split.

Response: We sincerely thank the reviewer for this insightful comment. In response to this concern, we have substantially revised our machine learning model in the updated version of the manuscript. Specifically, we have now exclusively included emissions, meteorological factors, and temporal descriptors as input variables, while removing the concentrations of $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, and CO from the model. This adjustment ensures a clearer and more appropriate attribution of contributions from emissions and meteorology to the changes in PM levels. We are pleased to note that the revised model still demonstrates strong performance, effectively capturing the variations in PM concentrations without the inclusion of other pollutant variables. This modification also aligns more logically with our study's objective of separating emission-driven and meteorology-driven influences. The relevant sections of the manuscript, including the Methods (See Lines 180–264) and Results (See Lines 348–377), have been updated accordingly to reflect these changes.

2.Cross-validation leakage. Authors mentioned 'a 5-fold cross-validation framework was implemented: the full training dataset was randomly partitioned into five mutually exclusive subsets' in this study. Random k-fold lets nearby days and the same cities appear in both train and test, inflating scores. That is not enough to check and avoid leakage for this study. I suggest that use blocked CV: leave-one-year/season out; leave-one-city out; ideally both. Report R, RMSE/MAE, and bias for each scheme.

Response: Thank you for your thoughtful comment. We fully agree that random k-fold cross-validation can introduce temporal leakage, especially when adjacent months share similar meteorological or emission conditions. To address this issue, we have revised the model framework and now adopt a leave-one-year-out (LOGO) cross-validation approach as the sole validation scheme in the manuscript. Under this strategy, each full year from 2015–2022 is held out as the test set while all remaining years are used for training. This design ensures strict temporal separation between training and testing, eliminates leakage from adjacent months or seasons, and evaluates the model under genuinely unseen meteorological and emission conditions. Because a separate model is trained independently

for each city, there is also no possibility of cross-city leakage. Model performance metrics (R, RMSE, MAE, and bias) under the LOGO scheme remain highly consistent with those reported previously, indicating that the model's predictive skill is not inflated by temporal dependence. The updated validation results have been incorporated into the revised manuscript (see Lines 349–354), and the detailed model training procedure is now described in Lines 242–255. This revision fully resolves the concern regarding cross-validation leakage and provides a more conservative and robust assessment of model generalization.

3."Importance doesn't mean variance explained." The analysis of variable importance is not explained well and clearly in this study. Tree importance (gain/splits) isn't "% of variation explained.". Authors should use SHAP or permutation importance and show partial-dependence (or ALE) plots. Reword claims to avoid "explains X% of variation."

Response: Thanks for this important comment. We agree that tree-based importance (gain/split counts) cannot be interpreted as "percentage of variance explained," and that our original wording was potentially misleading. In the revised manuscript, we have removed all expressions such as "explains X% of variation" and now base the attribution analysis entirely on SHAP values. As described in the Methods (Lines 266–281), for each model the SHAP value $s_{i,j}$ is used to represent the marginal contribution of feature j to the prediction for sample i, and feature importance is quantified by the (sample-size–weighted) mean absolute SHAP magnitude across cities, which we interpret only as a relative contribution strength within the model. In addition, rather than relying on tree-based importance, we now use SHAP-derived temporal and dependence patterns (Figs. 6, 7 and 9) to illustrate how key predictors relate to $PM_{2.5}$ and PM_{10}, which serves a similar role to partial dependence/ALE plots while remaining fully consistent with the SHAP framework.

4.Trend numbers/units look off. Very small annual rates don't match the multi-year drops shown. Authors should recheck units and decimals. Report both absolute (μg m^{-3} yr^{-1}) and relative (% yr^{-1}) trends with uncertainty, using a consistent method.

Response: We thank the reviewer for pointing out the inconsistency in our original trend values. We have re-calculated the interannual trends using ordinary least squares and now report both absolute (μg m^{-3} yr^{-1}) and relative (% yr^{-1}) trends together with their uncertainties. The uncertainties are quantified as 95% confidence intervals derived from the standard errors of the regression slopes. The updated results are provided in Table S3.

5.Inventory selection. The reason for choosing CEDS as inventory is not clear. CEDS is a global inventory, and the city grids for China are not representative enough. I recommend to compare it with the MEIC inventory, which is a China-specific inventory.

Response:

Thank you for this helpful comment. We agree that MEIC is a China-specific inventory with finer spatial detail, and we appreciate the need to justify the use of CEDS in our analysis. Our study requires a long, continuous, and fully sector-resolved monthly emission time series from 2015–2022, and CEDS is currently the only dataset that provides complete temporal coverage for all required species over this full period. In contrast, the publicly

available MEIC inventory does not provide emissions for the most recent years (2020–2022), and therefore cannot support the full temporal window analyzed in this study. Previous work has shown that CEDS and MEIC exhibit highly consistent interannual emission trends over China. For example, *Comparison of emissions inventories of anthropogenic air pollutants and greenhouse gases in China* (Saikawa et al., 2017), *A global anthropogenic emission inventory of atmospheric pollutants from sector- and fuel-specific sources (1970–2017): an application of the Community Emissions Data System (CEDS)* (McDuffie et al., 2020), and *An integrated view of correlated emissions of greenhouse gases and air pollutants in China* (Lin et al., 2023) all report broadly similar year-to-year trajectories across multiple inventories, including MEIC and CEDS, for key air pollutants and greenhouse gases in China. These results indicate that CEDS reliably captures the temporal evolution of Chinese anthropogenic emissions at regional scales. Given that our machine-learning attribution focuses on interannual variability and relative temporal changes rather than absolute magnitudes, using CEDS as a consistent, long-term emission dataset is therefore appropriate for this study.

References:

Saikawa, E., Kim, H., Zhong, M., Avramov, A., Zhao, Y., Janssens-Maenhout, G., Kurokawa, J.-i., Klimont, Z., Wagner, F., Naik, V., Horowitz, L. W., & Zhang, Q. (2017). Comparison of emissions inventories of anthropogenic air pollutants and greenhouse gases in China. Atmospheric Chemistry and Physics, 17(10), 6393–6421. https://doi.org/10.5194/acp-17-6393-2017

McDuffie, E. E., Smith, S. J., O'Rourke, P., Tibrewal, K., Venkataraman, C., Marais, E. A., Zheng, B., Crippa, M., Brauer, M., & Martin, R. V. (2020). A global anthropogenic emission inventory of atmospheric pollutants from sector- and fuel-specific sources (1970–2017): An application of the Community Emissions Data System (CEDS). Earth System Science Data, 12(4), 3413–3442. https://doi.org/10.5194/essd-12-3413-2020

Lin, X., Yang, R., Zhang, W., Zeng, N., Zhao, Y., Wang, G., Li, T., & Cai, Q. (2023). An integrated view of correlated emissions of greenhouse gases and air pollutants in China. Carbon Balance and Management, 18(1), 9. https://doi.org/10.1186/s13021-023-00229-x

6.Tone down causal claims. Linking the 2019–2020 drop mainly to policy may overstate causality, especially with COVID shocks. Authors should add a check excluding the year 2020.

Response: Thank you for raising this important point. We fully agree that attributing the 2019–2020 changes primarily to policy actions may overstate causality, especially given potential COVID-related perturbations. In the revised manuscript, we carefully re-examined the interannual and monthly time series trends of $PM_{2.5}$ and PM_{10} in both BTH and YRD (see Fig. S3). These updated sequences show no anomalous or disproportionate drop in 2020 relative to adjacent years. Instead: (1) The largest decreases occur during 2015–2017, (2) The decline slows after 2018, (2) And 2020 does not exhibit an isolated or unusually sharp reduction in either region. This pattern indicates that the multi-year downward trend is gradual and continuous, not dominated by the COVID-19 period.

Therefore, the revised text now avoids any causal claims linking the 2019–2020 changes to policy alone. The manuscript explicitly focuses on long-term structural emission reductions as the dominant driver, consistent with the monotonic trend observed across the entire 2015–2022 period, rather than any single-year factor. Corresponding statements have been softened or removed in the revision.

7.Clarify the contribution. The time coverage is outdated. Specify what is new (e.g., data, model, scale, or attribution design) versus prior studies, cite those studies, and show how your results change or add value.

Response: We thank the reviewer for raising this important point regarding the novelty of our work and for providing the list of relevant literature. We agree that several excellent studies have explored the separation of emission and meteorological contributions to air pollution in China. However, our manuscript provides significant advancements and novel insights in the following key aspects, which distinguish it from the existing body of work: First, extended and more dataset: Our analysis extends the investigation to the period 2015–2022. This more recent timeframe captures the crucial later phases of China's Air Pollution Prevention and Control Action Plan, as well as the unique emission variations associated with the COVID-19 pandemic and subsequent economic recovery, which are not covered by the cited studies (which end in 2020 or earlier). Second, comprehensive Analysis of Both $PM_{2.5}$ and PM_{10}: The studies listed by the reviewer primarily focus on either $PM_{2.5}$ or PM_{10}. A key novelty of our work is the simultaneous and comparative analysis of both particulate matter species within a unified methodological framework. This allows for a direct investigation of the differing drivers and behaviors of fine and coarse particles, providing a more holistic understanding of particulate air pollution. Third, Application of SHAP for long-term attribution and physical consistency evaluation. Although machine-learning models have been used previously, the application of SHAP-based feature attribution to multi-year PM variability is still very limited. Our study extends SHAP usage from short-term prediction contexts to long-term trend attribution, allowing transparent quantification of how key predictors contribute to interannual PM changes. Importantly, we also demonstrate that emission features and their SHAP attributions exhibit strong temporal consistency ($R \approx 0.89–0.95$) across species such as SO_2, NOx, and BC, providing a physically interpretable connection between emission evolution and model-inferred contributions. This strengthens confidence in the mechanistic fidelity of ML-derived conclusions. Fourth, Cross-validated, out-of-sample year-by-year reconstruction. Unlike many studies that train and evaluate models within the same period, our leave-one-year-out design produces true out-of-sample predictions for each individual year, enabling more credible reconstruction of meteorology-only and emission-only scenarios for trend decomposition.

Overall, rather than proposing a completely new attribution paradigm, our contribution lies in extending the observational period, integrating $PM_{2.5}$ – PM_{10} analyses, and introducing SHAP-based long-term mechanistic attribution with strict temporal cross-validation. These elements together provide new quantitative evidence on how anthropogenic precursors and meteorology shaped PM evolution during the most recent decade.

Minor comments

1. Clarify feature groups (meteorology vs. emissions/activity vs. concentrations) and which ones go into each counterfactual. State any lags.

Response: Thank you for this helpful comment. In the revised manuscript, we now explicitly clarify how the predictor groups are defined and how they are used in each counterfactual experiment. The LightGBM model in this study uses only meteorological variables, anthropogenic-emission variables, and temporal descriptors as predictors; all pollutant concentrations are excluded to avoid leakage. The complete list of input features is provided in Table S2. Meteorological features reflect real atmospheric conditions, while emission features—including species-level totals and their derived indicators (sdiff and detr)—represent changes in anthropogenic activities. Temporal descriptors (month_sin, month_cos, season) encode annual periodicity but contain no pollution information. For the counterfactual experiment used to separate meteorological and emission contributions, only the emission-related features (those listed as emission variables in Table S2) are held fixed at their 2015 values, while meteorological and temporal predictors retain their actual values for each target year. This design ensures that the counterfactual series reflects meteorology-driven variability alone. No lagged features are included in the model. This clarification has been incorporated into the revised manuscript (See in Lines208-240).

2. List LightGBM hyperparameters, seeds, data splits and more details.

Response: Thank you for your comment. Because each city–pollutant pair is trained independently under our Leave-One-Year-Out cross-validation scheme, the LightGBM hyperparameters vary across models, and listing all fold-specific configurations in the main text would be impractically long. For transparency and reproducibility, we summarize the full ranges of optimized hyperparameters in Supplementary Table S3, which provides a compact and comprehensive overview of all parameter settings used in this study.

3. Check variable names/units (e.g., T2M is near-surface air temperature, not "maximum"). It is confused that this paper shows: 'Line 24: 2-m temperature (T2M)'. Line 127: '2-m maximum air temperature (T2M)'.

Response: Thank you for pointing out this inconsistency. We have corrected the description of T2M throughout the manuscript to consistently refer to it as 2-m air temperature, and the erroneous phrase "2-m maximum air temperature (T2M)" has been removed (See lines 141).

4. For city averages, give station counts, completeness rules, and weighting (simple mean vs. population/land-use weights).

Response: Thank you for your comment. In this study, only meteorological and emission variables require spatial aggregation, and both are derived from gridded datasets following consistent city-level extraction procedures. For meteorological variables, which are intensive state quantities (e.g., T2M, QV2M, U10M, V10M), city-level values are obtained as the simple arithmetic mean of all GEOS-FP grid-cell centers falling within each city polygon. Because the study domain is located in mid–low latitudes, the variation in grid-cell area is relatively small, and this center-based averaging introduces negligible bias. The detailed extraction description is provided in the manuscript (see Lines 167–179). For

emission variables, which are expressed as surface fluxes (kg m^{-2} s^{-1}), city-level emissions are calculated using an area-weighted integration across all grid cells overlapping each city boundary. This ensures consistency between flux-based emissions and city-scale totals. The full formulation and computational steps are described in the manuscript (see Lines 155–166).These two procedures—simple averaging for meteorology and area-weighted integration for emissions—ensure spatially consistent, physically meaningful city-level inputs for subsequent model training.

5. Figure3 and relevant description: explain and show which "importance" metric you use; add SHAP/PD plots.

Response: Thank you for this helpful suggestion. In the revised manuscript, we have clarified the definition and computation of the feature-importance metric. As detailed in the Methods section (Lines 266–281), all importance values used in this study are based on SHAP contributions, computed from the Shapley additive framework rather than tree-based gain/split metrics. Figure 3 has been updated to explicitly represent the SHAP-derived feature importance ranking, and the figure caption and text now clearly state that SHAP values are the importance metric used.

In summary, this study requires a more defensible attribution design, leakage-safe validation, and stronger uncertainty treatment before the conclusions can be considered robust.

Response: Thanks for your comments. We hope our responses above have addressed your concerns.