Reviewer Comments: Black font

Author Responses: *Black italics*

Manuscript Revisions: *Blue italics*

Response to Reviewer #1

1. Re-reviewing this manuscript with the authors' responses to the first review does not provide any more encouragement that the paper adds significant new knowledge to the atmospheric chemistry domain. In terms of regulatory issues, it may be fine, but the extreme limitation of their diagnostics makes the paper useless for diagnosing the problems with modeling surface ozone. Further, calculating human impacts from 6-month OSDMA8 ozone when you show the incredible biases of the modeled ozone seems like going too far, e.g., the numbers in Table 3 have no uncertainties related to the obvious model bias, so how can they be published? "The ozone seasonal daily maximum 8-hour average mixing ratio " = six-month running monthly mean). This diagnostic totally obscures all issues of extremes and even hides the seasonal cycle, removing most of the inter-month variability. The use of a six-month DMA8 because some of the models only did that is still a poor excuse. Not looking at the diel cycle and extreme days means you cannot understand why the models fail to match DMA8. The CAMS and TCR-2 results seem sensible (2-hr and 3-hr ozone) and at a more reasonably resolution to make a useful comparison.

Response:
We thank Reviewer #1 for reviewing our manuscript a second time, and for their thoughtful comments. We have made significant changes to the paper in response to reviews. These revisions include:

1) *We have changed the order of Sections 4 and 5 in the manuscript, and have modified several results to include uncertainty. By reordering these sections, we discuss the performance evaluation of the different datasets with respect to measurements first, and then include estimates of uncertainty from the performance evaluation in results for the inter-comparison of the different datasets.*
2) *As Reviewer #1 suggested, we now add a grid-to-grid comparison method in addition to the grid-to-point method used in earlier versions of the manuscript. The grid-to-grid method follows from the methods of Schnell et al. (2015)*

In this comment, Reviewer #1 recommends that the paper be rejected since our use of an aggregate annual ozone metric obscures diurnal and seasonal cycles, and therefore the evaluation with respect to observations is inadequate. We agree that for some air quality related research, evaluating these short-term timescales is essential. However, the primary goal of this paper is to intercompare global datasets specifically for their use in long-term health exposure

related studies. For this purpose, the 6-month seasonal daily maximum 8-hour average (OSDMA8) ozone is a standard metric, as used by the Global Burden of Disease (GBD) and in the World Health Organization (WHO) Air Quality Guidelines. This metric is designed to capture the long-term exposure during these peak seasons which is most relevant for assessing long-term health impacts. In contrast, short-term exposure studies use daily or hourly metric to investigate the health risk related to acute air pollution events. Reviewer #1 also suggests that we conduct a traditional model evaluation for only the chemical reanalyses, since they report 2- or 3-hourly ozone. But such comparisons with observations have already been published (Jones et al., 2025; Miyazaki et al., 2025) and this is not the point of this paper. Rather, we compare these datasets for an annual metric OSDMA8 that is used by GBD and in the WHO Air Quality Guidelines. In doing so, we aim to compare datasets of global ground-level ozone for this health-focused metric, that have been generated by different methods – geostatistical data fusion, machine learning, and chemical reanalyses. Estimating how biases in these available ozone mapping products lead to systematic differences in exposure estimates is a significant and important topic for the community. Because these datasets have not been compared systematically, our work makes a direct contribution to this ongoing challenge. We have carefully considered Reviewer #1's suggestion regarding using a short-term metric like 2-hr and 3-hr ozone for intercomparison but addressing this would require a fundamental change to the purpose and scope of this paper, which we believe would not be appropriate here.

Regarding Reviewer #1's comments about uncertainty, we have made substantial changes to the paper. We have changed the order of Sections 4 and 5 to present dataset evaluation with respect to observations first, before the dataset intercomparison. We have also used estimates of uncertainty from the evaluation section (now Section 4) to quantify uncertainties that are now presented in the intercomparison section (now Section 5). We thank Reviewer #1 for these valuable comments. In this way, we expand our discussion to provide clearer interpretations of how the uncertainties from evaluation impact on the paper's main findings. With these interpretations, we draw more meaningful conclusions.

2. Looking at these two responses shows that there is little understanding of the problem:
"However, the goal of our work is to assess the accuracy of gridding products in estimating measured ground concentration point values.
"That is why we have adopted this grid-to-point evaluation approach. Any gridded product, such as that of Schnell et al, uses an interpolation of a point value that introduces its own uncertainties and biases. To avoid these additional uncertainties, we directly compared observed point-level ozone values to the nearest available grid estimates.
You did not "avoid" the uncertainties with the problem that Schnell dealt with, you simply ignored them. There is nothing here that attempts to deal with the problems of creating a high-resolution map based on the site measurements and then integrating it to get the cell average.

This problem is what Schnell's first paper spent most of its effort on. I may have missed it but I find no serious effort to optimize the point-area comparison.
“Previous research has adopted a 1º×1º grid-cell-averaged hourly ozone data from TOAR observations to evaluate global chemistry model performance over North America and Europe, which is suitable for analyzing extremes and validating seasonal and diel ozone cycles (Schnell and Prather, 2017; Schnell et al., 2015).”

“We adopted a grid-to-point evaluation approach, where the data from each TOAR-II observation site was matched with a corresponding grid cell in each dataset. For grid cells with a TOAR-II observation but no valid estimate in a dataset (NA value), we used the nearest valid estimate instead.”
Minor point – Schnell gathered all the AQ station data in N.Am. and EU, not specifically the TOAR data. Really, a "grid-to-point evaluation" is simply saying that every point in a cell should have the same value as the mean of that cell. This is quite apparently false when you have several nearby sites. The algorithm here (last sentence) does not really address how "far" a single station can reach? or why? "nearest" valid estimate is not provide a scientific comfort level.
"our focus is explicitly on assessing whether global ozone mapping products can reasonably estimate point-level concentrations at locations lacking monitoring stations.
I cannot see how you can begin to do that without models that resolve station-to-station differences (1 km) and so you really cannot say this.

Response:

Reviewer #1 questions our use of a grid-to-point comparison of observations with gridded ozone estimates, suggesting that we spatially average observations to a grid before comparing with the six ozone datasets. Such grid-to-point comparison methods are widely used in our field, and the specific dataset developed by Schnell et al. is limited to North America and Europe and so does not match the spatial or temporal scope of our study. We thank Reviewer #1 pointing out this concern, and we agree with Reviewer #1 that these issues pose a challenge when evaluating coarse-resolution datasets against observations. To address this issue, we have implemented two evaluation methods in our revised manuscript.

1. *Grid-to-Grid evaluation: We have re-gridded the TOAR-II observations onto the native grid resolution of each dataset. In doing so we use methods similar to those of Schnell et al. (2015), but here we gridded observations for the whole TOAR-II dataset globally and for more years than had been done previously. This method addresses the representativeness issue by comparing the value of re-gridded observational grid cell to the value of dataset’s grid cell for the same spatial resolution.*

2. *Grid-to-Point evaluation: This is the traditional method used in the original manuscript where the value of the dataset’s grid cell is compared to all observations within that cell.*

This method can ensure that all evaluations are the same sample size given by the number of observations.

We thank Reviewer #1 for their comments that led us to include the grid-to-grid methods, which we feel strengthened the manuscript.

We have revised the methods as follows:

Line 214: Considering that the six datasets have different resolutions and are designed for different applications, we adopted a dual evaluation strategy to provide a comprehensive assessment of their performance. The first method is a grid-to-grid evaluation. Similar to the approach of Schnell et al. (2015), we re-gridded TOAR-II observations to a 0.1° x 0.1° resolution by an inverse distance weighted method and then aggregated them to match the native resolution of each of the six datasets. In this approach, the sample size for each evaluation varies reflecting the varying resolution of the datasets; for 2016, BME had 173,718 grid cell pairs, NJML had 7,099, UKML had 162,419, CAMS had 4,614, GEOS-Chem had 782, and TCR-2 had 2,195. We also adopted the grid-to-grid evaluation method for regional evaluations, as it provides better spatial representativeness over large areas. To quantify the uncertainty of the six datasets' estimates, we determined the lower and upper bounds (95% confidence interval), derived from the grid-to-grid regression analysis performed between the TOAR-II observations and each of the six datasets at their native resolutions.

Line 224: The second method is a standard grid-to-point evaluation. This approach ensures a consistent sample size across all datasets by comparing each dataset's estimate at the grid cell containing an observation location. For grid cells containing a TOAR-II site but no valid estimate (NA value), we used the nearest valid estimate instead. This method captures a penalty for missing data and coarse resolution, only BME, NJML, and UKML had a small number of missing estimates at TOAR-II locations. The grid-to-point method was used to evaluate model bias, as it ensures a consistent sample size across all datasets when performing evaluations on different quantiles of the TOAR-II observations.

By presenting results from both methods in section 4.1, we provide a more robust and comprehensive assessment of dataset performance. Overall, our main conclusions and the relative performance rankings of the datasets remain largely consistent across both evaluation methods. As expected, the grid-to-grid approach generally results in lower RMSE and higher R^2 values, as it averages out localized errors that are more prominent in the direct grid-to-point evaluation. However, the difference of two methods does not change the fundamental takeaways of our analysis.

We have updated the results throughout the manuscript to describe the evaluation from both the grid-to-grid and grid-to-point methods. The following are some key passages from the manuscript, updated to reflect the revisions:

Line 275: For 2016, BME outperforms other datasets in both evaluation method, with the highest R^2 (0.75 for grid-to-grid, 0.63 for grid-to-point) and lowest RMSE (4.25 ppb for grid-to-grid, 5.28 ppb for grid-to-point), its density cores intersecting the y=x line.

Line 280: In Fig. 1(a), all three chemical reanalysis datasets exhibit a moderate R^2 ranging from 0.51 to 0.60 for grid-to-grid and 0.35 to 0.41 for grid-to-point, comparable to the performance of the machine learning datasets, which have R^2 values of 0.50 and 0.56 for grid-to-grid, 0.37 and 0.38 for grid-to-point. Among these five datasets, CAMS has the lowest RMSE (6.00 ppb for grid-to-grid and 7.59 ppb for grid to point), which is better than other chemistry reanalysis products but relatively low R^2 (0.51 for grid-to-grid and 0.35 for grid-to-point). Its density cores slightly below the y=x line suggests CAMS estimates are marginally lower than TOAR-II observations. GEOS-Chem and TCR-2 demonstrate adequate performance, albeit with higher RMSE values of 8.47 ppb and 10.26 ppb for grid-to-grid, 10.27 ppb and 13.23 ppb for grid-to-point, respectively.

Line 362: In grid-to-grid evaluation, GEOS-Chem shows an overall better performance in R^2 than CAMS, TCR-2 and UKML.

Line 369: From 2006 to 2016, the performance rankings derived from R^2 values varied significantly between the two evaluation scenarios, whereas the RMSE based rankings were nearly consistent.

Line 550: In instances of missing model estimates, we default to the nearest valid estimate to evaluate with TOAR-II observations or re-gridded grid cell. For datasets with coarse spatial resolution, this method may increase or reduce bias by double counting.

3. The authors have done nothing to address the primary problems with this analysis: the use of 6-month averages of MDA8 to compare with models; and the fundamental methodology of how to compare a mean grid-cell value with point measurements, especially when there are several in a cell. The latter is clearly major unresolved problem (here at least). What happens if you have three different sites within a cell, each with three different values – how does one compare and derive R^2? I realize that the authors may not be able to do this given the material they have, and thus this paper belongs in an air quality management journal that deals with meeting regulations, not in a science journal like ACP.

Response:

Please see our responses above to (#1) the overall purpose and scope of the paper, and (#2) the evaluation method.

Reviewer #1 argues that our paper is outside of the scope of ACP. We acknowledge Reviewer #1's concern regarding the scope of ACP. However, we strongly believe that our paper fits well within the journal's scope for the following reasons:

1. *The datasets we evaluate and compare are of significant interest to the ACP readership, which includes atmospheric scientists, climate modelers, and policy makers interested in the impacts of ozone exposure on health, agriculture, and ecosystems. Therefore, our work is directly relevant to the ACP community.*
2. *Our findings of significant differences among the datasets despite the fact that some methods use some of the same inputs, which has not been shown previously, highlights the importance of continued research on global ozone distributions. Machine learning and other methods have not converged on a single correct global ozone distribution. This is a direct contribution to atmospheric research community.*
3. *This paper was submitted as part of the TOAR-II special issue, and our work has been presented at TOAR-II online meetings, as well as at the CMAS conference. Our work has been received enthusiastically, with colleagues commenting both that it was conducted thoroughly, and that such a systematic comparison of different datasets that are used widely is overdue and important.*

Response to Reviewer #2

The authors have addressed some of the reviewers' suggestions, and there are some improvements in the revised manuscript. However, several important issues remain and require further revision. Failure to address the comments from the first round seriously compromises the value of this study. Unless these critical points are properly dealt with, the contribution of the work remains questionable.

Response:
We thank Reviewer #2 for their careful reading of the manuscript and for providing critical feedback. We agree with the reviewer that our revised manuscript did not sufficiently address the comments in first round. We have thoroughly revised the manuscript to address these critical points.

1. Impact of data uncertainty on related analysis and reorganizing structure (i.e. third comment in the 1st round): Regarding this issue, I do not agree with your response. While readers can draw their own conclusions, it is the authors' fundamental responsibility to provide clear and specific interpretations based on the findings. Leaving key aspects of the analysis entirely up to the reader undermines the completeness and clarity of the work. This approach also deviates from the stated aims and title of the study, which emphasize evaluating uncertainty and accuracy through comparisons between observation and datasets. One of the core values of this research should be to account for these uncertainties and to draw meaningful conclusion – particularly about implications for public health and agriculture. Moreover, Sections 4 and 5 lack a logical flow, which further weakens the clarity and impact of the manuscript. Without a through and well-reasoned interpretation of the results (i.e., the uncertainty and its impact on trend analysis and

ozone exposed population assessment), the manuscript does not fulfill its stated objectives and, in its current form, is not suitable for acceptance in ACP.

Response:

Thank you for this valuable suggestion to improve the manuscript's flow, and we apologize for this oversight in the first revision. Now we have reversed the order of sections 4 and 5, to present dataset evaluation with respect to observations first, before the dataset intercomparison. To directly address your concern, we have also used the results of the model evaluation to quantify uncertainty in some of the results presented in the dataset intercomparison (now Section 5). Specifically, for the exposure assessment (now in Section 5.4), we add the lower and upper bound of population exposure to the OSDAM8 level in Supplementary Figure S12. In Table 3, for share of population exposure to different ozone thresholds, we now add the lower and upper bound based on the evaluation results.

In addition to changing the order of Sections 4 and 5, we have also changed the text as follows:

Line 266: To quantify the uncertainty in our exposure analysis, we established lower and upper bounds for all population exposure and share of population estimates. The OSDMA8 95% confidence interval (CI) for each dataset is determined through a grid-to-grid linear regression between each dataset and the re-gridded TOAR-II observations based on 0.1°× 0.1°grid cells.

Line 438: We also calculated the distribution of population regarding the lower and upper bounds of OSDMA8 from 2006 to 2016 for each dataset, as shown in Figure S12.

Line 452: Results are presented as the estimate with the lower and upper bound in parentheses (e.g., 42% [24%, 66%]). Detailed table of population share for each year (2006 to 2016) are shown in Table S10.

We have also substantially expanded our discussion to provide a clearer interpretation of how the evaluation results impact on the paper's main findings. For example, the ozone trend analysis has been revised to include how systematic biases found in our evaluation (some datasets are overestimate) contribute to the divergent long-term trends among datasets. With these interpretations, we draw more meaningful conclusions.

Line 471: In addition, for chemical reanalysis datasets, there is a clear trade-off between capturing the spatial pattern and the accuracy. As shown in Fig. 2, TCR-2, GEOS-Chem all have widespread overestimation, but they often capture spatial patterns more effectively (higher R^2). Conversely, CAMS exhibits low bias in RMSE but shows worse spatial correlation in China. All six datasets show a reduced performance at higher ozone concentrations (>50 ppb), which may complicate their accuracy for assessing long term high-pollution exposure. Furthermore, most datasets perform better in regions with lower monitoring density (e.g., the United States and China) than in those with higher density (e.g., Japan and South Korea), which suggests that resolving high-resolution local ozone distributions remains challenging even with a good amount

of observational data. The performance of each dataset impacts the accuracy of trend analysis (Fig. 5 and Fig. 6) and population exposure assessment (Fig. 10), shown as uncertainty in these Figures, which may lead to different results when compared to the WHO guideline and interim target.

Line 490: From the comparison, the large disagreements among the six datasets regarding ozone trends, population exposure, and concentration estimates are a direct consequence of the systematic biases and performance issues identified in the evaluation.

Line 497: These uncertainties critically undermine the reliability of population exposure assessment.

Line 503: More importantly, the evaluation reveals that all datasets perform poorly at high ozone levels (> 50 ppb). This highlights the importance of removing systematic biases from these data sets before applying them to exposure estimates.

Line 509: And from the evaluation, we find that all datasets perform well in the United States, which makes the downward trend more reliable.

Line 570: Regionally, all datasets show a downward trend in North America, and the evaluation results make this trend more reliable. Only BME and NJML datasets demonstrate a downward trend in East Asia, and they also fit well with TOAR-II observations in population density distribution.

Line 579: The coarse-resolution datasets, GEOS-Chem and TCR-2, perform well in grid-to-grid evaluations at their native resolutions, making them effective for studying long-term regional ozone effects. However, because of their coarser resolutions, these two datasets cannot capture site-level distributions and exhibit greater bias than the higher-resolution BME, CAMS, and NJML datasets. UKML, despite its relatively fine resolution (0.125°), shows larger biases and a lower R^2. The superior performance of BME and NJML should be noted with the fact that both datasets use observational data for input or training, which gives them an inherent advantage in these evaluations.

2. Again, regarding the fourth comment in the first round, I don't think the authors thoroughly evaluate the performance of the datasets. The authors missed their scientific discussion and the implications.

Response:

We thank the reviewer for this excellent suggestion. We agree that a direct comparison of population exposure at the observation sites provides a more thorough evaluation and strengthens the paper's scientific discussion. We have now done this, but we are limited in doing this to the 3 world regions with a density of observation sites. We re-gridded TOAR-II observations to a 0.1° x 0.1° resolution by an inverse distance weighted method. The results are

now presented in Figure 3 and discussed in Section 4.2. We use these results to draw clearer conclusions about the implications for regional exposure studies and to better identify the uncertainty in the following exposure assessments. We also present these results for the lower and upper bounds for the population exposure estimates, given uncertainty based on the evaluation of each dataset in Figure S12. We think that this is a valuable addition to the paper, so thank you for the suggestion.

The text is revised as follows:

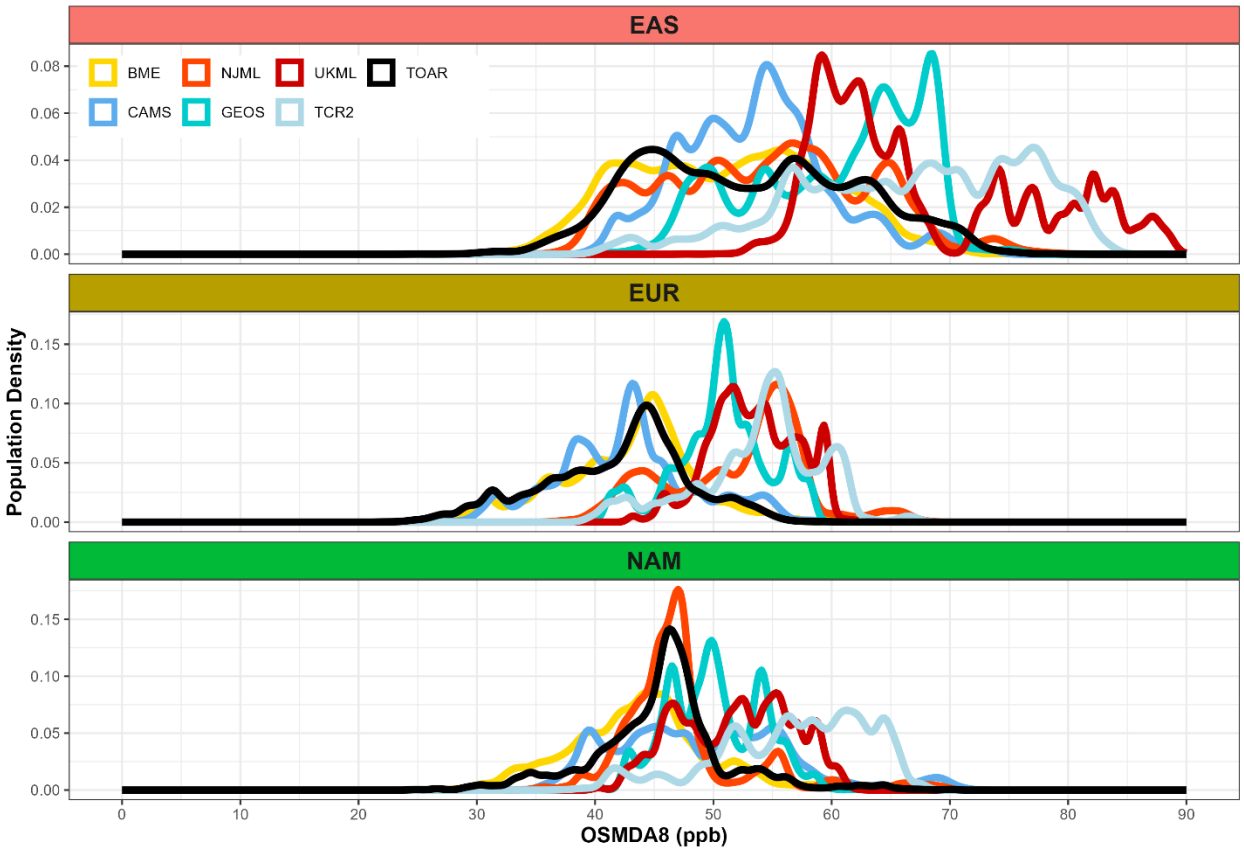

Figure 3: Population-weighted exposure distributions for OSDMA8 in 2016 in three regions: East Asia (EAS), Europe (EUR), and North America (NAM) (regions defined in Table S7). Each panel compares the distribution derived from the TOAR-II observations (black line) with estimates from six datasets (colored lines), calculating the population-weighted kernel density estimate, only for grid cells where TOAR-II measurements exist.

Line 317: Fig. 3 presents the distribution of population exposure calculated from six datasets and the gridded TOAR-II observations in three world regions with a high density of observations, for 2016. Here we calculate the population-weighted kernel density for population exposure to OSDMA8 concentrations, based on the 0.1° x 0.1° resolution for each region, only for grid cells where the re-gridded TOAR-II data have a value. Corresponding plots for other years (2006 to

2015) are shown in Figure S6. Overall, the datasets are widely distributed, and the estimated exposure peaks vary. In East Asia (EAS), the population is exposed to high ozone concentrations. The concentration distribution is broad and has multiple peaks from TOAR-II observations, indicating a complex pollution environment, with a large population exposed to concentrations frequently exceeding 50 ppb, even 70 ppb. BME and NJML show a similar distribution as TOAR-II. Significant differences exist between UKML, CAMS and GEOS-Chem with the TOAR-II data for EAS. In Europe (EUR), exposure is concentrated between 40 and 50 ppb, indicating a more moderate and uniform exposure. The BME and CAMS have the best fit with the TOAR-II. NJML, UKML, GEOS-Chem, and TCR-2 show a peak at a higher ozone concentration range of 50–60 ppb. In North America (NAM), exposure peaks sharply in the 40 to 50 ppb range, which is slightly higher and more concentrated than in Europe. The NJML dataset agrees best with the shape of the TOAR-II distribution, and GEOS-Chem and BME capture the overall shape of the major exposure peaks well.

Line 440: Populations in regions such as East Asia and South Asia appear to be exposed to higher ozone concentrations in all datasets compared to other regions, which supports our findings from exposure based on TOAR-II observations in Fig. 3.

3. The numbering of the supplementary table (Table S) is not presented in sequential order, which can cause confusion. Ensuring correct and consistent numbering of all tables and figures is a basic requirement that should be addressed prior to submission. Please revise the manuscript carefully to ensure that all Tables (S) and Figures (S) are accurate and properly ordered.

Response:

We thoroughly reviewed the supplementary information, and ensure all tables and figures are now numbered accurately and appear in sequential order in main paper.

4. Units are missing in Table 2. For clarity and scientific accuracy, please add the relevant units to all applicable columns or data entries.

Response:

We thank the reviewer for catching this oversight. We have now added the appropriate units (ppb/year) for the trend slope and its confidence interval. Please note that this table is now Table 3 in the revised manuscript.

5. Even in the revised version, some statements still lack objective explanations based on consistent criteria. For example, the manuscript highlights a downward trend only for NJML (in Line 264); however, BME also shows a concurrent downward trend in both area- and population-weighted metrics. If the authors interpret the populated-weighted trend of BME (i.e., -0.04) as

insufficient to indicate a clear trend, then by the same standard, none of the remaining five datasets show an increasing trend either. In this context, the statement in line 415 needs to be revised. More importantly, the manuscript lacks a clear definition of what constitutes a "clear" or "unclear" trend. The criteria used to make such classifications should be explicitly stated—whether based on slope magnitude, statistical significance, or another method. Without a consistent and objective basis for trend interpretation, the analysis risks appearing arbitrary and subjective. Please review the manuscript carefully to ensure that all claims are supported by clear, objective, and consistently applied analytical reasoning. In addition, the authors are advised to carefully review all statements throughout the manuscript.

Response:

We agree that our initial manuscript lacked a clear and consistent definition for interpreting trends, which led to the issues you identified. To fix this issue, we explicitly define our criteria for trend and certainty in the Methods section (section 3). The interpretation is based on the statistical significance (p-value) of the trend's slope, categorized by levels of certainty (e.g., very high, high, low, etc.).

Line 255: We calculated the yearly ozone trend using 50% quantile regression for each dataset using both population-weighted and area-weighted approaches, with details of the calculation methods provided in Text S2. In this study, the trend is interpreted from the slope of the quantile regression, and confidence in the trend is determined by its p-value: $p \leqslant 0.01$ is considered very high certainty; $0.01 < p \leqslant 0.05$, high certainty; $0.05 < p \leqslant 0.1$, medium certainty; $0.1 < p \leqslant 0.33$, low certainty; and $p > 0.33$, no evidence.

Line 266: To quantify the uncertainty in our exposure analysis, we established lower and upper bounds for all population exposure and share of population estimates. The OSDMA8 95% confidence interval (CI) for each dataset is determined through a grid-to-grid linear regression between each dataset and the re-gridded TOAR-II observations based on 0.1°× 0.1°grid cells.

*Using these objective criteria, we have carefully revised all trend-related statements throughout the manuscript for consistency. For the specific example you noted, the text now clarifies that while both **NJML** and **BME** show downward trends, they do so with different levels of statistical confidence ('very high certainty' vs. 'low certainty').*

The text is revised as follows:

Line 382: In Table 3, focusing on 2006 to 2016, we find that NJML was the only dataset to exhibit a downward trend with very high certainty for both area- and population-weighted mean ozone concentrations. In contrast, TCR-2 and UKML only show increasing trends in population-weighted mean ozone during this period with very high certainty. However, while the BME

dataset shows a negative slope for the area-weighted mean, this downward trend has only low certainty; for the population-weighted mean, there is no evidence of a decreasing trend.

Line 477: Despite this, the three chemical reanalysis datasets unexpectedly outperform the machine learning datasets in R^2 (TCR-2, GEOS-Chem) and in RMSE (CAMS) over the full year 2016.

Line 493: NJML demonstrates a very high certainty decreasing trend in global population-weighted and area-weighted yearly mean over the 2006-2016 period. While TCR-2 and UKML exhibit very high certainty increasing trends in global population-weighted mean which relates to their overestimation.

6. (Sect 3.3, Line 224-225) Regarding 6th comment in the 1st round, the concern is not about missing data. Of course, reanalysis datasets like GEOS-CHEM, CAMS, and TCR-2 have complete spatial coverage. The issue is the lack of spatial representativeness when comparing coarse-resolution grid cells with single (or multiple) observation sites. A single or (multiple) monitoring station(s) may not adequately represent the entire of a coarse grid cell. The authors need to justify how the address this potential mismatch in spatial representativeness.

Response:

We thank the reviewer #2 for pointing out this concern regarding the potential mismatch in spatial representativeness. We agree this is a challenge when evaluating coarse-resolution datasets against observations. To address this issue, we have implemented two evaluation scenarios in our revised manuscript.

1. *Grid-to-Grid evaluation: We have re-gridded the TOAR-II observations onto the native grid resolution of each dataset. In doing so we use methods similar to those of Schnell et al. (2015), but here we gridded observations for the whole TOAR-II dataset globally and for more years than had been done previously. This method addresses the representativeness issue by comparing the value of re-gridded observational grid cell to the value of dataset's grid cell for the same spatial resolution.*

2. *Grid-to-Point evaluation: This is the traditional method used in the original manuscript where the value of the dataset's grid cell is compared to all observations within that cell. This method can ensure all evaluations are the same sample size given by the number of observations.*

By presenting results from both methods in section 4.1, we provide a more robust and comprehensive assessment of dataset performance. We have revised as follows:

Line 214: Considering that the six datasets have different resolutions and are designed for different applications, we adopted a dual evaluation strategy to provide a comprehensive assessment of their performance. The first method is a grid-to-grid evaluation. Similar to the approach of Schnell et al. (2015), we re-gridded TOAR-II observations to a 0.1° x 0.1° resolution by an inverse distance weighted method and then aggregated them to match the native resolution of each of the six datasets. In this approach, the sample size for each evaluation varies reflecting the varying resolution of the datasets; for 2016, BME had 173,718 grid cell pairs, NJML had 7,099, UKML had 162,419, CAMS had 4,614, GEOS-Chem had 782, and TCR-2 had 2,195. We also adopted the grid-to-grid evaluation method for regional evaluations, as it provides better spatial representativeness over large areas. To quantify the uncertainty of the six datasets' estimates, we determined the lower and upper bounds (95% confidence interval), derived from the grid-to-grid regression analysis performed between the TOAR-II observations and each of the six datasets at their native resolutions.

By presenting results from both methods in section 4.1, we provide a more robust and comprehensive assessment of dataset performance. Overall, our main conclusions and the relative performance rankings of the datasets remain largely consistent across both evaluation methods. As expected, the grid-to-grid approach generally results in lower RMSE and higher R^2 values, as it averages out localized errors that are more prominent in the direct grid-to-point evaluation. However, the difference of two methods does not change the fundamental takeaways of our analysis.

We have updated the results throughout the manuscript to describe the evaluation from both the grid-to-grid and grid-to-point methods. The following are some key text from the manuscript, updated to reflect the revisions:

Line 275: For 2016, BME outperforms other datasets in both evaluation method, with the highest R^2 (0.75 for grid-to-grid, 0.63 for grid-to-point) and lowest RMSE (4.25 ppb for grid-to-grid, 5.28 ppb for grid-to-point), its density cores intersecting the y=x line

Line 280: In Fig. 1(a), all three chemical reanalysis datasets exhibit a moderate R^2 ranging from 0.51 to 0.60 for grid-to-grid and 0.35 to 0.41 for grid-to-point, comparable to the performance of the machine learning datasets, which have R^2 values of 0.50 and 0.56 for grid-to-grid, 0.37 and 0.38 for grid-to-point. Among these five datasets, CAMS has the lowest RMSE (6.00 ppb for grid-to-grid and 7.59 ppb for grid to point), which is better than other chemistry reanalysis products but relatively low R^2 (0.51 for grid-to-grid and 0.35 for grid-to-point). Its density cores slightly below the y=x line suggests CAMS estimates are marginally lower than TOAR-II observations. GEOS-Chem and TCR-2 demonstrate adequate performance, albeit with higher RMSE values of 8.47 ppb and 10.26 ppb for grid-to-grid, 10.27 ppb and 13.23 ppb for grid-to-point, respectively.

Line 362: In grid-to-grid evaluation, GEOS-Chem shows an overall better performance in R^2 than CAMS, TCR-2 and UKML.

Line 369: From 2006 to 2016, the performance rankings derived from R^2 values varied significantly between the two evaluation scenarios, whereas the RMSE based rankings were nearly consistent.

Line 550: In instances of missing model estimates, we default to the nearest valid estimate to evaluate with TOAR-II observations or re-gridded grid cell. For datasets with coarse spatial resolution, this method may increase or reduce bias by double counting.

7. The amount of supplementary material is excessive and may overwhelm the reader. The authors are encouraged to condense and streamline the supplementary information by summarizing key findings and removing redundancies. Presenting only the most relevant and necessary content will improve the clarity and accessibility of the supporting materials. In particular, Figure S4 does not appear to add meaningful value.

Response:

We have revised the supplement to remove some tables and figures that do not add much meaningful value. The specific tables and figures removed are as follows, based on their original numbering:

Table S12. Number of NA value at all TOAR-II stations of six datasets in 2016.

Table S13. Yearly trends of area weighted, and population weighted global mean of ground-level ozone for six datasets with 95% confidence intervals and p-values over full time period of each dataset.

Figure S4. Population weighted ozone (OSMDA8) trends per decade for six datasets, calculated over the full period of each dataset.

Figure S8. Heatmaps of pairwise correlation (Pearson R) between each dataset from 2006 to 2016.

Figure S9. Heatmaps of pairwise Root mean square difference (RMSD) between each dataset from 2006 to 2016.

Figure S10. Population-weighted ozone (OSMDA8) for each year from 2006 to 2016 in different regions. The horizontal axis represents ozone exposure concentrations, and the vertical axis represents population size.

Figure S12. Performance evaluations of six datasets with TOAR-II observations for OSDMA8 for each year from 2006 to 2015. The evaluation only includes monitor stations above 50ppb in TOAR-II network for each year (2006 to 2015).