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:

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.

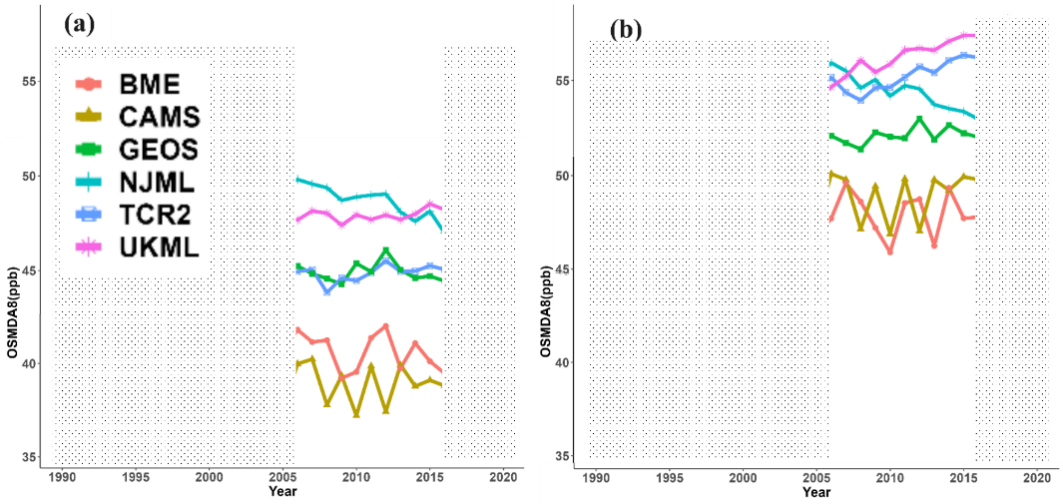

Fig. R1. Six trends of OSMDA8 modified from Figure 1 in the manuscript.

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.

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.

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.

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

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.

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.

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.

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

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.

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

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.
2. L108: Provide an explanation of what M3Fusion is.
3. OSDMA8 and OSMDA8: These terms are used interchangeably. Check if it is correct, and if not, check the spelling.
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.
5. Y-axis in Figure 1: To avoid confusion, make the y-axis 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’.
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.
8. Colors in Figures 1 and S3 (and Figures 8 and S14): To reduce confusion, use consistent color for each dataset across the figures.
9. L325: It seems that Figure S7 is mistakenly referenced and should be corrected to Figure S11.

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