Dear authors,

as a representative for the ozonesonde network, I'm really grateful for your effort in filtering and cleaning up the ozonesonde data, which are available in three (major) existing international archives, but in different formats and not necessarily identical. The Unified Ozonesonde Dataset is certainly a step in the right direction to ease data access and data use!

As for every ground-based (or in-situ) dataset, one of the major challenges of the ozonesonde data is the spatial and temporal representativeness for measuring (here) stratospheric ozone. In this study, an attempt has been done to analyze its impact on stratospheric ozone trends by considering two different subsets of the Unified Ozonesonde Dataset: the "Large coverage" and Large and medium coverage" datasets, and this for different latitude bands.

General comments

The authors should also give some remaining caveats associated with the Unified Ozonesonde Dataset:

- It should be clearly mentioned that "unified" is a different concept than "homogeneous". As you know, the ozonesonde community has been undertaken a homogenization activity to process all the ozonesonde data across the world according to the same principles (e.g. removal of known biases by referring all data to the same standards, consistent correction procedures for known effects, uncertainty calculation). These homogenized data are stored on a ftp-server with access details provided on the HEGIFTOM website, and also on SHADOZ, and for a handful of sites only in WOUDC or NDACC.
- As a consequence of the previous point, the Unified Ozonesonde Dataset will merge data at a given site that have been processed differently (e.g. homogenized in SHADOZ vs. non-homogenized in WOUDC or NDACC; the Payerne data has been corrected for total ozone normalization in one archive, but not in the other archive), which can give rise to biases between data points. This should be mentioned.

Although the primary focus and expertise of the authors lie in the data polishing to construct the Unified Ozonesonde Dataset, the authors also want to contribute to the calculation of stratospheric ozone trends from this dataset. They apply common approaches or assumptions, but no justification or interpretation is given why

- two different trend periods are considered (1978-1999, 2000-2022)
- a linear regression trend model is used
- zonal trends are considered (i.e. different trends for tropics, mid-latitudes, polar regions)
- different vertical trends are calculated (i.e. between 300-200, 200-100, 100-50, and 50-1 hPa)

The authors should give an interpretation or explanation why trends differ (or not) between the two periods, between the different latitudinal zones, and between the different vertical ranges. The latest WMO Scientific Assessment of Ozone Depletion can be used as guidance here (https://ozone.unep.org/sites/default/files/2023-02/Scientific-Assessment-of-Ozone-Depletion-2022.pdf).

The assessment of the representativeness of the unified stations in terms of representing the total ozone column field (Fig. 9) is a nice feature, but also raises some important questions:

- what is the consequence of these correlations for the representativeness of the LMC and LC datasets for the ozone concentrations in the vertical ranges considered (300-200, 200-100 hPa, etc)? Should this analysis not be performed for those vertical ranges (e.g. making use of the MLS satellite vertical ozone retrievals) instead?
- Can we draw the conclusion from this graph that LC dataset is not representative for the tropical total ozone band, while the LMC might be?
- Can we draw the conclusion that neither the LMC or the LC datasets are representative for the entire NH mid-latitude total ozone band, and what does this mean for the NH mid-latitude trends calculated from these datasets in the different vertical ranges?

On top of this correlation analysis, to assess the impact that the LMC and LC datasets have on the different vertical ozone trends, it would also be good to show in the manuscript (or provide in the supplementary material) the monthly anomaly time series (and calculated trends) for both datasets (i.e. LMC for Figs. 10-11 and LC for Figs. 12-13), as in the PhD Thesis this manuscript is based on. Those plots would be more illustrative and informative (i.e. better guidance for interpretation) than the tables 7 to 13). Based on those plots in your PhD Thesis, when comparing the ozone time variability (not only the linear trend) in different layers and for different latitude bands (Figs. 32-39), a very different time variability (in terms of features, spread in the variability) arises between the "Large coverage (LC)" and "Large and Medium coverage (LMC)" clusters, illustrating the strong dependence of this time variability on the chosen individual sites in the clusters. The merging of individual sites with different onsets of the ozonesonde observation program (as is done in the comparison between the LC and LMC clusters) gives a different spatial and temporal distribution of ozonesonde sites (or at least a different weighting of the contributions of individual sites) between the beginning and end of the periods for which ozone anomalies/trends have been calculated. This has of course consequences for the calculated trends, but this impact has not been investigated in enough detail.

The discussion of the trend differences (based on the tables 7 to 13) between the different trend estimation tools and the different datasets or vertical ranges really needs the inclusion (in the table and in the interpretation) of the trend uncertainties! I know that the significance of the trend is mainly assessed by using the MK test statistic, but every used linear regression trend estimation tool provides an uncertainty estimation on the slope coefficient (i.e. trend), which should be displayed and used in the discussion between the trend estimates. In the trend discussion, it is also important to note that "no significant trend" can be a physical result, and the presence of a trend does not mean that one dataset is more "reliable", more "robust", or "outperforms" another one. In my specific comments, I'll point to these wordings.

Nowadays, most trend estimation tools for stratospheric ozone trend assessment make use of different proxies for attributing some dynamical behavior (QBO, ENSO, solar cycle, stratospheric AOD) in the ozone time series (e.g. LOTUS multiple linear regression model). In the ozone monthly anomaly time series plots (Figs. 10 to 13), which show a distinct non-linear behavior, only the impact of volcanic eruptions on the ozone data has been discussed. The authors might also discuss the possible impact of e.g. QBO and ENSO

on the tropical ozone time series or, even better, use a multiple linear regression technique to give more credibility to the estimated trends and better enable the interpretation of the long-term time behavior.

At several places, the authors do not give enough details and more clarifications or interpretation is needed. In the specific comments, I'll give an overview.

Specific comments

- Lines 42-43: nowadays, ozonesonde measurements are not normalized to the total ozone column anymore
- In the introduction: mention the existence of the homogenized ozonesonde data (O3S-DQA activity or HEGIFTOM website) and the difference in concept between "unified" and "homogenized".
- Line 100: it is not clear which selection criterion is used here. Please specify "the dataset's maturity and availability of measurement uncertainties on ozone concentration profiles selection criteria"!
- Figures 4 and 5: ozone(sonde) profiles are commonly displayed with the ozone partial pressure (in mPa) on the horizontal axis (x-axis), with the vertical coordinate (pressure here, decreasing in amplitude) on the vertical axis (y-axis).
- Lines 165-170. The criteria for the different datasets (LC, MC, SC) are incomplete and perhaps even not objective. As it is written now, Praha and Macquarie Island sites should be LC stations, not a SC and MC station, respectively. It looks like the criterion is "20 full years" for a SC station, and Macquarie Island is "degraded" to a MC station because it has a 3-month gap in 2003. This rises two questions: (i) how flexible have you been with those criteria for other sites (this is not obvious and some subjective assessment cannot be ruled out → make it objective), (ii) to which extent would a 3-month gap (or larger) has a significant impact on the calculated trend over an at least 20 year time period?
- Figures 7 and 9: as your different latitudinal regions are marked by 30°, use an increment of 30° instead of 25° (Figure 7) or 50° (Figure 9) for the latitude axis.
- Line 219: same comment as Reviewer 1: I don't understand the completeness check criterion, as most ozonesonde sites have weekly launches.
- Table 3: the difference between the flagged NDACC percentages for the ozone partial pressures and ozone concentrations is striking. It is not because the ozone concentrations (in ppmv) are not provided, that they could not be easily calculated from the ozone partial pressures (in Pa). The flagged percentages for those two variables should be very similar, as for SHADOZ. So, what is the meaning (and relevance) of this 86.22% flagged NDACC data for the ozone concentrations?
- Line 229: vertical completeness check. "At least one point every 50 m in an ozone profile is required". With a typical rise rate of 5 m/s and an inherent response time of the ozonesonde of about 18-28s, the effective vertical resolution of the ozone profile is 100-150 m. So, how relevant is the 50m criterion?
- Line 238: Give more details on the single quality flag you generated for each ozone profile: what are the possible values and how have the different quality checks been compiled to obtain one single flag?

- Fig 8: specify that the bad profile in the right panel is the brown one.
- Section 2.5: Several clarifications/additions are needed here. First, clearly mention if the correlations between the total ozone column measurements have been done for deseasonalized monthly values. Secondly, it is also important to add that/if the total ozone column from the ozonesonde profiles have been obtained by integrating the profile, and that the ozone column above the burst altitude is missing w.r.t. the TOMS-EP total ozone measurements. Thirdly, also add how the correlation maps in Fig. 9 have been obtained (you calculate for each grid point the correlation coefficient with each unified station, so which of those correlation coefficients is displayed in Fig. 9? Maximum correlation coefficient, as suggested by Weatherhead et al.?). And finally, mention from which correlation coefficient we can speak of good correlation (according to Weatherhead et al., I think it is higher than 0.7).
- Caption Fig. 9: you should mention that the (linear) Pearson correlation coefficients are calculated between the deseasonalized monthly averages from the total column ozone measurements (please confirm).
- Section 3.3: this section does not provide what is announced in its first sentence (lines 317-318). As far as I understand from this sentence (and from my experience with statistical breakpoint detection tools as SNHT), you compare the individual ozonesonde time series with the corresponding unified ozonesonde time series at the same site and the SNHT test looks for a breakpoint in the difference time series (a significant shift in the mean of the segments before and after the breakpoint). The presence of such breakpoints in the unified vs. individual time series will bring very relevant information. Instead, in Table 5, you provide the mean and maximum values of the SNHT test statistic, but it is far from clear what those values represent and no real interpretation is given. Much more useful information would be the time epochs for which breakpoints in the difference time series are detected. A breakpoint (shift in the mean) in one time series w.r.t. the other would result in trend differences between the two time series, which is very relevant for the remaining of the paper. This section should be seriously rewritten to better reflect its aim, as was outlined in the first sentence.
- Tables 7 to 13: I find it very strange that you include the latitudinal zones SP and SH in your analysis, as you write in lines 201-204: “Due to the limited availability of LC and MC stations in the Southern Polar (SP) and Southern Hemisphere (SH), the analysis presented hereinafter is not conducted on these two regions. Although the related estimated trends are significant, the scarcity of stations’ number and data availability in these sectors substantially enhances the uncertainties on the calculated trends, making them unreliable.” This is really a contradiction with including the trends of these regions in the tables. Therefore: drop those zones in the tables and discussion, you gave the arguments for doing so.
- Tables 9, 10, 11, 12, 13: add the (two-sigma) uncertainties of the trend estimations, also in the discussion!
- Line 356: argue why you consider two different periods 1978-1999 and 2000-2022 for calculating trends.
- Line 357: in your analysis, you do not explicitly discriminate between geographical and temporal sampling impacts on trends, so add “temporal” here.

- Line 367-368: as you do not consider the MC cluster trends (time series not long enough, see line 172), this statement cannot be verified.
- Line 371: I don't have a clue what is meant with "In the TR, results resemble those of the SP."
- Line 372 "the LC lacks significant trends in certain pressure ranges due to the limited number of stations", Line 375 "LC showing slightly better performance via the MK test", Line 381 "show comparable abilities to estimate significant changes", Line 418 "This performance improvement", Line 425 "trend estimates are considered unreliable", Line 429 "Notably, the performance of the LMC cluster in the TR sector outperforms that of the LC cluster in this layer", Line 440 "providing a more robust dataset for trend estimation, …: all these statements assume that no trend means a worse performance of a dataset for trend calculation, and do not take into account that no significant trend can be a true, physical result for some zones, vertical ranges or time periods. The entire discussion in this section should be realigned in this sense. As mentioned earlier, the comparison of the monthly anomaly time series plots for the two datasets (LC vs. LMC) will reveal more insight whether or not they both display a similar time variability (or trend).
- Line 420-421: what is meant with "enhanced representativeness". This cannot be deduced from the trend analysis presented here, right?
- Line 439-440: "The most suitable vertical ranges for trend analysis are 50-1 hPa and 100-50 hPa due to their richer data content, providing a more robust dataset for trend estimation." It is not the amount of points in the vertical range that plays a role, but the dispersion/variability between the points, because the means are considered to calculate trends. If you want to make a comparison between the different vertical ranges, the standard deviation of the means would be a better metric to compare instead of the number of data points.
- Section 4.1: when comparing your trends with the estimates from other studies, you only consider the trend uncertainties from the other studies; include yours in the comparison as well.
- Section 4.2, lines 502-503: which tropical stations have data before 1998? Mention those!