Review of the manuscript of egusphere-2024-3723 titled “Intercomparison of global ground-level ozone datasets for health-relevant metrics” written by Wang et al.

This study conducted a variety of analyses, including assessing ozone-exposure populations using extensive reanalysis and AI-derived ozone concentration data. While the analysis method itself is not entirely novel, the study is meaningful in its comparison of AI-based data with chemical reanalysis data. However, the authors have some issues that require improvement in the manuscript for publication. The following are the reviewer’s concerns:

Response:
We thank the reviewer for their careful reading of the manuscript and thoughtful comments.

Major comments

1. Correct trend calculation and null hypothesis: Trends can vary depending on the selected time range. For instance, as shown in the figure below (Fig. R1), when restricting to the time range of GEOS data, the trends of BME and CAMS seem to be stagnant or declined, unlike those described in the manuscript. Consequently, if this time range is not properly justified, the calculated one itself may be questionable. Therefore, I strongly recommend that the authors provide a clear reason for selecting the different time ranges used in the trend calculation and assess its statistical significance.

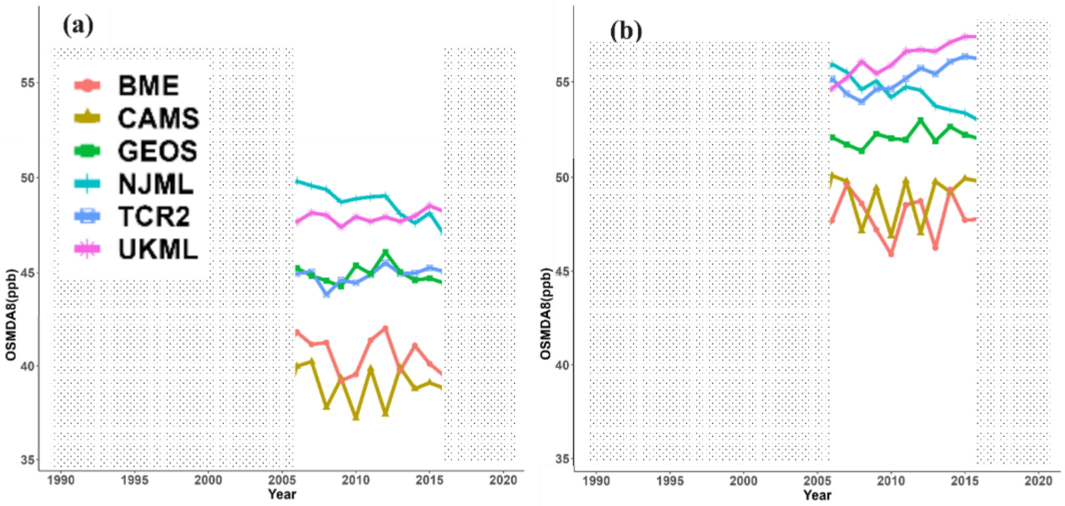

Fig. R1. Six trends of OSMDA8 modified from Figure 1 in the manuscript.

Response:
Thank you for this comment. We agree with your points that the selection of the time range can significantly influence trend calculations, and that choosing a uniform time range over all datasets gives the most consistent basis for comparison. We added a table (Table 2) that limits analysis of trends to 2006 to 2016, showing the area-weighted trend and population-weighted trend for six datasets with 95% UI in the main manuscript. We also maintain Figure 1 because it remains valuable for illustrating the trends over each dataset’s full coverage period. And we add a table (Table S13) for the full time period with 95% UI in the SI. The text is revised in some places to discuss results when focusing on 2006-2016.
Revised:

Line 263: “In Table 2, focusing on the period from 2006 to 2016, we find that NJML is the only dataset showing a downward trend in both area-weighted and population-weighted mean *ozone concentrations, with very high certainty. In contrast, TCR-2 and UKML show increasing trends in population-weighted mean ozone during this period with very high certainty."*

Line 415: "NJML demonstrates a decreasing trend in global population-weighted and area-weighted yearly mean over the 2006-2016 period, while the five others exhibit either increasing trends or no clear trend."

Line 498: " In terms of long-term trends over 2006 to 2016 period, UKML and TCR-2 show a consistent upward trend globally, while NJML shows a downward trend."

2. Figure 2: Regarding the first comment, the comparison among the six datasets in Figure 2 (and discussion in Section 4.1) is meaningless since their temporal ranges are different.

Response:

In the original draft, Figure S4 shows the regional trend from 2006 to 2016. We now move that figure to the main body to replace Figure 2 and move the original Figure 2 to Figure S4 with a note that time periods are inconsistent. We also add a table in SI (Table S11) to show the trend of six datasets in each region from 2006 to 2016, with 95% UI.

Revised:

Line 22: "For example, in Europe, the two chemical reanalyses show an increasing trend while the other datasets show no increase."

Line 275: "From Table S11, we observe that some regions exhibit a clearer trend from 2006 to 2016, with very high certainty across six datasets. In East Asia, BME and NJML observe decreasing trends, whereas the other 4 datasets display increasing trends. In North America, all datasets display a downward trend, and in Europe, BME, NJML, UKML and TCR-2 show a decline, contrasting with increases in CAMS and GEOS-chem. Recent analyses using TOAR observations indicate that from 2005 to 2016, most sites in North America experienced decreasing ozone, while many sites in East Asia exhibited significant positive trends."

Line 416: "Divergence among datasets becomes even more evident in the analysis of regional ozone trends (Fig. 2). The ozone concentrations decreased in Europe from 2006 to 2016 according to BME, NJML, UKML, and TCR-2, yet increase in the other chemical reanalysis datasets"

Line 429: "In Fig. 2, all datasets exhibit a downward trend in North America over 2006 to 2016."

Line 499: "Regionally, all datasets show a downward trend in North America, and only BME and NJML datasets demonstrate a downward trend in East Asia; In Europe, BME, UKML, NJML and TCR-2 report a downward trend, while the other two chemical reanalysis datasets reveal an upward trend that is not seen in observations."

3. Impact of data uncertainty on related analysis and reorganizing structure: The accuracy of predicted O_3 concentrations in each dataset significantly affects trend analysis, spatial distribution, and assessments of ozone-exposed populations. The substantial differences in uncertainty among the predicted datasets, as demonstrated through the comparison between TOAR-II observations and various predicted datasets in Figure 7, significantly hamper trend analysis and understanding of the ozone-exposure population. However, this study does not reflect or discuss the uncertainty in the several analyses presented in Section 4. Therefore, I recommend that the authors explicitly address the impact of dataset uncertainty on trend analysis and ozone-exposed population assessments. In addition, to discuss this efficiently, Section 4 and Section 5 should be re-arranged.

Response:

We appreciate your feedback. We agree that the analyses of trends, spatial distributions, and population exposure among the different datasets in Section 4 can be informed by comparisons of each dataset with observations in Section 5. The information in Section 5 provides some guidance as to which dataset in Section 4 is likely to be closer to reality. However, there is also logic in showing how each dataset compares with the others comprehensively, showing differences among datasets and for application to population exposure, before comparing datasets with observations. Therefore, we have chosen to keep the original organization in Sections 4 and 5. Readers who wish to can view the agreement with observations in Section 5 to make their own judgements about the likely veracity of the different datasets shown in Section 4. Then following Sections 4 and 5, causes of uncertainties in the datasets and their relevance for trend analysis and population exposure assessments are discussed in Section 6.

In fact, we attempted to use the biases identified in Section 5 to interpret and discuss the comparative results presented in Section 4; however, we did not find a clear and effective approach. Instead, in Section 6, we explicitly discuss how each model's overestimations and underestimations impact the differences observed in the comparative analyses of Section 4. And we add that uncertainties in each dataset impact the accuracy of trend analyses and population exposure assessments in Section 6.

Revised:

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."

4. Also, regarding the third comment, one idea might be to compare the population exposure to ozone (i.e., Figure 6) calculated based on observations and six analysis datasets for the ozone observational (TOAR-II) sites.

Response:

We appreciate the suggestion to compare population exposure based on observations with that derived from the six analysis datasets. However, directly calculating population ozone exposure from TOAR-II observations is subject to high uncertainty because the monitoring stations are sparsely distributed, and some method would be needed to interpolate between the observations, and this is similar to what the geostatistical and machine learning datasets do.

Additionally, our analysis focused on comparing the six datasets to understand their differences in section 4, and we have thoroughly evaluated the performance of each model against TOAR-II data across different concentrations, regions, and years in section 5.

5-1. Figure 7b. Why is the standard set at 50 ppb? What are the intended messages from the analysis in Figure 7b?

Response:

We selected 50 ppb because it corresponds to the long-term air quality interim target established by the WHO, as stated in the description we added in the Section 2 data part. Figure 7b is intended to demonstrate each dataset's capability to capture ozone concentrations exceeding this ozone level, highlighting their ability to detect years of high ozone. We have revised the discussion of Figure 7b in the second paragraph of Section 5.5 for better clarity.

Revised:

Line 108: "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 358: "Fig. 7(b) focuses only on TOAR-II sites with OSDMA8 value above 50 ppb, showing that R^2 is reduced compared to the comparison of all ozone measurements (Fig. 7(a)) for all six datasets, suggesting overall weaker agreement between modeled and observed ozone distributions at higher concentrations."

Line 361: "However, the change of biases varies among datasets at higher concentrations. Specifically, overestimation is reduced in the UKML, NJML, GEOS-Chem, and TCR-2 datasets when observations exceed 50 ppb. Conversely, we observe increased underestimation in the BME and CAMS datasets at ozone levels above 50 ppb."

5-2. Fig. 7b (and Figure 8). If it is significant that the accuracy of prediction is lowered, particularly over 50 ppb, then how should the results in Figure 7b (or Figure 8) be considered in the analysis of Figure 6? It is also regarding the third comment.

Response:

Yes, we agree with your point that model's accuracy varies at high ozone levels. The comparison in Figure 6 is mainly to address the fact that researchers would typically use any of the six models as the basis for health-related studies on ozone concentrations. We should take into account differences in exposure estimates among the datasets without recalibrating or correcting them.

We have clarified our discussion of Figures 7 and 8 by explicitly noting that the poorer performance of some datasets at higher ozone concentrations will influence the distribution of ozone exposure across the population, as presented in Figure 6.

Revised:

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."

6. Sect. 3.3 (Lines 224-225). For this case mentioned in lines 224-225, the observation data lack representativeness due to the coarse grid resolution in the GEOS-CHEM, CAMS, and TCR-2 datasets. Thus, the authors need to justify it.

Response:

For grid cells with TOAR-II observations, the GEOS-CHEM, CAMS, and TCR-2 reanalysis datasets did not have any missing values. Only the BME, NJML and UKML dataset exhibited some NA values (at finer resolutions). We add a table (S12) detailing the number of NA values and sample sizes for each dataset in SI.

Original text:

For grid cells with a TOAR-II observation but no valid estimate in a dataset (NA value), we used the nearest valid estimate instead.

Revised:

Line 244: "For grid cells with a TOAR-II observation but no valid estimate in a dataset (NA value), we used the nearest valid estimate instead. Table S12 displays the number of missing values in each dataset in 2016 at TOAR-II locations, showing that only BME, NJML and UKML have a small number of missing estimates."

7. L283 - 294. I would like to ask the authors to describe the purpose of separating Groups A and B in Figure 5. Additionally, please specify the criteria used to assign NJML to Group B. If the criterion is a correlation of ~0.83, what is the rationale behind this choice? Why was the RSMD criterion deemed unsuitable? Considering the statement in lines 289-290, the criteria appear to be arbitrary.

Response:

We separated the datasets into Groups A and B to compare their spatial distribution patterns of ozone estimates. Our grouping method is based on pairwise correlation rather than RMSD because our focus is on spatial similarity, not absolute magnitude differences. Although a correlation value around 0.83 is mentioned, it is not used as a strict criterion. The objective is to ascertain the grouping combination that maximizes the difference between the mean of the within-group correlations and the mean of the out-of-group correlations. The details of the grouping method are described in Text S1. Moreover, even though datasets in Group A show similar spatial distributions, the high RMSD values among them reveal significant differences in the ozone estimates. We add more descriptions of this grouping method in the main manuscript.

Revised:

Line 232: "The idea of this grouping is to distinguish the spatial similarity between the datasets, which is based on the pairwise correlation. For each grouping combination, 4 variables are computed: the sum of pairwise correlations within groups (C_i), the sum of pairwise correlations outside the groups (C_o), the number of dataset pairs within groups (N_i), and the number of dataset pairs outside the groups (N_o). The objective is to ascertain the grouping combination that maximizes the difference between C_i/N_i and C_o/N_o. More details of the calculation can be found in Text S1."

8. L328-338. Some statements lack objective descriptions based on consistent criteria. For example, it is stated that the TCR-2 shows adequate performance, whereas UKML has a significant overestimation. However, both datasets demonstrate similar performance in terms of correlation, RMSE, and slope for each year (refer to the tables below, with values taken from Figures 7 and S11). In fact, the lower slope in TCR-2 indicates a greater overestimation, so the description needs to be corrected.

R^2	**BME**	**NJ**	**UK**	**CAMS**	**GEOS**	**TCR2**
2006	0.62	0.28	0.14	0.19	0.11	0.16
2007	0.68	0.31	0.27	0.31	0.25	0.33
2008	0.66	0.28	0.29	0.27	0.26	0.23
2009	0.59	0.18	0.39	0.23	0.35	0.27
2010	0.52	0.38	0.2	0.17	0.1	0.19
2011	0.6	0.43	0.12	0.33	0.15	0.19

2012	0.59	0.38	0.21	0.25	0.19	0.25
2013	0.51	0.34	0.27	0.29	0.19	0.19
2014	0.53	0.37	0.3	0.29	0.25	0.22
2015	0.58	0.36	0.27	0.24	0.25	0.23
2016	0.63	0.38	0.37	0.35	0.38	0.41
Slope	**BME**	**NJ**	**UK**	**CAMS**	**GEOS**	**TCR2**
2006	0.94	0.54	0.49	0.45	0.48	0.46
2007	0.97	0.61	0.68	0.57	0.66	0.68
2008	0.94	0.62	0.56	0.66	0.64	0.52
2009	0.89	0.52	0.74	0.46	0.7	0.59
2010	0.8	0.62	0.53	0.47	0.4	0.41
2011	0.91	0.65	0.37	0.6	0.53	0.48
2012	0.89	0.69	0.52	0.65	0.52	0.55
2013	0.79	0.68	0.56	0.57	0.47	0.4
2014	0.8	0.73	0.52	0.6	0.54	0.43
2015	0.93	0.75	0.49	0.51	0.54	0.42
2016	0.96	0.80	0.6	0.63	0.72	0.58

RMSE	**BME**	**NJ**	**UK**	**CAMS**	**GEOS**	**TCR2**
2006	4.8	12.2	12.6	8.21	9.3	11.89
2007	4.58	12.17	12.86	7.66	8.74	11.16
2008	4.44	10.84	13.1	8.12	8.48	10.53
2009	4.84	10.72	11.67	8	8.48	11.34
2010	4.93	11.34	13.09	7.54	9.93	12.01
2011	4.63	11.23	14.08	6.53	9.49	12.07
2012	4.72	10.69	13.75	7.55	10.44	11.32
2013	5.07	10.44	12.36	6.48	10.24	12.59
2014	5.26	10.24	13.45	6.23	10.41	12.67
2015	5.53	9.87	14.5	8.61	11.82	14.88
2016	5.28	8.63	13.49	7.59	10.27	13.23

Response:

We acknowledge that our original description was misleading. In fact, TCR-2 indicates a greater overestimation compared to UKML. We revise the manuscript accordingly to provide a more objective description based on performance as shown in Figures 7 and S1.

Revised:

Line 353: "UKML exhibits the highest RMSE of 13.49 ppb, and its density core region is above the y=x dashed line, indicating an overestimation. This is because the UKML algorithm emphasizes higher ozone pollution levels in rural and remote areas compared to adjacent urban districts, which consequently leads to an overestimation especially in population-weighted metrics."

9. L329: I disagree with the characterization of the decreased as "minor". The R^2 value decreased significantly, from 0.63 to 0.51, which cannot be considered minor.

Response:

We change description to "significantly". After re-running the evaluation, the R² improves to 0.53. This time, we excluded all sites located at observation points previously used as BME input. In the initial version of the manuscript, we removed the nearest sites to the BME observations points if they were within a 1-degree radius. Compared to other datasets, 0.53 is still good performance.

Revised:

Line 344: "After excluding all sites located at observation points previously used as BME input, using 3911 observations for validation, BME performs well compared to another datasets, though its R2 decreases significantly to 0.53."

10. L330: The phrase "relatively good" is inappropriate. The performance is not good. It is better described as moderate.

Response:

We agree that "relatively good" overstates the performance. We revise the description to "moderate," which more accurately reflects the performance by the dataset.

Revised:

Line 346: "In Fig. 7(a), all three chemical reanalysis datasets exhibit a moderate R^2 ranging from 0.35 to 0.41, comparable to the performance of the machine learning datasets, which have R^2 values of 0.37 and 0.38."

Minor comments

1. Tables 1 – 6 are not mentioned in the manuscript. The authors need to check the order and ensure proper mention of all tables and figures.

   Response:
   I add the Table numbers when I mention Tables S1-S6 in the main manuscript.

2. L108: Provide an explanation of what M3Fusion is.

   Response:

   M3Fusion is a composite of multiple chemistry models by weighting based on their performance against TOAR observations.

   Revised:

   Line 114: "M^{3}Fusion (Measurement and Multi-Model Fusion) is a statistical method developed to improve estimates of global surface ozone distributions by integrating observational data from TOAR and outputs from multiple chemistry models. Specifically, the method assigns weights to multiple atmospheric chemistry models based on their regional accuracy compared to observed ozone values."

3. OSDMA8 and OSMDA8: These terms are used interchangeably. Check if it is correct, and if not, check the spelling.

   Response:
   We correct OSMDA8 to OSDMA8.

4. In Section 4.1: Clarify what "area-weighted" and "population-weighted" mean or describe how they are calculated. Regarding this in Fig. 1, explain why the population-weighted mean increases more rapidly than the area-weighted one.
   Response:
   We add explain the potential reason that lead to rapidly increases in population-weigthed mean. We add the explanation of "area-weighted" and "population-weighted" in Text S2 in the SI with the calculation methods.

   Revised:

   Line 216: "We calculated the yearly ozone trend for each dataset using both population-weighted and area-weighted approaches, with details of the calculation methods provided in Text S2."

   Line 263: "The faster increase in the population-weighted mean compared to the area-weighted mean appears to be driven by rising ozone levels in highly populated regions."

5. Y-axis in Figure 1: To avoid confusion, make the y-axis the same.
   Response:
   We change the Y-axis to the same.

6. L269: Modify the phrase to "in the multi-model average over 50 ppb" in Line 269. Remove a dot before the 'over'.

   Response:
   We remove the dot and revised the phrase.

   Revised:
   Line 284: India, China, and the Middle East are estimated to have the world's highest average ozone concentrations, exceeding 50 ppb in the multi-model average.

7. Figures 7 and S11 – S13: The observation-prediction data points are shown in blue,

which can be confused as indicating density. Thus, it would be better to change their color to black or gray for clarity.

Response:

We used blue color to distinguish the y=x line from the regression line. We have changed the data points to grey color.

8. Colors in Figures 1 and S3 (and Figures 8 and S14): To reduce confusion, use consistent color for each dataset across the figures.

   Response:
   We have changed to use the same color.

9. L325: It seems that Figure S7 is mistakenly referenced and should be corrected to Figure S11.

   Response:
   We change it to S11.

10. Significant digits in Figures 7 and S11 – S12: Ensure that significant digits are presented consistently.

    Response:

    We changed the significant digits to be consistent for Figures 7 and S11 – S12.