The authors would like to thank the referee for taking the time to review this paper and for the many helpful comments that will be used to improve it. The referee's comments/concerns are listed below in red text, while the authors' responses to each comment are written below in black text.

The manuscript describes a new version of a regression model, but the model is nowhere described, neither briefly nor in detail. Much of it might be described in detail in Damadeo et al. (2014), however, it now is used for the first time to not just use SAGE II measurements, but simultaneously also HALOE and ACE-FTS measurements, and I think that should definitely be described mathematically.
As mentioned before, I was sometimes missing details about the methodology description and during the result discussion. I would recommend that the authors read through the manuscript again with this in mind (after considering the more specific comments below) to clarify any remaining items that are described too briefly.

An appendix has been added that summarizes the technique from Damadeo et al. (2014) and adds some additional detail regarding how multiple data sets are incorporated simultaneously. Hopefully this inclusion as well as some other edits throughout the paper make things clearer.

Section 2: I would recommend to remove the description of the data sets that are not used in the analysis that is described here in detail (POAM II, POAM III, SAGE III), as well as the description of their filtering. The results of the STS regression with all six data sources are only mentioned briefly once, and therefore the detailed description of those data sources seems unnecessary. It is always possible to refer to their description in the literature.

The same STS regression analysis was performed for all six data sets. The resulting trends are very similar and thus are not discussed here but some things like the seasonal cycle and diurnal variability of the ACE-FTS instrument were detrimentally affected slightly in their inclusion. This is why the paper discusses their exclusion. The resulting trends and the residual plots are also shown in the new supplement.

Page 4, line 11-12: Was the analysis done for more unit systems than just number density versus altitude? If yes, this should be mentioned in more detail. If no, I don't think the information about other unit systems is relevant here.

The same STS regression analysis was performed for the three main data sets in mixing ratio on pressure. The resulting trends are similar and thus are not discussed here but are shown in the new supplement.

Page 4, line 11-12: What vertical resolution is used for the different regression analyses? 1km?

The data was interpolated to 0.5 km increments. This has been added to the paper.

Page 4, line14: What is the spatial resolution of the daily means (for the STS)?

For occultation instrument sampling, the spatial resolution varies with latitude and ranges from ~3 degrees in latitude at the turnaround to up to ~10 degrees in the tropics.

Page 5, line 32 to page 6, line 16: Here is a detailed description of the correlated residuals that is not shown in a graph. It is very hard to follow the discussion without being able to look at something. I would recommend to either drop the paragraph, or add a figure that shows the correlated residuals.

This is a good point. Investigation of the total residuals do not have that much value compared to analyzing both the correlated and uncorrelated. However, in the interest of figure size, resolution, and space it is best to only show two types. The figure now shows the correlated and uncorrelated instead of the total and uncorrelated since these are what the paper discusses.

Page 7, line 33 to Page 8, line 2: The discussion about the SAGE II data filtering and the conclusion about the HALOE filtering is not clear. This section would benefit from some more details (how the conclusion was drawn) or some rephrasing.

This section has been rephrased and is hopefully clearer now.

Section 4.4: How do the results from Maycock et al. (Maycock, A. C., Matthes, K., Tegtmeier, S., Thiéblemont, R., and Hood, L.: The representation of solar cycle signals in stratospheric ozone – Part 1: A comparison of recently updated satellite observations, Atmos. Chem. Phys., 16, 10021-10043, https://doi.org/10.5194/acp-16-10021-2016, 2016.) compare to the findings described here?

Like other recent analyses of the response of stratospheric ozone to the solar cycle using SAGE II data, Figure 4b of Maycock et al. (2016) appears to have similar results as this paper though the comparison requires a factor of 2 adjustment (peak-to-peak versus amplitude). This is now noted in Section 4.4.

Page 9, line 3-4: How much did the trends change in Millan et al. (2016) between considering the sampling bias and not considering it?

The thing about Millan et al. (2016) is that it didn't really consider the sampling bias. It chose a single "representative year" of sampling and repeated it 30 times (i.e., over 30 years) and then ran the sampling through a model. This was done so that multiple instruments that may or may not overlap in time could be evaluated on the same time scale. However, this is not the same as the actual sampling of those instruments as they change from year to year. As such, Millan et al. (2016) essentially only considers a hypothetical scenario that is not representative of how the different data sets behave. It is, however, informative in discussing the potential problems non-uniform sampling could create.

Page 9, line 7-15: It is not clear to me what that calculated bias is based on. It is given in percent, but is it percent of ozone? Or percent of something else? If it is ozone, how was the difference between the biased value and the "centered" MZM value (middle of the month and middle of the latitude band) calculated? More details would be helpful here.

The temporal and spatial offsets between the actual average of sampling and the "centered" values refer to differences in time and location. The biases are computed as ozone values, looking at the difference between the regression fits between these two times/locations, which is now clarified in the paper. In other words, after the data is regressed via the STS method and the coefficients are retrieved, a fit value to any time and place can be computed. These differences (i.e., biases) are the differences between these fit values at the two different locations and times (i.e., actual average location and time versus "centered" location and time).

These figures (6 and 7) are meant to be illustrative of the effect, but the actual correction used later is performed for each individual profile before any daily zonal means are created.

Page 10, line 17: The latitude band "20S-20N" should be "15S-15N" here? At least Figure 8 shows the results for "15S-15N".

The figure is correct and now the paper agrees.

Page 11, line 5-17: It would be good to be a bit more detailed in the description of the different methods here. It is not very obvious from the text that there are indeed 4 different regression models discussed.

The corrections that are applied are described in the paper and the regression that is applied to each of them is the same (only the input data changes with the corrections). This is now noted in the paper.

Page 11, line 18-23: It was not clear from the supplementary material how the text would change here with the updated way of calculating the uncertainties on the trends.

This paragraph has been rewritten to reflect the new methodology as detailed in the supplementary material and an appendix has been added to the paper to mathematically describe the process.

Page 11, line 33: The increase of about 1%/decade in the NH mid-latitudes is not very obvious in the updated plot. Is this a remnant description of the old plot? If not, I would recommend to adjust the description to ensure the reader knows where exactly the 1%/decade increase takes place.

It is now clarified that analyzing the impact of the diurnal correction implies comparing "MZM DCorr" and "MZM Raw" while analyzing the impact of the "seasonal" correction implies comparing "MZM DSCorr" and "MZM DCorr".

Page 13, line 6-8: It is referred here to the "recovery trend results in Fig 11", however it is not specified which results exactly. STS results? MZM? For all four results shown in Figure 11, I am not sure I see the pattern that is supposed to match the ACE-FTS drift pattern. Should this be Figure 12? If not, could you explain in more detail here where the similarities in the figures are?

The paper now mentions that it is the STS results we are comparing to. Unfortunately having some drift between the different data sets does not create a direct correlation with equal patterning into the resulting trends. The different drift patterns for the different instruments over the different time periods will alias into the different proxies in different ways. As such, the changes between Figures 11 and 14 may not be readily apparent simply from looking at Figure 13. However, Figure 14 does illustrate the aggregate effect of ignoring the possibility of drifts (but not offsets in the mean) between the different data sets.

Page 13, line 14: Where exactly are the results changed by up to 2%/decade?

The differences occur at various locations, most notably where the "recovery" trends were strongest in Figure 11.

Page 13, line 17: "limitations in these regression techniques" -> which regression techniques are referred to here? The ones used in this analysis? All four of them have the same limitations?

Page 13, line 25-26: “only a single uniform seasonal cycle should be used for these analyses” -> which analyses exactly are referred to here? Any regression model? Only the ones used here?
Page 14, line 1: “This study also highlights the limitations inherent in these techniques: : :” -> which techniques are referred to here?

It seems the word “these” implied specificity, whereas we really meant regression techniques in general, not just this one. The instances of the word “these” has been removed and the sentences corrected.

Page 14, line 8-9: “With sufficient overlap: : :” -> does MLS provide a sufficient overlap with SAGE II and HALOE to allow the suggested analysis?

This is actually both a good and very difficult question. There were/are many instruments with measurements after ~2000 that can be used to try to determine potential recovery trends. For most of these instruments, an instrument like MLS does have sufficient overlap to try to do this kind of analysis. Unfortunately, only a few of these instruments also provided data prior to 2002. The current problem is that the representation of the solar cycle can have a significant impact on derived trends and truthfully more than one solar cycle needs to be sampled by the data used in the regression to adequately constrain it. This is achievable with current measurements, but only by including data prior to 2002. Otherwise, we’ll need to wait for more measurements. This is where the difficulty of the question comes in: Is there enough overlap between any high-sampling instrument in the modern record and SAGE II or HALOE to link the time periods? While the assumption in previous works (both in regression analyses and in the creation of merged data sets such as GOZCARDS and SWOOSH) is “yes,” a definitive answer has, to our knowledge, never been investigated and is beyond the scope of this work.

Page 23: Maybe add “filtering” in the last line of the figure caption, “: : :, though results with filtering are similar”

This has been added to the paper.