Comment on acp-2022-252

Robert Damadeo (Referee)

Referee comment on "Stratospheric ozone trends for 1984–2021 in the SAGE II–OSIRIS–SAGE III/ISS composite dataset" by Kristof Bognar et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2022-252-RC2, 2022

Summary:

The authors describe, and to an extent assess, a new version of the OSIRIS ozone data set (v7.2). They then combine this data with ozone data from SAGE II and SAGE III/ISS to create a composite data set to be used for ozone trending analysis. This composite incorporates a sampling correction to attempt to mitigate the potential impacts sampling biases can have on trend analyses. Trend analyses are performed using, and comparing, both MLR and DLM methods with most of the analysis focused on the DLM results. The overall trend patterns are largely consistent with those from other studies over similar time periods. However, the DLM results show a curious second turnaround in the middle-to-upper stratosphere at northern mid-latitudes starting in the mid-to-late 2010s. The authors also demonstrate, through trend analysis in tropopause-relative coordinates, that dynamical effects, particularly a positive trend in tropopause height, are primarily responsible for continuous negative tropical ozone trends. Lastly, the authors consider the potential impact of instrument drifts on resulting trend significances by comparing the data composite to separate measurements from MLS.

Overall this paper is well-written, the work is well-thought out using commonly accepted techniques, and the conclusions are mostly reasonable. However, I do have some general questions and minor comments with respect to the methodology as well as to the overall structure of the paper that I would like to see addressed prior to publication.

General Comments:

The description of changes to the OSIRIS algorithm / data versions feels very disjointed between the text and appendix. It appears that this is the first time the v7.2 data are being used and so I believe the authors intend for this paper to also act as a data reference of sorts. However, the description of changes feels rushed and incomplete. If there's enough material, a separate paper detailing the changes and comparisons for validation purposes would make more sense, though some of the biggest influences seem to be already detailed in other papers so perhaps that is not the case. At the very least (i.e., in lieu of a separate paper), though this is a request outside of this work, I would recommend the authors also put together some sort of release notes with a more singular story regarding improvements (perhaps with some standard comparison validation figures) that could also be referenced for future works. In its current form, this paper does not tell a cohesive story regarding how OSIRIS data has changed between versions and what the impacts to the data quality are.

I think it's worthwhile to change how the QBO is incorporated into this work. The relatively recent series of ozone papers (specifically Ball -> Chipperfield -> Ball) demonstrated that the DLM's adaptable shape is more susceptible to influence from endpoint anomalies caused by potentially improperly capturing natural variability. I am glad to see the inclusion of seasonal cross-terms with the QBO, despite this being outside the original LOTUS framework, as the seasonal cycle modulates the frequency of the QBO, particularly at mid-latitudes. What I'd really like to see though is a change to the number of principal components / EOFs that are used. I know in the past I have recommended using, and this group in particular has used, more than the leading 2 EOFs. Prior to about 2015, this seemed like it may have only yielded a minimal change to the results. However, since that time there have been two major disruptions to the QBO and, perhaps more importantly, the leading 2 EOFs are no longer sufficient to capture all of the variability associated with the QBO. Anstey et al. (2021; DOI: 10.1029/2021GL093058) demonstrates that the leading 4 EOFs are necessary to capture these new features. I wonder what kind of an impact not adequately capturing the QBO has on the DLM's trend results and I think that it's time to start taking this effect into account as a standard practice. I am particularly curious to see if there is any change to the somewhat peculiar behavior of the DLM shown in Fig. 3 at mid-latitudes in the middle to upper stratosphere where the trends appear to turn around again.

While stated in the paper, it is difficult to use the word "trend" when discussing the DLM results. The authors compute the difference between start and end years (am I interpreting that correctly?) and call that the trend (obviously with the proper unit conversion), but this means any meaningful decreases in the last year (or last few years) could yield significantly different results than the MLR that yields a rough average of the change over the time period. This means that the MLR is generally better than the DLM, at
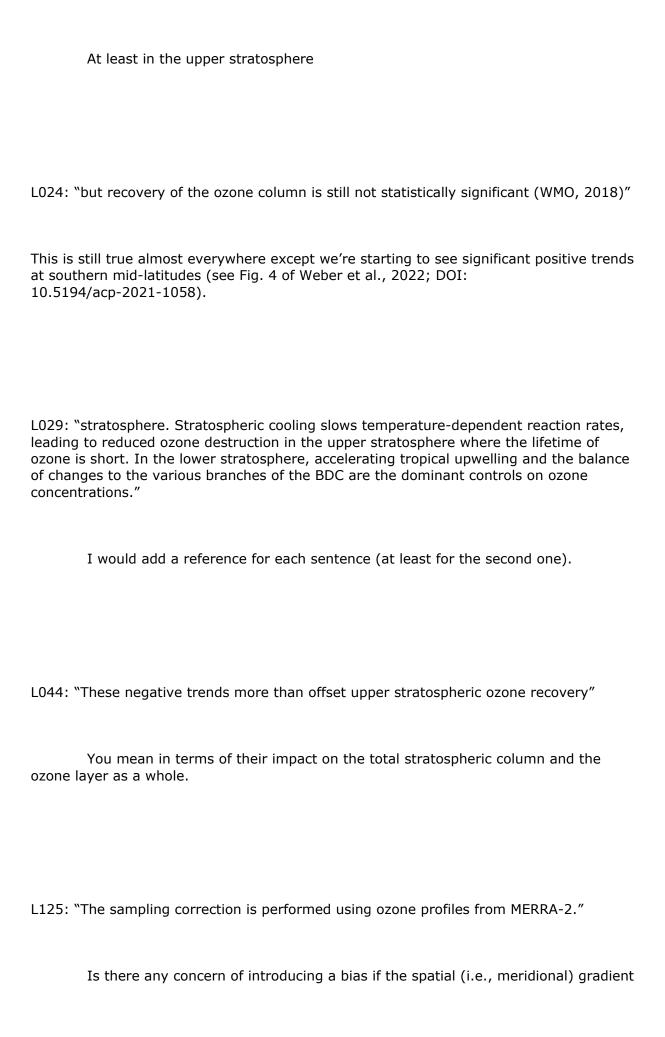
least for this calculation, at stating the overall trend results while the DLM is generally better than the MLR at stating the overall difference between any two years. I just think it's important to mention the caveats about the interpretation of results between the two methods if results from the two methods are going to be directly compared. Right now there is just the statement on Line 239 about there being a positive bias, which implies that the MLR is worse at representing the "trend" because it assumes linearity.

The "adjusted" trend distributions using MLS data are a tad odd to me. I understand the desire to assess potential drifts in the data set, but, even with the improvements between data versions that specifically mention removing some potential drifts, the verbiage in the paper reads like the implicit underlying assumption with this is that OSIRIS data has a drift and MLS doesn't and that biases the trend results. Such a statement would make the reader question the utility of the SOS data set for trend analyses and wonder why one would use OSIRIS data instead of MLS data for a data composite, particularly data after 2015 (i.e., optics temperature dependency). It would be interesting to see if any drift is also visible in just the differences between OSIRIS and SAGE anomalies over the roughly 4.5 years of data overlap as an independent verification.

Operating under the assumption that the drift between OSIRIS and MLS data is entirely a result of anomalies in the OSIRIS instrument brings up some additional questions. The difference between Fig. 2b and Fig. 5a obviously stems from the fact that the SOS data shows a second turnaround that begins in the early-to-middle 2010s, particularly at northern mid-latitudes. Comparing this (mental) difference to Fig. B1c seems to suggest that this second turnaround feature present in the SOS composite may be, at least partly if not entirely, the result of instrument drift and not anything physical. Figure 5b shows that accounting for these MLS differences doesn't change the trend significance, but I would also be interested to see another category of shading on this plot, namely "Remains not significant (oppositely sign)", and how it might manifest at NH mid-latitudes.

Specific Comments:

L023: "stratospheric ODS loading reached its maximum in the mid to late 1990s (Chipperfield et al., 2017). The ozone decline stopped around the same time …"

At least in the upper stratosphere

L024: "but recovery of the ozone column is still not statistically significant (WMO, 2018)"

This is still true almost everywhere except we're starting to see significant positive trends at southern mid-latitudes (see Fig. 4 of Weber et al., 2022; DOI: 10.5194/acp-2021-1058).

L029: "stratosphere. Stratospheric cooling slows temperature-dependent reaction rates, leading to reduced ozone destruction in the upper stratosphere where the lifetime of ozone is short. In the lower stratosphere, accelerating tropical upwelling and the balance of changes to the various branches of the BDC are the dominant controls on ozone concentrations."

I would add a reference for each sentence (at least for the second one).

L044: "These negative trends more than offset upper stratospheric ozone recovery"

You mean in terms of their impact on the total stratospheric column and the ozone layer as a whole.

L125: "The sampling correction is performed using ozone profiles from MERRA-2."

Is there any concern of introducing a bias if the spatial (i.e., meridional) gradient

in ozone is not adequately captured by the model? Also, it's interesting how the standard deviations appear to increase in the tropical lower stratosphere in Fig. 1. Why do you suppose it gets worse?

L127: "Multiplying each the OSIRIS profile with …"

"Multiplying each of the OSIRIS profiles with …"

L129: "The correction does not attempt to remove longitudinal variability (the dominant variability in the MZM), which is well sampled by OSIRIS."

I think it's worth mentioning that nor does it remove any of the random variance that is naturally present in the data. This sounds vaguely familiar except you're using model data for the correction instead of the regression fit itself. I suppose that makes sense since your regression is being applied to each latitude bin separately and thus you don't have a measure of the seasonally varying spatial gradient.

L132: "The method includes a simple diurnal correction as well, since the local time of the MERRA-2 reference profiles is fixed to noon, whereas average descending node local times for OSIRIS are between 6–8 am. It should be noted that the ozone diurnal cycle is not represented in MERRA-2 before 2004, i.e. before day and night measurements from MLS were available for assimilation (Wargan et al., 2017)."

Can you elaborate more on how the diurnal correction is derived / applied? Also, if the diurnal cycle is not represented in MERRA-2 prior to 2004, what do you do with the OSIRIS data during this time period?

L161: "If multiple tropopauses are present, the second tropopause …"

I think we all know this, but it's probably worth stating somewhere that the second tropopause is higher than the first. Maybe simply "… the second, (i.e., higher) tropopause …"

L161: "data up to and including the altitude level that contains the tropopause are discarded."

At least for SAGE data, we usually recommend excluding data up to and including 1 km above the tropopause altitude since that is roughly the vertical resolution of the data.
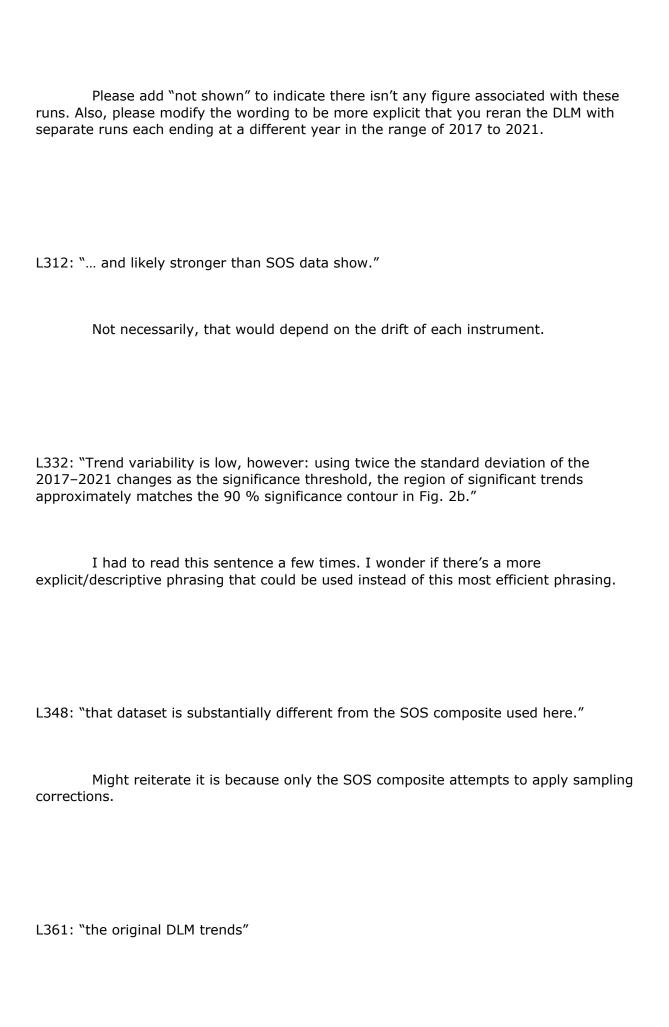
L168: "SAGE II and SAGE III/ISS data are then bias-corrected using OSIRIS as the baseline, and the deseasonalized anomalies are averaged."
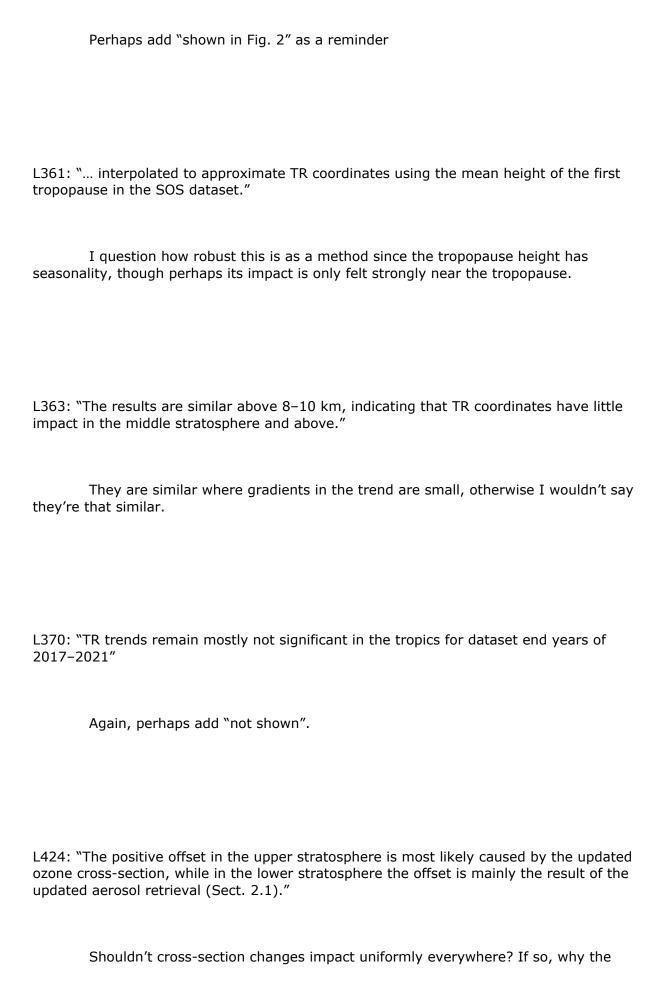
I assume this is done by adjusting their mean values to be the same in the overlap period? If the regressions are performed on anomalies, why notAlso, as a general recommendation on wording choice, I would avoid using the word "bias" when possible (i.e., perhaps use "offset" instead) as the connotation around the word "bias" is that one instrument is inherently wrong.

L184: "The GloSSAC data (v2.1, 1979-2020) is extended to 2021 by extrapolating the last value."

Perhaps state that GloSSAC is extended "through" 2021 (i.e., December 2021) as the current wording could be misconstrued to mean just up to 2021 (i.e., December 2020). Normally I wouldn't advise simply extending the aerosol data set using the last value considering how things have been recently, but 2021 was a comparatively benign year (for future work and reference, 2022 is not a benign year). Also just as an FYI, there

is a v2.2 that adds 2021 (https://asdc.larc.nasa.gov/project/GloSSAC).

L211: "The weighted MLR trend results, however, are sensitive to the exact correction method chosen …"

And how the data from the different instruments are merged together. The LOTUS Report specifically mentioned how using a weighted MLR was almost impossible with a pre-merged data set as the heteroscedasticity correction would ideally need to be applied to each individual instrument's data set separately.

L246: "The variable turnaround dates highlight one disadvantage of MLR, where the turnaround period is a fixed parameter that leads to endpoint anomalies in the trend results"

This is somewhat dependent on the trend model as the ILT method helps to mitigate this effect by generally avoiding fitting a trend line near the potential turnaround and the EESC EOFs don't specify a turnaround time at all (though they are constrained to the range of potential turnaround times governed by mean age of air).

L252: "… but the negative changes indicate a pause in ozone recovery nonetheless."

Is there any physical reason to believe in this second turnaround in the NH mid-latitude upper stratosphere?

L278: "To test the importance of the selected end year, we recalculated the SOS dataset and the corresponding DLM fit with data ending in 2017–2021."

Please add "not shown" to indicate there isn't any figure associated with these runs. Also, please modify the wording to be more explicit that you reran the DLM with separate runs each ending at a different year in the range of 2017 to 2021.

L312: "… and likely stronger than SOS data show."

Not necessarily, that would depend on the drift of each instrument.

L332: "Trend variability is low, however: using twice the standard deviation of the 2017–2021 changes as the significance threshold, the region of significant trends approximately matches the 90 % significance contour in Fig. 2b."

I had to read this sentence a few times. I wonder if there's a more explicit/descriptive phrasing that could be used instead of this most efficient phrasing.

L348: "that dataset is substantially different from the SOS composite used here."

Might reiterate it is because only the SOS composite attempts to apply sampling corrections.

L361: "the original DLM trends"

Perhaps add "shown in Fig. 2" as a reminder

L361: "… interpolated to approximate TR coordinates using the mean height of the first tropopause in the SOS dataset."

I question how robust this is as a method since the tropopause height has seasonality, though perhaps its impact is only felt strongly near the tropopause.

L363: "The results are similar above 8–10 km, indicating that TR coordinates have little impact in the middle stratosphere and above."

They are similar where gradients in the trend are small, otherwise I wouldn't say they're that similar.

L370: "TR trends remain mostly not significant in the tropics for dataset end years of 2017–2021"

Again, perhaps add "not shown".

L424: "The positive offset in the upper stratosphere is most likely caused by the updated ozone cross-section, while in the lower stratosphere the offset is mainly the result of the updated aerosol retrieval (Sect. 2.1)."

Shouldn't cross-section changes impact uniformly everywhere? If so, why the

negative change in middle stratosphere?

L452: "Here we include a seasonal term with annual and semiannual components …"

Is there still some constraint on this or is it incorporated like a standard MLR proxy? I only ask because if there is some constraint / prior, the sampling-induced differences may not adhere to this.

Fig. 2 Caption: "represents lack of statistical significance"

Do you mean at the 95% CI?

The contour lines drawn in several of the figures are almost impossible to distinguish with zooming in extensively. Please modify the graphic to increase the spacing between dots/dashes so that the reader can tell which is which when viewed at normal resolution.

Since the trend plots do not have data contour lines on them (i.e., they have significance contour lines), it might be helpful to set a specific color increment (e.g., something like 0.5–1.0% per color transition). The continuous color gradation makes it very difficult to have a sense of what the numbers might actually be.