Response to reviewer's comment on "Meteorological influence on surface ozone trends in China: Assessing uncertainties caused by multi-dataset and multi-method" by X. Wang et al. (Ms. Ref. No.: EGUSPHERE-2025-1880)

Response to Referee #1

This study presents an analysis of the meteorological drivers of surface ozone (O_3) trends in China from 2013 to 2022, based on an observational dataset of ozone and various supporting analyses, including statistical analyses, simple machine-learning and chemical transport modeling using GEOS-Chem.

The authors highlight the role of meteorological conditions in driving seasonal and regional ozone increases, and use these analyses to begin a discussion of the uncertainties arising from applying these different supporting datasets. The paper will be of interest for those using such large-scale observational datasets to isolate the drivers of air quality trends and may be of interest to policymakers. The use of a consistent metric is interesting. The paper represents a significant effort in gathering and providing an interesting high-level analysis of different ways to analyse the data.

Response:

We sincerely thank the referee for the decision and constructive comments. The manuscript has been revised accordingly, and our point-by-point responses are provided below. The referee's comments are shown in black, and our replies are highlighted in blue. A tracked-changes version of the revised manuscript is also clearly showing the changes made.

The main result of the study is to assess the consistency between approaches using a coefficient of variability metric, in which higher CVs indicate lower consistency of meteorologically driven O_3 trends derived from different datasets or methods. Initially this is used as a comparator between datasets, but towards the end of the MS the authors use this more quantitatively, with thresholds of 0.5 and 1.0 being applied to indicate consistency. How were these numbers chosen? What do they mean?

Response:

The CV eliminates the influence of dimensionality across different models by normalizing the standard deviation with the mean, thereby enabling meaningful comparisons of the dispersion degrees among various models. Mathematically, a "CV < 1" indicates that results from different models are closely clustered around the mean, signifying a high consistency across the models. Both Chen et al. (2019) and Wang et

al. (2022) adopted a threshold of 1.0 for classifying variability levels. In this study, to more rigorously assess the uncertainties caused by multi-dataset and multi-method, we introduced an additional stringent threshold of 0.5. This refinement allows for a more nuanced demonstration of the consistency in multimodal results.

We have added the description on thresholds of 0.5 and 1.0 as follows: "To give a more quantitative assessment, consistency levels were classified as strong and weak with CV<0.5 and CV>1.0, respectively (Wang et al., 2022a)." *[Lines 228–229 in the tracked-changes version of the revised manuscript]*

What other metrics could be used as a metric for comparison?

Response:

Besides CV, other widely used measures of spread include the range, inter-quartile range (IQR), and standard deviation (SD) (Chattamvelli and Shanmugam, 2023). The range and IQR, calculated as the difference between the maximum and minimum values, and the difference between the 75%-quartile and 25%-quartile, respectively, can be sensitive to outliers and heavy-tailed distributions (Högel et al., 1994). In contrast, SD quantifies total variation around the mean and is less responsive to tail behaviour. Compared to SD, the CV is a unit-free measure that expresses variation relative to the mean as a percentage.

We have added the following statement to clarify the advantages of CV over other metrics: "Compared to other comparators (e.g. range, inter-quartile range, and SD), the CV is a unit-free measure that quantifies percentage variation relative to the mean and is less sensitive to outliers and heavy-tailed distributions (Högel et al., 1994; Chattamvelli and Shanmugam, 2023)." *[Lines 225–227 in the tracked-changes version of the revised manuscript]*

Most time is spent discussing an analysis using meteorological reanalyses with the ML and CTM work in a supporting role as challenger methods to the MLR analysis.
In section 2, the methods used are described. In the regression-based statistical analysis, the authors first use a time-series filter to retrieve trends in ozone and other fields, and then a MLR-based model to derive the drivers of these trends. I was not able to find further details of the method used as it is in a separate publication that is incorrectly referenced.

Response:

Following the Reviewer's suggestion, we have refined the description of the KZ-

MLR methodology as follows: "After establishing MLR models for the short-term and baseline components in each season, we obtain their respective residual terms. The total residuals, which represent the sum of residuals from baseline variables and short-term variables, primarily reflect anthropogenic influences. We then applied a $KZ_{(365,3)}$ filter to these aggregated residuals to derive long-term emission-driven and meteorology-driven O_3 variations. Finally, the meteorology-driven O_3 trends and emission-driven O_3 trends were obtained through Least Square Method." *[Lines 146–150 in the tracked-changes version of the revised manuscript]*

Additionally, we have revised the caption of Figure S1 to more clearly illustrate the KZ-MLR workflow.

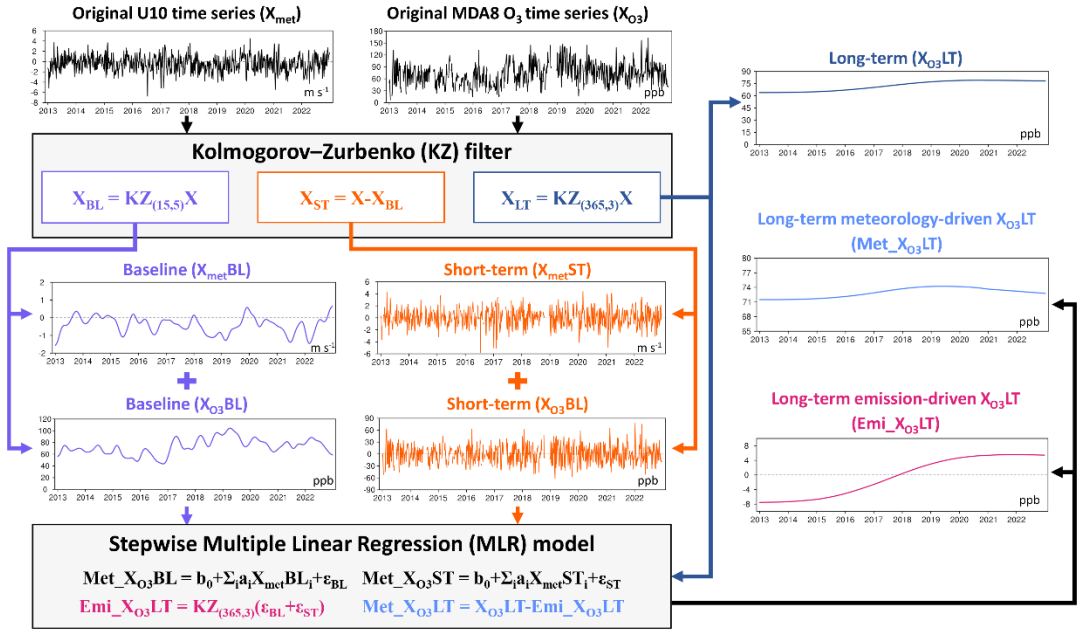

Figure S1. Flowchart of the Kolmogorov-Zurbenko - Multiple linear regression (KZ-MLR) model, which decomposes the observed MDA8 O_3 time series into meteorology-driven and emission-driven long-term components. Shown on the figure is an example: MDA8 O_3 data from Station_1015A and U10 data from ERA5 during the summer.

The ML study is perhaps the least well justified - six of the predictors are proxies for time, with a further six (pressure, temp, wind speed, RH and PBLH) being deemed sufficient to capture the meteorological drivers of ozone. I have reservations about this approach because the RF model is trained on MDA8 O_3 concentrations. Are the authors satisfied that this model is sufficiently accurate that it can be used for attribution of driver and yield confident results? If so, what is the justification? What is the basis for explaining 50% of the variance to be a threshold for inclusion? I'd like to see more here,

particularly the basis for exclusion of e.g. trends in emissions or atmospheric composition which may be drivers. It would seem much more appropriate if they had used RF to predict the recovered LT O_3 trend and then used the meteorological data as predictors for the trend.

Response:

Thank you for thoughtful comments regarding ML study. We would like to address the concern as follows:

1) Given the demonstrably limited impact of algorithmic differences in O_3-Meteorology analyses (Wang et al., 2024a), we employ a RF-based weather normalisation framework, which is a validated method for quantifying meteorological influences on O_3 concentrations (Grange et al., 2018; Vu et al., 2019). Our predictor set integrates six temporal proxies (capturing long-term anthropogenic drivers including emission trends and atmospheric composition changes) with six key meteorological parameters (pressure, temperature, U10, V10, relative humidity, and PBLH) that demonstrably influence O_3 variability.
2) We maintain high confidence in the model's attribution capability, as it skillfully reconstructs observed MDA8 O_3 concentrations (Fig. S2b), achieving $R^2 > 0.5$ at over 70% of state-controlled stations in all seasons, which is consistent with the 0.4–0.6 range reported in comparable studies (Weng et al., 2022; Lu et al., 2024). The $R^2 > 0.5$ inclusion threshold represents a deliberate compromise between model reliability and spatial coverage, systematically excluding stations where RF performance could introduce significant attribution uncertainty (Varoquaux and Colliot, 2023).
3) For meteorological normalisation, we implement the protocol of Vu et al. (2019): daily meteorological variables undergo 1000 resampling iterations from ±14-day observational windows while temporal proxies remain fixed. The mean predicted O_3 under average meteorological conditions refers to the emission-driven concentration. The residual (the difference between observation and emission-driven O_3) constitutes the meteorology-driven component, with its long-term trends derived through $KZ_{(365,3)}$ filter followed by Least Square Method. This integrated methodology robustly separates meteorological and anthropogenic drivers.
4) We have added the statement on R^2 as follows: "Over 70% of state-controlled stations showed $R^2 > 0.5$ in all seasons (Fig. S2b), which is consistent with the 0.4–0.6 range reported in comparable studies (Weng et al., 2022; Lu et al., 2024).

Stations with $R^2 < 0.5$ were excluded to avoid significant attribution uncertainty that could be introduced by the RF performance." *[Lines 171–173 in the tracked-changes version of the revised manuscript]*

5) More details about the RF model were also added as follows: "For meteorological normalisation, we implemented the protocol of Vu et al. (2019). Meteorological variables were resampled by randomly selecting data from the two weeks before and after the specified date, while temporal proxies remained fixed. To derive the de-weathered MDA8 O_3 concentration for a given day (e.g. March 1, 2013), the random resampling process was iterated 1000 times. The mean predicted O_3 under average meteorological conditions, which refers to de-weathered O_3, corresponds to the emission-driven O_3 concentration. The meteorology-driven MDA8 O_3 concentrations for each season were computed as the difference between observed concentrations and de-weathered concentrations. Detailed processes were shown in Fig. S2(a). The $KZ_{(365,3)}$ filter was then applied to obtain long-term components, and meteorology-driven O_3 trends were derived using Least Square Method." *[Lines 179–187 in the tracked-changes version of the revised manuscript]*

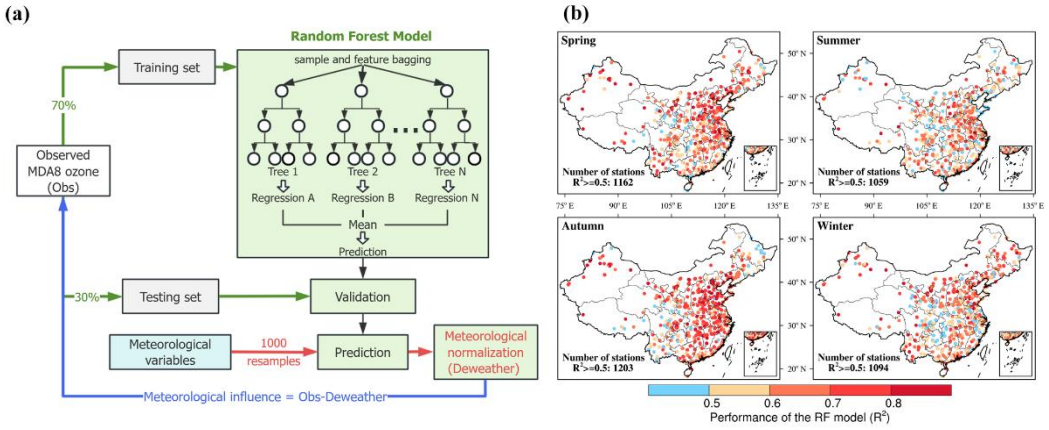

Figure S2. (a) Conceptual diagram of obtaining the meteorological influence based on the Random Forest (RF) algorithm, and (b) the performance of the RF model for the testing dataset at each state-controlled station during four seasons. The number of state-controlled monitoring stations with the coefficient of determination (R^2) greater than or equal to 0.5 is also presented.

L167 specifies how MDA8 was calculated, but needs much more detail on how the trends were computed.

Response:

Trends in this study were all calculated using the Least Square Method. We have added the corresponding statement in Sections 2.2.2, 2.2.3, and 2.2.4. *[Lines 150, 187,*

215 in the tracked-changes version of the revised manuscript]

The use of GEOS-Chem is interesting and the experiment is well-conceived, and the model is well validated in the supporting information. No information on the extraction of the trend data from the GC experiments is given, and this should be included in the main MS.

Response:

Following the Reviewer's suggestion, we have added the information on the extraction of the trend data from the GC experiments as follows: "The FixE2013 simulation is designed to obtain the MDA8 O_3 concentrations driven solely by meteorological changes and further quantify the meteorological influence on O_3 variations in four seasons. After applying the $KZ_{(365,3)}$ filter to derive the long-term meteorology-driven series, trends were calculated through Least Square Method." *[Lines 213–216 in the tracked-changes version of the revised manuscript]*

The MERRA2 reanalysis was used to drive the CTM. Given the scope of the MS, why just one reanalysis? It seems that there's an opportunity here to expand the analysis of the uncertainty in the GC trend on meteorological product, and it is certainly necessary to discuss how the lack of independence of the GC and MLR (MERRA2) results affects the analysis in this paper.

Response:

Thank you for the suggestion. We would like to address the concern as follows:

1) To our knowledge, GEOS-Chem is conventionally driven by meteorological inputs from NASA's Global Modeling and Assimilation Office (GMAO) Goddard Earth Observing System (GEOS), such as MERRA2. MERRA2 currently represents the latest and most widely adopted NASA/GMAO reanalysis product. Consequently, utilizing MERRA2 to drive GEOS-Chem aligns with established methodological practices in the field. Employing alternative meteorological fields (e.g. ERA5 or FNL) would necessitate converting these datasets into formats compatible with GEOS-Chem, which is a process requiring substantial time investment and posing technical challenges.
2) Furthermore, a key objective of this study is assessing uncertainties in quantifying meteorological impacts arising from the use of different datasets. This objective is addressed through MLR models driven by MERRA2, ERA5, and FNL. Thus, additional simulations using alternative reanalyses to drive GEOS-Chem were deemed beyond the scope of this work.

3) While prior studies (e.g. using WRF-CMAQ) have explored reanalysis-dependent variability in the CTM (Wang et al., 2024b), such investigations remain limited for GEOS-Chem.
4) We fully acknowledge the value of your suggestion and intend to pursue this avenue in future research by systematically evaluating different reanalysis products within the GEOS-Chem framework.

Section 3.1 details the results, and leaves some questions unanswered. Please include a discussion of what the analysis says about which are the main drivers, etc. At present, this discussion is more of a comparison with other findings. In fact, the authors note that most of the outcomes are already published elsewhere (L214-L223), which reinforces the need for novel analysis in this section. I believe the MS would be improved by reporting drivers of the trends, particularly as Section 3.2 lumps all these drivers together as the meteorological impact on the MDA8 O_3 trends. Maybe a figure showing the contribution of each driver would be useful here.

Response:

We appreciate your constructive comments regarding driver analysis in Section 3.1. As established in Section 1, O_3 variations are primarily modulated by emissions and meteorology. Our manuscript therefore employs a structured analytical progression: Section 3.1 characterizes observed O_3 trends in China through comparison with previous studies, establishing the essential context for subsequent driver attribution. This foundational approach intentionally precedes Sections 3.2–3.3, where we assess the uncertainty in meteorology-driven O_3 trends caused by multi-dataset and multi-method, and further conduct driver quantification.

The current framework ensures our core focus—systematic uncertainty analysis in meteorology-driven O_3 trends, which can remain central while maintaining a clear separation between observational benchmarking and driver attribution. Introducing attribution results prematurely in Section 3.1 would disrupt this logical flow and create redundancy, particularly since comprehensive driver contributions were already visualized in Figure 5 (Section 3.2).

To strengthen narrative coherence and enhance analytical continuity, we have added the following transitional statement in Section 3.1: "As mentioned in Section 1, variations in O_3 concentrations are fundamentally modulated by emissions and meteorology. This section mainly documents observed O_3 trends, and the quantitative contributions of emissions and meteorology to MDA8 O_3 variations will be discussed

in Section 3.2." *[Lines 249–251 in the tracked-changes version of the revised manuscript]*

Section 3.2 addresses the consistency of the MLR results across different reanalyses. I don't understand why the uncertainty in the derived trends is not included here. Could it not be calculated? I suggest it's included, not least to visually assess the consistency/difference between calculated trends and support the CV analysis. If it can be calculated, please add it as an error bar to the figure.

Response:

Thanks! Following the Reviewer's suggestion, we have now incorporated error bars into Figure 3 and Figure 6 to address the uncertainty in the derived trends. Corresponding descriptions of the error bar methodology is detailed in the updated figure captions: "Error bars indicate ±1 standard deviation (SD) of site-level trends calculated from all available monitoring stations within each region."

We have added the following description about error bars:

"with the multi-dataset mean trends ranging from +0.19 (±0.47) ppb yr^{-1} to +0.55 (±0.45) ppb yr^{-1}." *[Line 264 in the tracked-changes version of the revised manuscript]*

"with trends ranging from +0.47 (±0.47) ppb yr^{-1} to +0.71 (±0.59) ppb yr^{-1}"*[Line 265 in the tracked-changes version of the revised manuscript]*

"During summer and autumn, meteorological influences on O_3 show the greater spatial heterogeneity (with higher SD) and larger variability among multi-datasets (with higher CV)." *[Lines 267–268 in the tracked-changes version of the revised manuscript]*

"the multi-dataset mean trends ranged from +0.09 (±0.38) ppb yr^{-1} to +0.33 (±0.13) ppb yr^{-1} in BTH, +0.18 (±0.20) ppb yr^{-1} to +0.68 (±0.56) ppb yr^{-1} in YRD, and +0.73 (±0.36) ppb yr^{-1} to +1.13 (±0.45) ppb yr^{-1} in PRD" *[Lines 278–279 in the tracked-changes version of the revised manuscript]*

"In the other three seasons, the multi-method mean trends, ranging from +0.17 (±0.37) ppb yr^{-1} to +0.26 (±0.27) ppb yr^{-1}, are 1.1 to 2.1 times lower than those computed by the three dataset-driven MLR models (Fig. 3a)"*[Lines 339–341 in the tracked-changes version of the revised manuscript]*

"multi-method mean trends of +0.17 (±0.08) to +0.47 (±0.22) ppb yr^{-1} and +0.10 (±0.12) to +0.83 (±0.19) ppb yr^{-1}" *[Lines 351–352 in the tracked-changes version of the revised manuscript]*

"In BTH, the three models perform consistently well only in winter, with meteorology-driven O_3 trends ranging from +0.09 (±0.07) ppb yr^{-1} to +0.26 (±0.15) ppb

yr^{-1} and a CV of 0.55." *[Lines 355–357 in the tracked-changes version of the revised manuscript]*

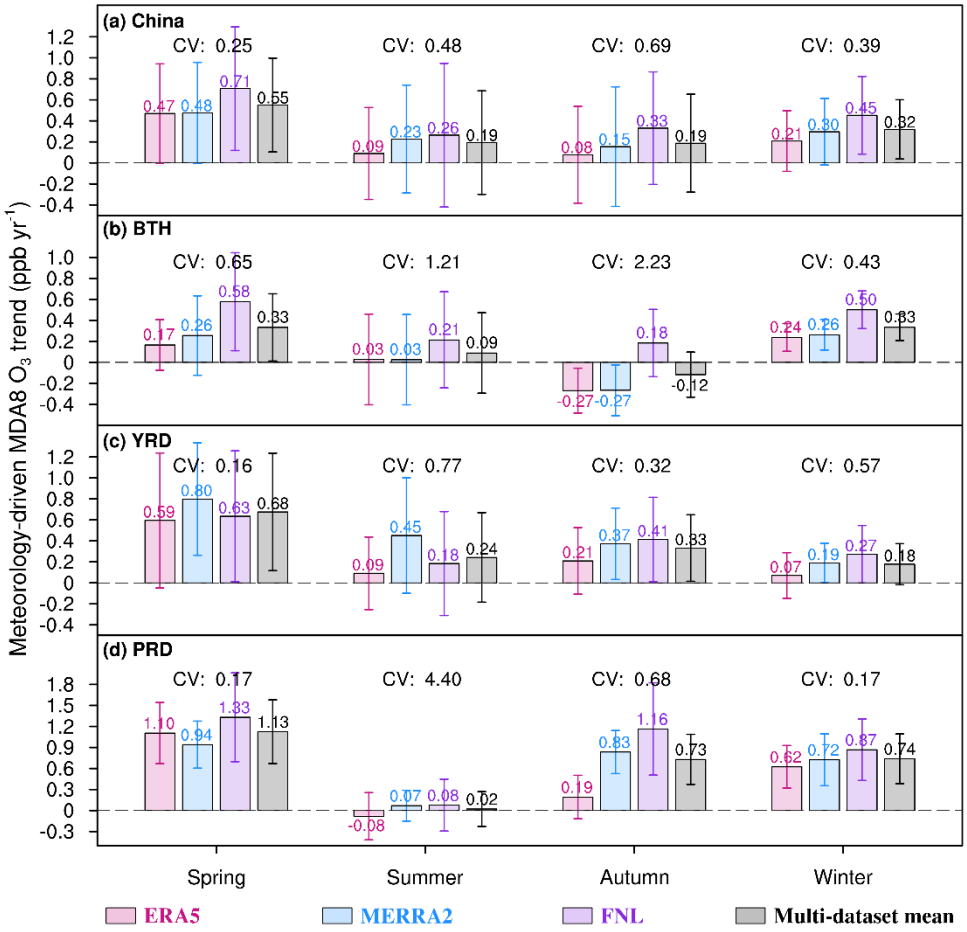

Figure 3. Meteorology-driven MDA8 O_3 trends in (a) the whole China, (b) BTH, (c) YRD, and (d) PRD during four seasons. Values in red, blue, and purple represent trends calculated by ERA5-, MERRA2-, and FNL-driven multiple linear regression (MLR) model, respectively. The fourth black bar represents the multi-dataset mean trend. Error bars indicate ±1 standard deviation (SD) of site-level trends calculated from all available monitoring stations within each region. The absolute value of the coefficient of variation (CV) for each season is also shown.

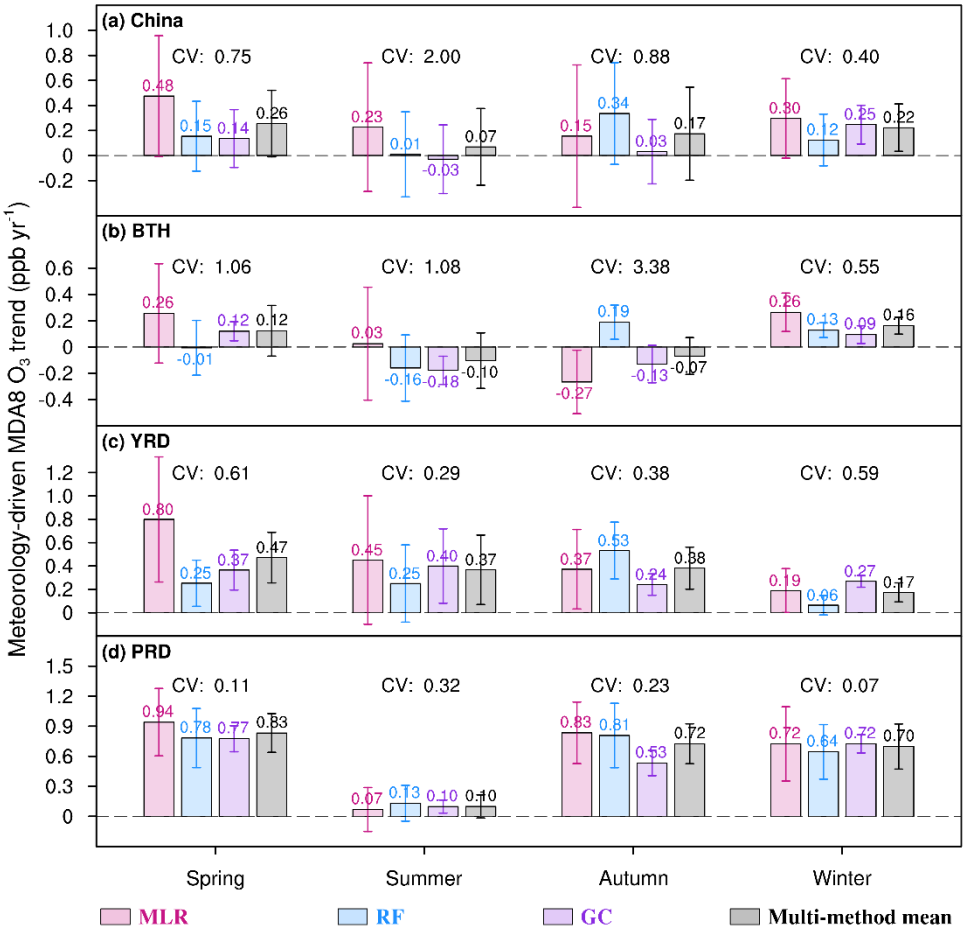

Figure 6. Meteorology-driven MDA8 O_3 trends in (a) the whole China, (b) BTH, (c) YRD, and (d) PRD during four seasons. Values in red, blue, and purple represent trends calculated by multiple linear regression (MLR), random forest (RF), and GEOS-Chem (GC) models, respectively. The fourth black bar represents the multi-method mean trend. Error bars indicate ±1 standard deviation (SD) of site-level trends calculated from all available monitoring stations within each region. The absolute value of the coefficient of variation (CV) for each season is also shown.

Section 3.3 confronts the MLR method with its challengers. Here the MS inter-compares the metrics and notes the difference across various domains. This provides a brief description of the uncertainty (ie spread) of results but stops short of providing a good assessment of the importance of individual drivers or in making broad recommendations as to which analysis is the most robust, reliable or useful. The analysis of the FNL results is interesting. My main concern here is with the ML/RF approach: it may be undermined by relatively low skill of the resulting model and resulting first-principles questions as to the robustness of these results - does a statistical model of relatively low skill permit us to say much about the drivers?

Response:

Thank you for your insightful comments regarding the multi-method analyses in Section 3.3. We would like to address the concern as follows:

1) Following the Reviewer's suggestion, we have made broad recommendations as follows: "To obtain a more reliable estimate, it is recommended to use MERRA2 reanalysis dataset due to its eclectic result (Fig. 3) and avoid using FNL because of the uncertainty brought by PBLH when separating meteorological and anthropogenic influences on O_3 concentrations in China." *[Lines 332–334 in the tracked-changes version of the revised manuscript]* and "The trends driven by RF model are eclectic in more cases (Fig. 6) and recommended to isolate meteorological and anthropogenic drivers. " *[Lines 375–376 in the tracked-changes version of the revised manuscript]*
2) The RF model's robustness was rigorously established through both model performance assessment (over 70% of state-controlled stations exhibited $R^2 > 0.5$ across all seasons) and consistency with previous studies on RF-based separation of meteorological influences. We are therefore confident that the insights derived from the RF model provide a meaningful foundation for evaluating meteorological influences on O_3 concentrations. See more details in our reply to the fourth comment above.
3) While our current focus is quantifying uncertainties in meteorology-driven O_3 trends caused by multi-method, we agree that deeper interrogation of individual drivers (e.g. temperature, wind speed, relative humidity) is essential. Future work will employ Lindeman-Merenda-Gold (LMG) indices to quantitatively resolve the contributions of specific meteorological variables, thereby strengthening mechanistic interpretations of O_3 variations.
4) Following the Reviewer's suggestion, we have expanded the limitation discussion in Section 4 as follows: "Finally, the Lindeman-Merenda-Gold indices can be employed to quantitatively resolve the contributions of specific meteorological variables. The mechanistic understanding of O_3 drivers would be improved by integrating additional variables, such as solar radiation, soil moisture, and climate indices (e.g. El Niño-Southern Oscillation)."*[Lines 401–403 in the tracked-changes version of the revised manuscript]*

Overall, the MS has a number of positive qualities: the use multi-dataset and multi-method approaches is welcome. The MS shows that the analysis is quite robust for some

regions and some seasons, and so has some policy relevance.

Response:

Thank you again for your positive comments on our manuscript.

The MS would be much improved if the analysis was extended to identify the drivers of the what the authors call uncertainty, ie intermodel spread. At present, the MS doesn't give enough information on how the ML and GC data were used to compute trends, and whether it was as statistically advanced as for the reanalysis data, separating processes at different timescales. In short, if the comparison is between similar quantities.

Response:

Following the Reviewer's suggestion, we have clarified the computational procedures as follows to enhance methodological transparency and ensure comparable trend quantification across approaches.

For the ML model (Section 2.2.3), we now explicitly state: "The $KZ_{(365,3)}$ filter was then applied to obtain long-term components, and meteorology-driven O_3 trends were derived using Least Square Method." *[Lines 186–187 in the tracked-changes version of the revised manuscript]*

For the GC model (Section 2.2.4), we specify: "The FixE2013 simulation is designed to obtain the MDA8 O_3 concentrations driven solely by meteorological changes and further quantify the meteorological influence on O_3 variations in four seasons. After applying the $KZ_{(365,3)}$ filter to derive the long-term meteorology-driven series, trends were calculated through Least Square Method." *[Lines 213–216 in the tracked-changes version of the revised manuscript]*

The application of identical $KZ_{(365,3)}$ filter and trend-calculation techniques to both ML-derived and GEOS-Chem isolated components ensures inter-model comparability.

It would be interesting to discuss the limitations of working with reanalysis datasets, and indeed the relative strengths and weaknesses of ML and GC data in deriving trends for comparison with observations. The ML and MLR analysis would be stronger if the role of additional chemical, meteorological and climate variables were included to capture a fuller picture of ozone drivers, e.g. solar radiation, soil moisture, vegetation cover, or climate indices like ENSO in driving uncertainty was quantified. Similarly, clustering techniques would be valuable to augment the region based approach and would provide better understanding of the similarity between stations.

Response:

Thank you for your insightful suggestions. Following your suggestion, we have expanded the limitation discussion in Section 4 as follows:

"While this study advances understanding of meteorological contributions to O_3 trends, several limitations warrant attention in future work. Though the reanalysis meteorological dataset is generated observationally, inherent constraints exist, including parameterization uncertainties affecting O_3-relevant physical processes (Janjić et al., 2018; Davidson and Millstein, 2022) and resolution constraints.

Regarding analytical approaches, machine learning efficiently captures nonlinear O_3-meteorology relationships without requiring explicit physicochemical parameterizations, enabling scalable multi-site analysis. However, its inability to resolve chemical mechanisms and sensitivity to predictor selection remain key constraints. Conversely, while GEOS-Chem mechanistically resolves chemistry-transport interactions and enables source attribution, it propagates uncertainties from emission inventories and chemical mechanisms into trend estimates.

Future studies could be improved in the following ways: First, more meteorological datasets and methods should be used to provide more robust uncertainty quantification in O_3-meteorology analyses. Second, implementing clustering techniques (e.g. K-means algorithm) could identify sub-regional drivers at ecotones, enhancing spatial resolution beyond our regional framework. Finally, the Lindeman-Merenda-Gold indices can be employed to quantitatively resolve the contributions of specific meteorological variables. The mechanistic understanding of O_3 drivers would be improved by integrating additional variables, such as solar radiation, soil moisture, and climate indices (e.g. El Niño-Southern Oscillation). Clustering techniques would be valuable to augment the region-based approach and would provide better understanding of the similarity between stations." *[Lines 388–405 in the tracked-changes version of the revised manuscript]*

To enhance its impact, in broad terms, I'd suggest to provide more detailed justifications for their methods, expand the analysis to include additional variables and uncertainties, and focus on identifying the main drivers of ozone trends. By addressing these points, the value of the study would be increased for researchers and policymakers working to mitigate ozone pollution under changing meteorological conditions.

Response:

We sincerely appreciate your overarching recommendations for enhancing this

study's scientific and policy impact. In response, we have comprehensively strengthened methodological descriptions throughout the manuscript (e.g. Lines 125–133, 146–150, 171–187 in the tracked-changes version of the revised manuscript), systematically expanded the limitation discussion in Section 4 (e.g. Lines 389–405 in the tracked-changes version of the revised manuscript), and refined driver attribution narratives (e.g. Lines 329–333, 375–376 in the tracked-changes version of the revised manuscript) to better support policy applications. We are confident these revisions significantly enhance the scholarly rigor and practical relevance of our work.

Finally, regarding data availability, the data do not conform to Copernicus policy which states that "access to data is by depositing them (as well as related metadata) in FAIR-aligned reliable public data repositories, assigning digital object identifiers, and properly citing data sets as individual contributions.". This needs to be addressed via a DOI via archiving through Zenodo or similar of the entire O_3 dataset.

Response:

Thank you for highlighting this important requirement. In full compliance with Copernicus policy, we have deposited the complete research data, including surface MDA8 O_3 observations, and results derived from MLR, RF, and GEOS-Chem analyses in Zenodo. These data are now publicly accessible via https://doi.org/10.5281/zenodo.15859028.

We have added the following new statement to the Data Availability section: "The MDA8 O_3 observations and analytical results derived from MLR, RF, and GEOS-Chem can be obtained from https://doi.org/10.5281/zenodo.15859028." *[Lines 447–448 in the tracked-changes version of the revised manuscript]*

Minor comments

L31 rapid not repaid

Response:

The typo has been revised. *[Line 32 in the tracked-changes version of the revised manuscript]*

L266 uncertainties caused by multi-model is not clear. How are they caused? what is 'multi-model' in this context?

Response:

We wish to clarify that the term "multi-dataset" (rather than "multi-model") appears in Line 266, referring specifically to the use of three meteorological reanalysis products (MERRA2, ERA5, and FNL) to drive MLR models. The calculated uncertainties shown by CV values are caused by the use of different meteorological reanalysis products (MERRA2, ERA5, and FNL).

L296 interesting, but please add reasons why PBLH in FNL introduces these issues.

Response:

The planetary boundary layer (PBL) serves as the primary interface where exchanges of heat, water, momentum, and mass occur between the free atmosphere and the Earth's surface. The intricate interplay between PBL turbulence and the vertical structure of thermodynamic variables presents a substantial challenge in determining the planetary boundary layer height (PBLH) (Teixeira et al., 2021). Consequently, uncertainties in PBLH within reanalysis datasets may stem from the adoption of divergent PBLH derivation methodologies. As suggested by Guo et al. (2021), when validating PBLH, the NCEP FNL dataset tends to be more vulnerable to the impacts of complex underlying surfaces compared to ERA5 and MERRA2.

We have added the reason as follows: "and that its performance may be constrained by complex underlying terrain and static instability (Guo et al., 2021)." *[Lines 329–330 in the tracked-changes version of the revised manuscript]*

L300 should read 'for the whole of China'

Response:

The usage has been modified. *[Line 338 in the tracked-changes version of the revised manuscript]*

References:

Chattamvelli, R., Shanmugam, R. Measures of Spread. In: Descriptive Statistics for Scientists and Engineers. Synthesis Lectures on Mathematics & Statistics. Springer, Cham. https://doi.org/10.1007/978-3-031-32330-0_3, 2019

Chen, L., Gao, Y., Zhang, M., Fu, J. S., Zhu, J., Liao, H., Li, J., Huang, K., Ge, B., Wang, X., Lam, Y. F., Lin, C.-Y., Itahashi, S., Nagashima, T., Kajino, M., Yamaji, K., Wang, Z., and Kurokawa, J.: MICS-Asia III: multi-model comparison and evaluation of aerosol over East Asia, Atmos. Chem. Phys., 19, 11911–11937,

https://doi.org/10.5194/acp-19-11911-2019, 2019.

Davidson, M. R. and Millstein, D.: Limitations of reanalysis data for wind power applications, Wind Energy, 25, 1646–1653, https://doi.org/10.1002/we.2759, 2022.

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., and Hueglin, C.: Random forest meteorological normalisation models for Swiss PM_{10} trend analysis, Atmos. Chem. Phys., 18, 6223–6239, https://doi.org/10.5194/acp-18-6223-2018, 2018.

Guo, J., Zhang, J., Yang, K., Liao, H., Zhang, S., Huang, K., Lv, Y., Shao, J., Yu, T., Tong, B., Li, J., Su, T., Yim, S. H. L., Stoffelen, A., Zhai, P., and Xu, X.: Investigation of near-global daytime boundary layer height using high-resolution radiosondes: first results and comparison with ERA5, MERRA-2, JRA-55, and NCEP-2 reanalyses, Atmos. Chem. Phys., 21, 17079–17097, https://doi.org/10.5194/acp-21-17079-2021, 2021.

Högel, J., Schmid, W., and Gaus, W.: Robustness of the standard deviation and other measures of dispersion, Biom. J., 36, 411–427, https://doi.org/10.1002/bimj.4710360403, 1994.

Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A., and Weston, P.: On the representation error in data assimilation, Q. J. R. Meteorolog. Soc., 144, 1257–1278, https://doi.org/10.1002/qj.3130, 2018.

Lu, X., Liu, Y., Su, J., Weng, X., Ansari, T., Zhang, Y., He, G., Zhu, Y., Wang, H., Zeng, G., Li, J., He, C., Li, S., Amnuaylojaroen, T., Butler, T., Fan, Q., Fan, S., Forster, G. L., Gao, M., Hu, J., Kanaya, Y., Latif, M. T., Lu, K., Nédélec, P., Nowack, P., Sauvage, B., Xu, X., Zhang, L., Li, K., Koo, J.-H., and Nagashima, T.: Tropospheric ozone trends and attributions over east and southeast Asia in 1995–2019: an integrated assessment using statistical methods, machine learning models, and multiple chemical transport models, https://doi.org/10.5194/egusphere-2024-3702, 17 December 2024.

Teixeira, J., Piepmeier, J. R., Nehrir, A. R., Ao, C. O., Chen,S. S., Clayson, C. A., Fridlind, A. M., Lebsock, M., McCarty, W., Salmun, H., Santanello, J. A., Turner, D. D., Wang, Z., and Zeng, X.: Toward a global planetary boundary layer observing system: the NASA PBL incubation study team report, NASA PBL Incubation Study Team, 134

pp, 2021

Varoquaux, G., Colliot, O.: Evaluating Machine Learning Models and Their Diagnostic Value. In: Colliot, O. (eds) Machine Learning for Brain Disorders. Neuromethods, 197. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-3195-9_20, 2023

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.

Wang, M., Chen, X., Jiang, Z., He, T.-L., Jones, D., Liu, J., and Shen, Y.: Meteorological and anthropogenic drivers of surface ozone change in the North China Plain in 2015–2021, Sci. Total Environ., 906, 167763, https://doi.org/10.1016/j.scitotenv.2023.167763, 2024a.

Wang, T., Li, N., Li, Y., Lin, H., Yao, N., Chen, X., Li Liu, D., Yu, Q., and Feng, H.: Impact of climate variability on grain yields of spring and summer maize, Comput. Electron. Agric., 199, 107101, https://doi.org/10.1016/j.compag.2022.107101, 2022.

Wang, X., Jiang, L., Guo, Z., Xie, X., Li, L., Gong, K., and Hu, J.: Influence of meteorological reanalysis field on air quality modeling in the Yangtze River Delta, China, Atmos. Environ., 318, 120231, https://doi.org/10.1016/j.atmosenv.2023.120231, 2024b.

Weng, X., Forster, G. L., and Nowack, P.: A machine learning approach to quantify meteorological drivers of ozone pollution in China from 2015 to 2019, Atmos. Chem. Phys., 22, 8385–8402, https://doi.org/10.5194/acp-22-8385-2022, 2022.

Response to Referee #2

This study focuses on attributing the increase in surface ozone concentrations in China using different datasets and methods. Specifically, it compares discrepancies in attribution results derived from multiple reanalysis datasets and approaches—including data-driven statistical models, machine learning models, and process-driven atmospheric chemistry transport models (GEOS-Chem). The work holds significant value for more accurately quantifying ozone attribution in China and elucidating inconsistencies among existing studies. Overall, the manuscript is clearly written, and the conclusions are well-supported. I recommend publication after minor revision.

Response:

We sincerely thank the referee for the decision and constructive comments. The manuscript has been revised accordingly, and our point-by-point responses are provided below. The referee's comments are shown in black, and our replies are highlighted in blue. A tracked-changes version of the revised manuscript is also clearly showing the changes made.

While the results present variations in attribution outcomes across datasets/methods, the discussion regarding the underlying causes of these differences can be sharper. One recent study from the Tropospheric Ozone Assessment Report (TOAR, https://egusphere.copernicus.org/preprints/2024/egusphere-2024-3702/) has similarly addressed methodological disparities in ozone trend attribution in China. It would be helpful if the authors could incorporate a comparative evaluation between the two studies to enrich the discussion of result discrepancies.

Response:

We sincerely appreciate your valuable suggestions to improve our manuscript. We have incorporated a comprehensive comparative evaluation as follows:

1) "Lu et al. (2024) compared meteorology-driven O_3 trends derived from ERA5- and MERRA2-driven MLR models during the summer of 2013–2019. Their findings revealed that ERA5-derived trends were lower than those from MERRA2 in YRD and PRD, whereas trends derived from ERA5 were comparable to those from MERRA2 in BTH. This inter-study consensus further validates the robustness of our methodological framework." *[Lines 271–274 in the tracked-changes version of the revised manuscript]*
2) "Lu et al. (2024) also demonstrated a high degree of consistency among the MLR, ML, and GC models in PRD during summer. Specifically, all three

models indicated that meteorology contributed approximately 25% of O_3 variability over the period 2013–2019." *[Lines 353–355 in the tracked-changes version of the revised manuscript]*

3) "The GC's systematic overestimation of O_3 concentrations, as well as underestimation of O_3 increases, were also reported by Lu et al. (2024), in which the GC captured 13.6 ~ 81.1% of the observed O_3 increases in China during the summer of 2000–2019." *[Lines 370–372 in the tracked-changes version of the revised manuscript]*

Specific Comments:

Line 14: Clarify the specific ozone metric (e.g., MDA8, annual mean).

Response:

We have clarified the ozone metric as the "maximum daily 8-hour average O_3". *[Line 14 in the tracked-changes version of the revised manuscript]*

Sections 2.2.2–2.2.4: The temporal scale of statistical analyses (daily, monthly, or seasonal) is unclear. Please specify.

Response:

Thanks for your suggestion. We have revised the following statements to make it clear:

"In the decomposition process, X(t) represents the original daily time series" *[Line 122 in the tracked-changes version of the revised manuscript]*

"In this study, all statistical analyses were performed at the seasonal scale (spring: March-April-May; summer: June-July-August; autumn: September-October-November; winter: December-January-February)." *[Lines 132–133 in the tracked-changes version of the revised manuscript]*

"After establishing MLR models for the short-term and baseline components in each season", and "The constructed MLR models driven by meteorological variables from ERA5, MERRA2, or FNL in each season will allow a comprehensive analysis of multi-dataset uncertainties." *[Lines 146, 151–152 in the tracked-changes version of the revised manuscript]*

"quantify the meteorological influence on O_3 variations in four seasons" *[Line 214 in the tracked-changes version of the revised manuscript]*

"For each season, when examining the uncertainties arising from different datasets" *[Lines 231–232 in the tracked-changes version of the revised manuscript]*

Figure 6: The GEOS-Chem model suggests a smaller meteorology-driven ozone trend compared to data-driven methods. To what extent might this stem from GEOS-Chem underestimating observed ozone trends? While Figure S3 indicates the model captures overall temporal variability, please provide a quantitative evaluation of trend performance.

Response:

Thank you for your suggestion. Quantitative trend evaluation has now been provided in Table S4. During 2013–2022, GEOS-Chem simulated substantially lower O_3 trends than observations in China and BTH, but captured observed O_3 increases in YRD and PRD. Compared with 2013–2017, trend performance notably improved in BTH during 2018–2022. These quantitative comparisons confirm that the smaller meteorology-driven trends from GEOS-Chem partially stem from its underestimation of observed O_3 trends, especially during 2013–2017.

Following the Reviewer's suggestion, we have added newly Table S4 to provide a quantitative evaluation of trend performance and updated the discussion about multi-method uncertainty as follows: "As shown in Fig. S3 and Table S4, this difference could partly be attributed to the higher O_3 levels and lower O_3 increases simulated by the GC model before 2018." *[Lines 368–370 in the tracked-changes version of the revised manuscript]*

Table S4. Trends in GEOS-Chem-simulated (Sim) and observed (Obs) monthly mean MDA8 O_3 concentrations (ppb yr^{-1}) during 2013–2022, 2013–2017, and 2018–2022.

Region / Time period	China		BTH		YRD		PRD	
	Sim	Obs	Sim	Obs	Sim	Obs	Sim	Obs
2013–2022	–0.05	+0.84	–0.11	+0.89	+0.40	+0.97	+0.39	+1.07
2013–2017	–0.84	+0.27	–0.82	+0.22	–0.40	+0.29	+0.22	+0.31
2018–2022	–1.78	–0.86	–2.86	–2.49	–2.00	–0.96	–1.38	+0.17

References:

Lu, X., Liu, Y., Su, J., Weng, X., Ansari, T., Zhang, Y., He, G., Zhu, Y., Wang, H., Zeng, G., Li, J., He, C., Li, S., Amnuaylojaroen, T., Butler, T., Fan, Q., Fan, S., Forster, G. L., Gao, M., Hu, J., Kanaya, Y., Latif, M. T., Lu, K., Nédélec, P., Nowack, P.,

Sauvage, B., Xu, X., Zhang, L., Li, K., Koo, J.-H., and Nagashima, T.: Tropospheric ozone trends and attributions over east and southeast Asia in 1995–2019: an integrated assessment using statistical methods, machine learning models, and multiple chemical transport models, https://doi.org/10.5194/egusphere-2024-3702, 17 December 2024.

Response to the comment on "Meteorological influence on surface ozone trends in China: Assessing uncertainties caused by multi-dataset and multi-method" by X. Wang et al. (Ms. Ref. No.: EGUSPHERE-2025-1880)

Response to Referee #1

I thank the authors for their time in responding to my comments. Regarding their response to the question about the R^2>0.5 criterion, I appreciate the need to balance model skill and spatial coverage. However, it would strengthen the justification if the authors could show how results change when this threshold is tightened or relaxed. For example, a table in the supplementary could include attribution results when the currently excluded stations are reinstated or when a higher bar such as R^2>0.6 is applied (more in line with the skill of models in Weng et al., 2022). This would allow assessment of whether lower-skill stations systematically bias the meteorology–emissions separation, dampen trends, or increase variability in the CV metric. Such a sensitivity analysis would also align with best practice in attribution studies, where robustness to model-skill thresholds is important for ensuring that conclusions are not an artefact of an arbitrary cutoff.

Response:

Thanks for your insightful suggestion. We fully agree that evaluating the robustness to different model-skill thresholds is essential for ensuring the reliability of our conclusions. Following the reviewer's suggestion, we have conducted a comprehensive sensitivity analysis by implementing two additional R^2 thresholds ($R^2 \geq 0.6$ and $R^2 \geq 0.4$) to filter unreliable stations across all seasons and regions. The complete results are presented in Table S3 and summarized as follows:

4) Tightening the threshold to $R^2 \geq 0.6$ reduced the number of available stations by 13.5% (autumn) to 39.4% (summer) compared to $R^2 \geq 0.5$, while relaxing to $R^2 \geq 0.4$ increased station inclusion by 4.1% (autumn) to 15.5% (summer).
5) Meteorology-driven MDA8 O_3 trends derived from the MLR, RF, and GC models exhibited minor variations across the three thresholds. The maximum difference was observed in the MLR model for summer in China, with trends of +0.23 ppb yr^{-1} ($R^2 \geq 0.5$), +0.31 ppb yr^{-1} ($R^2 \geq 0.6$), and +0.20 ppb yr^{-1} ($R^2 \geq 0.4$). The overall directional consistency of meteorology-driven O_3 trends across thresholds confirms that our primary conclusions are not artifacts of an arbitrary cutoff.
6) The uncertainty metric (CV) for multi-method spread showed minor changes

across R^2 thresholds, indicating that methodological uncertainties are robustly quantified regardless of the station inclusion criteria.

We have added the following statements to the manuscript:

"To evaluate the robustness of the $R^2 \geq 0.5$ criterion, we performed sensitivity analyses using thresholds of $R^2 \geq 0.6$ and $R^2 \geq 0.4$, to ensure that our conclusions are not an artifact of an arbitrary cutoff (Table S3)."*[Lines 171–173 in the tracked-changes version of the revised manuscript]*

"Meteorology-driven MDA8 O_3 trends exhibited minor variations across different R^2 thresholds (Table S3), indicating that the trends are not an artifact of an arbitrary cutoff." *[Lines 335–336 in the tracked-changes version of the revised manuscript]*

"The low uncertainties are further corroborated by consistent CV estimates derived under different RF's R^2 thresholds (Table S3). " *[Lines 344–345 in the tracked-changes version of the revised manuscript]*

Table S3. Comparison of meteorology-driven MDA8 O_3 trends (ppb yr^{-1}) derived from multiple linear regression (MLR), random forest (RF), and GEOS-Chem (GC) models using different model performance criteria (the R^2 of the RF model ≥ 0.5, 0.6, and 0.4) across seasons and regions.

Season	R^2	Available stations [a]	China				BTH				YRD				PRD			
			MLR	RF	GC	CV [b]	MLR	RF	GC	CV	MLR	RF	GC	CV	MLR	RF	GC	CV
Spring	$R^2 \geq 0.5$	1162	+0.48	+0.15	+0.14	0.75	+0.26	–0.01	+0.12	1.06	+0.80	+0.25	+0.37	0.61	+0.94	+0.78	+0.77	0.11
	$R^2 \geq 0.6$	882	+0.50	+0.14	+0.14	0.81	+0.26	–0.00	+0.12	1.07	+0.86	+0.25	+0.37	0.66	+0.89	+0.75	+0.77	0.10
	$R^2 \geq 0.4$	1250	+0.47	+0.16	+0.14	0.73	+0.26	–0.01	+0.12	1.06	+0.77	+0.25	+0.36	0.60	+0.89	+0.75	+0.77	0.09
Summer	$R^2 \geq 0.5$	1059	+0.23	+0.01	–0.03	2.00	+0.03	–0.16	–0.18	1.08	+0.45	+0.25	+0.40	0.29	+0.07	+0.13	+0.10	0.32
	$R^2 \geq 0.6$	642	+0.31	+0.02	–0.01	1.69	+0.06	–0.20	–0.17	1.40	+0.51	+0.21	+0.34	0.42	+0.08	+0.13	+0.10	0.25
	$R^2 \geq 0.4$	1223	+0.20	+0.00	–0.04	2.39	+0.04	–0.15	–0.19	1.24	+0.43	+0.25	+0.39	0.27	+0.06	+0.13	+0.09	0.34
Autumn	$R^2 \geq 0.5$	1203	+0.15	+0.34	+0.03	0.88	–0.27	+0.19	–0.13	3.38	+0.37	+0.53	+0.24	0.38	+0.83	+0.81	+0.53	0.23
	$R^2 \geq 0.6$	1041	+0.18	+0.36	+0.04	0.85	–0.27	+0.19	–0.14	3.40	+0.40	+0.55	+0.24	0.39	+0.85	+0.83	+0.54	0.23
	$R^2 \geq 0.4$	1252	+0.15	+0.33	+0.03	0.90	–0.27	+0.19	–0.13	3.38	+0.36	+0.52	+0.24	0.38	+0.83	+0.81	+0.53	0.23
Winter	$R^2 \geq 0.5$	1094	+0.30	+0.12	+0.25	0.40	+0.26	+0.13	+0.09	0.55	+0.19	+0.06	+0.27	0.59	+0.72	+0.64	+0.72	0.07
	$R^2 \geq 0.6$	738	+0.33	+0.13	+0.24	0.41	+0.26	+0.13	+0.10	0.53	+0.20	+0.06	+0.27	0.59	+0.70	+0.66	+0.73	0.05
	$R^2 \geq 0.4$	1217	+0.28	+0.12	+0.25	0.40	+0.26	+0.13	+0.10	0.55	+0.17	+0.06	+0.27	0.62	+0.70	+0.63	+0.72	0.07

[a] "Available stations" denotes the number of state-controlled monitoring stations with the R^2 of the Random Forest model ≥0.5, 0.6, and 0.4.

[b] The absolute value of the coefficient of variation (CV) is calculated by the standard deviation of the trends derived from MLR, RF, and GC models divided by the mean.

Minor note: the reference "Wang 2024c" is still incomplete.

Response:

Thanks for your suggestion. We have corrected the reference "Wang 2024c" as follows:

Wang, X., Zhu, J., Li, K., Chen, L., Yang, Y., Zhao, Y., Yue, X., Gu, Y., and Liao, H.: Meteorology-driven trends in $PM_{2.5}$ concentrations and related health burden over India, Atmos. Res., 308, 107548, https://doi.org/10.1016/j.atmosres.2024.107548, 2024c.

[Lines 623–624 in the tracked-changes version of the revised manuscript]