We thank the reviewer for their attention to this manuscript and thoughtful comments.

L23: Given the large bias errors in these data sets, comparing the population exposed to a threshold value, like 50 ppb is meaningless. Do these differences impact "health" as stated or one's analysis of health effects. This is a bit sloppy writing.

Response:

We selected the 50 ppb threshold because it corresponds to the long-term air quality interim target established by the WHO. Our analysis focused on comparing the six datasets to understand their differences. It is important to evaluate whether the large differences between ozone products result in different conclusions. We add more discussion in section 6 (Discussion), that biases in the datasets will affect assessments like the population exposure above 50 ppb. Then there is a need for future work that will reduce the bias of ozone products.

Revised:

Line 25: "These differences are large enough to impact assessments of health impacts and other applications."

Line 109: "OSDMA8 is GBD's ozone metric for quantifying health effect from long-term ozone exposure (Brauer et al., 2024), and it is the metric used in the World Health Organization's air quality guidelines, with values of 30 ppb for the guideline and 50 ppb for the interim target (World-Health-Organization, 2021)."

Line 251: "We selected the 50 ppb as the threshold for high ozone concentration because it corresponds to the long-term air quality interim target of WHO."

Line 454: "The performance of each dataset can impact the accuracy of trend analysis (Fig. 1 and Fig. 2) and population exposure assessment (Fig. 6), which may lead to very different results when compared to the WHO guideline and interim target."

L25: very good point, but you should compare the gridded data you have with Schnell's gridding of the EU and NAm. Comparing points to grid-cell averages that you have from the global data sets is a serious science problem -? not the way you treat it here.

How can you get an R2 for surface sites vs grid-cell means?? This is not sensible.

Response:

Thank you for raising this important point. We acknowledge your concern regarding the comparison of grid-cell averages to point measurements. However, the evaluation approach we employed, using grid-cell averages of model output to evaluate model performance at point locations of observations, is widely used for evaluation. The goal of *our work (Section 5) is to assess the accuracy of gridding products in estimating measured ground concentration point values.*

Schnell et al.'s approach of creating 1º×1º grid-cell averages of TOAR observational data is valuable and effective for regions with dense monitoring networks, such as Europe and North America. However, there are several reasons to not use it in our study. First, we specifically focus on evaluating global ozone mapping products against the most recent TOAR-II observations. However, Schnell's dataset is inadequate to evaluate globally, and it was created using TOAR-I data. Second, most datasets included in our comparison have finer spatial resolutions than 1º×1º.

In this context, the grid-cell-average-to-station-point comparison represents an accepted method. We explicitly acknowledge the limitations of this approach in our manuscript and clarify that the performance metrics, including R^2 values, should be interpreted considering this spatial representativeness uncertainty.

Revised:
Line 252: "These performance metrics should be interpreted considering the spatial representativeness uncertainty that is caused by the grid-to-point evaluation approach."

L27: You just said your data is worst at overestimating O3 at low abundances, but here you say it is worse for >50 ppb??

Response:

This decline in performance at higher ozone concentrations (>50 ppb) arises not primarily from increased overestimation but rather from reduced agreement between modeled and observed ozone distribution at the higher ozone concentration (>50 ppb).

Thus, there is no contradiction: the datasets typically overestimate ozone concentrations at lower observed levels, but the R^2 deteriorates more significantly at higher ozone concentrations due to increased uncertainty and reduced agreement at these more extreme conditions. We clarify this distinction more explicitly in the revised manuscript.

Revised:

Line 25: "Comparing with Tropospheric Ozone Assessment Report (TOAR) II ground-level observations, most datasets overestimate ozone, particularly at lower observed concentrations. In 2016, across all stations, R^2 ranges among the six datasets from 0.35 to 0.63, and RMSE from 5.28 to 13.49 ppb. Agreement between modeled and observed ozone distributions is reduced at ozone concentrations above 50 ppb."

L29: "highlighting the importance of continued research on global ozone distributions"

Response:

The referee refers to the original text without comment. The large discrepancies found among datasets suggest that it is important to continue research on global ozone distributions through more widespread measurements, improved modeled estimates, etc. We retain the original text here.

L38: Oh really. The number of regions could be much much greater if you picked smaller regions. The key issue is the area fraction NOT the number of regions.

Response:

Thank you for highlighting the importance regarding area fraction. Gaudel et al. find that ozone is increasing over all 11 NH regions that they defined and analyzed, and they did not find decreasing or flat trends in any region.

Revised:

Line 40: "Gaudel et al. find that since the mid-1990s, tropospheric ozone above the surface has increased across all 11 study regions in the Northern Hemisphere that they defined and analyzed (Western North America, Eastern North America, Southeast North America, Northern South America, Northeast China/Korea, The Persian Gulf, India, Southeast Asia, Malaysia/Indonesia, Europe, Gulf of Guinea) (Gaudel et al., 2020)."

L42: 30 ppb is basically the minimum background level – this is not a useful statement and it implies that pollution is the cause here.

Response:

Thank you for this clarification. We show results for population exposure above 30 ppb because this is the WHO air quality guideline. We agree that this guideline is near the background ozone level, although typical estimates of preindustrial (without human influence) ozone are lower. The intention here was not to imply that ozone near 30 ppb results from pollution.

L44: The quality of writing (logic, not English) is poor: You just quoted all these results that rely on estimates of surface ozone and then you say you lack knowledge of surface ozone.

Response:

Thank you for pointing out the potential confusion. We recognize how the final sentence might seem logically inconsistent with the preceding context. Our intention was not to suggest that no knowledge exists regarding surface ozone concentrations; rather, we intended to highlight that despite existing assessments, substantial uncertainties remain *due to observational gaps, especially in remote and developing regions. The paragraph that follows this one focuses on recent research on ozone mapping products. We revise the last sentence of this paragraph to clarify this point explicitly.*

Revised:
Line 48: "Despite existing assessments, substantial uncertainties remain due to observational gaps, especially in remote and developing regions."

L70: Great. This is the most important statement. Could be up front

Response:

Thank you for this comment. We agree that this statement is important, and we move it up to the end of second paragraph of the Introduction.

L71: "Potential" – there are most assuredly inconsistencies.

Response:

Thank you for highlighting this point. We acknowledge that inconsistencies among ozone datasets assuredly exist. Our original use of "potential" was too cautious. To reflect this clearly, we delete the "Potential".

Revise:

Line 77: "Inconsistencies in these datasets could significantly impact public health research, especially in assessing the risks of ozone-related health impacts, and may impede the development of effective environmental policies and ozone management strategies."

L75: the biases and errors certainly come from the process. I hope you are not using 'data' to describe the assimilation modeling here.

Response:

Thank you for highlighting this point. We fully agree that biases and errors primarily arise from the processes. In our original phrasing, "data" referred specifically to the "input data" utilized by each ozone mapping product. However, we recognize that the input data themselves can also contribute to these biases and errors. In the subsequent discussion, we explicitly note that chemical reanalysis products are constrained by limitations in satellite observations, while machine learning and geostatistical methods are constrained by the spatial distribution of ground-level monitoring stations.

Revised:
Line 79: "Although each dataset incorporates a considerable amount of observational

information and model simulations through various methodologies, each inherently incorporates biases from these input data sources during the fusion processes. "

L85-95: This exposes the fundamental flaws in the focus of this paper. The use of OSDMA8 totally washes out the key fundamental information about ozone that can be tested with the real surface direct observations. The 24-hour diel cycle is a must that needs to be simulated in any modeled ozone product (all of your six sets are modeled products). Likewise the variability of ozone (including MDA8) is critical in evaluating health/agric. impacts and there needs to be a test of your six 'sets' as to their ability of match extremes.

Response:

This is a very good point. We agree that although the OSDMA8 metric fulfills specific needs for many scientists, regulators, epidemiologists, and policymakers, it certainly is not the only metric of interest. As you have pointed out, metrics capturing the full 24-hour diel cycle of ozone are essential for robust evaluation and validation of ozone from chemical transport models or other models. However, our study is not intended to perform a model evaluation as one would do for a chemical transport model. Rather, we focus here on intercomparison and model evaluation for a single yearly metric (OSDMA8) that is important as a metric adopted by the WHO air quality guidelines, and by the Global Burden of Disease studies. Doing so is necessary because the UKML and NJML datasets estimate monthly average DMA8, and not a finer temporal resolution, and the BME dataset estimates OSDMA8. The chemical renalyses estimate ozone at finer timescales (Table 1), and they have been evaluated comprehensively with respect to observations previously ((Miyazaki et al., 2024; Sekiya et al., 2024; Jones et al., 2024)). We clarify this explicitly in introduction of revised manuscript.

Revised:

Line 95: "Our study specifically utilizes the OSDMA8 metric because we focus on evaluating long-term ozone exposure, an aspect not comprehensively compared previously among global ozone mapping products."

Line 155: "Detailed comparisons of these reanalyses for ozone over the entire troposphere at finer timescales have been conducted by the TOAR-II chemical reanalysis working group (Sekiya et al., 2024; Jones et al., 2024; Miyazaki et al., 2024), but without a focus on the ground level and long-term metric as analyzed here."

Line 210: "The OSDMA8 metric is used for long-term ozone exposure given its utility and wide acceptance in health impact studies, despite the inherent loss of shorter temporal dynamics."

L195ff: ibid. This is a mistake to smooth out the fundamental ozone cycles (diel and synoptic).

Response:

Please see the response to the previous comment.

L220ff: "We adopted a point-to-grid 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." This seems to ignore the previous TOAR-related work by Schnell where for the high-density of surface sites in EU and N.Am., a 1ºx1º grid-cell averaged, hourly surface ozone product was created.

This data set was used to assess extremes and to test the CMIP model's accuracy in seasonal and diel cycle of ozone. The cell average is the only way to do a fair comparison with the surface sites because of their irregular – sometimes oversampling and sometimes under sampling – many regions. Comparing surface sites with model cells is dangerous, especially since in this paper their appears to be a lack of understanding of the problems with this approach. The Schnell data are the obvious choice to validate your six model-data sets, even if it is only for EU and NAm:

doi:10.5194/acp-14-7721-2014

doi:10.5194/acp-15-10581-2015

doi:10.1002/2016GL068060

doi:10.1002/2017GL073044

doi: 10.1073/pnas.1614453114.

Then you can go after the rest of the world (which is very important).

Response:

Thank you for emphasizing this important point. We recognize the significance and validity of the Schnell et al. dataset, which provides 1º×1º grid-cell-averaged hourly ozone data, particularly suitable for analyzing extremes and validating seasonal and diel ozone cycles. However, the goal of our work is to assess the accuracy of gridding products in estimating measured ground concentration point values. The reason for doing this is because the *point value is of interest in some applications. For example, exposure scientists are frequently concerned with assessing ozone exposure at specific locations or points. While point values are available at monitoring sites, they are not available away from monitoring sites. Thus, while Schnell's dataset effectively addresses the challenge of spatial representativeness by providing grid-cell ozone values, our focus is explicitly on assessing whether global ozone mapping products can reasonably estimate point-level concentrations at locations lacking monitoring stations. 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. We explicitly acknowledge and clarify this in our methodology part of the revised manuscript. We have also included a Table (S12) that lists the NA values for each dataset. The chemical reanalysis datasets at coarse resolution have no NA values. For the other datasets at finer resolution, NA values are mainly along coasts and in the large majority of cases where an NA value exists, an adjacent grid cell is selected for comparison with observations.*

Revised:

Line 239: "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)."

Line 243: " 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."

Line 249: " We assessed the performance of each dataset using the coefficient of determination (R^2) between ozone estimates and observations, and root mean square error (RMSE) as the primary metrics. We selected the 50 ppb as the threshold for high ozone concentration because it corresponds to the long-term air quality interim target of WHO. These performance metrics should be interpreted considering the spatial representativeness uncertainty which is caused by the grid-to-point evaluation approach."

This paper is based on comparing 6 different modeled surface ozone dataset with one another and with the TOAR set of surface sites (Table 3). The comparison of individual sites with grid-cell averages fails to recognize the difficulty of the task and ignores the extensive efforts to develop unbiased grid-cell means from high-density observations. The authors further corrupt the data set by averaging and smoothing to destroy the fundamental information on ozone variability that is critical for testing the modeled ozone products. The use of these 6 sets, varying in resolution from 0.1 to 2.5 degrees, to map population exposure is premature.

I can not recommend publication of this work as is.

Response:

Thank you for your thoughtful comments. In general, we focus here on an annual metric that is recognized as important for health, which all 6 datasets estimate, whereas the machine learning and BME datasets do not provide estimates at finer temporal resolution. Evaluating fine temporal resolution (daily cycle) is important for typical model evaluations, but that is not the purpose of this study. While the work of Schnell et al is valuable, we do not think it provides a stronger basis for model evaluation than the comparison of individual monitoring sites with grid cell averages from the ozone mapping products, which is a standard method used in our field. Our work is valuable in showing that current products using different methods of estimating ground-level ozone differ from one another and vary in performance against observations, suggesting that further work to better constrain ground level ozone remains important.

Reference

Jones, D. B. A., Prates, L., Qu, Z., Cheng, W. Y. Y., Miyazaki, K., Inness, A., Kumar, R., Tang, X., Worden, H., Koren, G., and Huijnen, V.: Assessment of regional and interannual variations in tropospheric ozone in chemical reanalyses, 2024.

Miyazaki, K., Bowman, K., Marchetti, Y., Montgomery, J., and Lu, S.: Drivers of regional surface ozone bias drivers in chemical reanalysis air quality revealed by explainable machine learning, 2024.

Sekiya, T., Emili, E., Miyazaki, K., Inness, A., Qu, Z., Pierce, R. B., Jones, D., Worden, H., Cheng, W. Y., and Huijnen, V.: Assessing the relative impacts of satellite ozone and its precursor observations to improve global tropospheric ozone analysis using multiple chemical reanalysis systems, EGUsphere, 2024, 1-35, 2024.