Response to RC1

Thank you for your thorough comments—please see our response in blue in the text below.

The manuscript uses ozone time series from the SBUV satellites along with model simulations from the Earth System Chemistry Climate Model (ESM) and the Chemistry Climate Model Intercomparison (CCMI) to investigate long-term trends in ozone and compare them with ozone variability from the model simulations. There are two major results: 1) wide averaging kernels of observations like SBUV mix information from different vertical levels. This can shift and distort vertical trend profiles. 2) Uncertainty estimates are necessary to determine if a trend is significant compared to natural variability. Again, wide averaging kernels combine information from different altitude levels, and this tends to result in underestimated variability. One example are ozone variations associated with the QBO. These are important for trend estimation in the atmosphere, but are reduced and smeared out in SBUV data. This tends to result in errors when accounting for the QBO, and in incorrect uncertainty estimates. Overall this is important information. The paper is well written and deserves publication in ACP.

There are a number of points that should be improved, though:

Abstract and other places in the text: The authors point out a number of problems with merged satellite records (sampling, calibration, instrumental differences, ..). I find this misleading, because the manuscript does not account for any of these "merging" issues. The only issue addressed here, are the SBUV averaging kernels. So I don't think the merging issues should be mentioned in the abstract. The first two sentences should be dropped. In line 4, "one merged" should be replaced by "the SBUV MOD". In line 11 "merged satellite records" should be replaced by "records from instruments with wide averaging kernels”. We understand the potential for a misunderstanding, and have modified the abstract to better reflect our focus (in particular, we changed “one merged record” to “one record”). We also mentioned “broad nadir averaging kernels” in that sentence, implemented the requested change at line 11, and changed the title to explicitly mention nadir viewing geometry. We do not mention SBUV MOD in the abstract as we prefer to avoid defining acronyms there.

Line 16/17: "continued recovery" I think this should be "beginning recovery". This also applies to other places where "recovery" is mentioned. We are just at the beginning of ozone recovery. We are far from "recovered" and, as explained in the paper, we are also far from significant recovery in many regions of the atmosphere.
We enacted this change.

Lines 20,22: delete "lower" and add "in the tropics" after "abundances". The main branch of the Brewer Dobson circulation transports ozone rich air in the mid- and upper stratosphere from the tropics to the extra-tropics. Enhanced upwelling in the tropics is decreasing ozone in the tropical lower stratosphere.
Implemented as suggested.

Line 29: suggest to replace "newly detected" by "recent illegal" We changed this to “recent noncompliant,” as in the WMO report, rather than “recent illegal.”

Line 30: "increasingly large" is too strong. I would say "possibly increasing" Done

Line 32: explicitly add "e.g. Hunga Tonga-Hunga Ha'apai in 2022". Done

Lines 46-47: Reword. You are not adressing merging challenges, you are only adressing the effects of wide averaging kernels. Done: "we turn to the nontrivial effects of error propagation in algorithms used to retrieve ozone abundances from space-borne nadir measurements."

Figure 1: Please explain why the power density of the ESM4 historical runs in the 2 to 20 year range is lower in the top panel and larger in the bottom panel. I guess it is due to applying the SBUV averaging kernel in the top panel. I think this needs to be said / clarified in caption and text. I looks like the averaging kernels reduce variability. Also change the text in the legend in the top panel e.g. to ESM4 historical@SBUV resolution. It needs to be different from ESM4 historical in the legend of the lower panel. The difference in the ESM4 historical run PSDs between the top and bottom panels is indeed due to differential sampling. In the top panel, the historical run is sampled with a 13-year sliding window, while a 165-year window is used in the bottom panel. As a result, the spectral resolution is much coarser in the top panel and several frequency peaks are not resolved in the 2-20 year range, as you point out. We note that no SBUV kernels were applied at this stage; only the temporal sampling is considered. We clarified this in the legend.

Line 98: better to say "pre-industrial simulations" instead of "these simulations" Done

Line 134: I would start a new paragraph after NOAA. It should also be pointed out here that SBUV-MOD and SBUV from NOAA have wide averaging kernels and use the same nadir-viewing satellite data. On the other hand, GOZCARDS, SWOOSH, and the other data sets use LIMB and occultation instruments, which have much finer altitude resolution. We agree on the importance of the distinction between the two categories; to avoid short paragraphs or mentioning related datasets across two paragraphs, we use a bullet list instead.

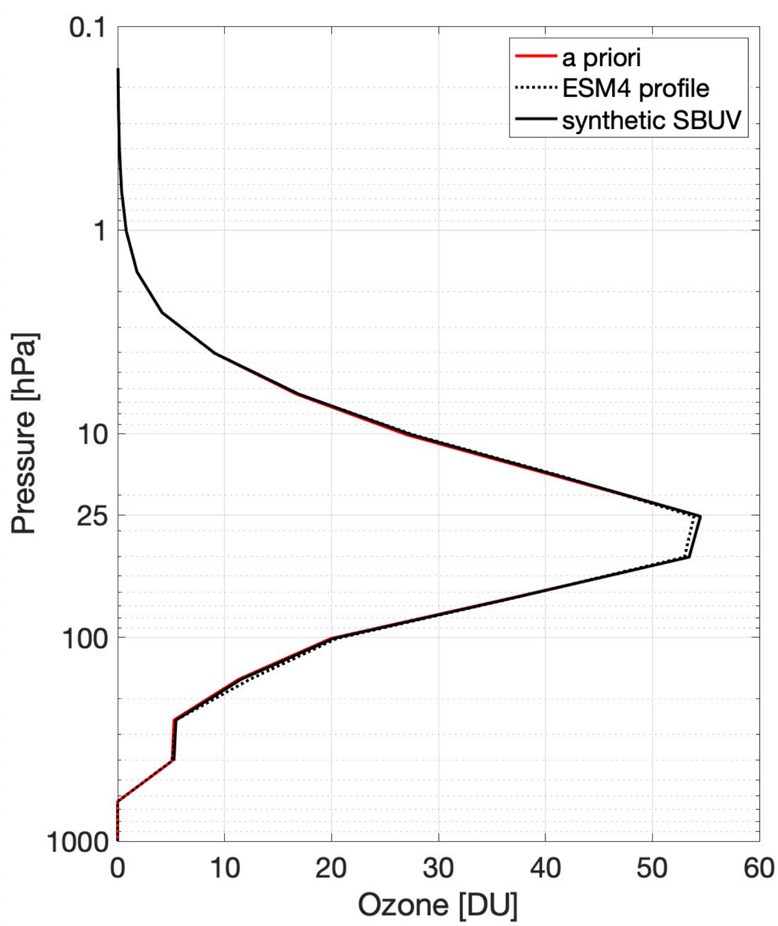

Figure 2: It would be good to have another panel showing the a-priori ozone profil and the two profiles, in addition to the panel shwing the deviation of the two profiles from the a-priori. The differences between a priori and raw/kernelized model ozone profiles are relatively small (especially since in this case, the a priori is equal to the mean), and therefore difficult to visualize as is (see attached plot). This is why we show the profiles as departures from the a priori, so as to best reveal the effect of the

averaging kernels (see caption). To address your point, we've added the order of magnitude of the errors shown in the caption of Figure 2.

Line 230: for clarification, after "70 hpa", add "from the model simulations"? As shown later the model QBO is quite different from the "real" QBO. We adopted this suggestion.

Line 233/234: "see NOAA ... June 2024" Again, I am assuming you are using the ENSO of the model simulations, not the "real" ENSO. So, while the NOAA page is a good reference, it is kind of misleading here. Please move, reword / clarify. The NOAA page also details how to calculate the ENSO index, but we agree it could be confusing so we removed it.

Line 275: It would be helpful to explain skewness and kurtosis a bit more here. What you are saying is that the residuals are often not normally distributed, with distributions leaning to the left (skewness greater than 0.5), and distributions that are narrower than a normal distribution (positive excess kurtosis). We rephrased to clarify. Note that O3 residuals can have negative or positive skewness, so we explain that the residual distributions are often asymmetric.

Line 294: would be helpful to add "(e.g. the red curve in Fig. 4c)" after "earlier", and " (the black curve in Fig. 4c)" after "itself". Done.

Figure 5: I find it difficult to see much in panel a.) I think it would be better to show here the ratio (standard deviation)/(average values), i.e. the relative standard deviation, e.g. as percent. The overall ozone distribution (average values) will be well known to the readers. The relative standard deviation (or variability) in percent will be much better to compare, e.g. to trends which are also in percent per decade. If the authors don't want to change panel a.) they should add another panel with the relative standard deviation. We've modified the figure as requested—showing the relative standard deviation is an effective way to visualize internal variability evenly across the globe.

Line 314: add "CCMI" before modeled? I assume you are talking about trends from CCMI here. Correct—done.

Line 353: "sampling and retrieval". The way I understand it from section 2.3 you are using SBUV-MOD monthly zonal mean data. I assume you are also using monthly zonal mean data from the model simulations, but without accounting for the specific times and locations of the individual SBUV measurements. Am I correct? Are you dropping polar night data? My guess would be that your model sampling is the same in both hemispheres / polar caps, so "sampling" differences should not play a role here. You only see differences due to the retrieval / averaging kernels, which mixes and redistributes stuff from different altitudes. But, in my understanding, you do not look at sampling differences, i.e. differences due to the specific times and locations of the SBUV observations. So delete "sampling and". SBUV kernels are not available during polar night (they are NaNs) and we now mention this at the end of Section 3.1.1. Our method therefore *does* capture the annual dependency of the SBUV sampling. However, since we are not using the SBUV orbital data to reproduce the specific sampling times and locations in the model, we changed "sampling" to "measurements." As a note, we are not using the SBUV orbital data because:

1. The pre-industrial model run was not available to us with better than monthly resolution at the time of the analysis;
2. SBUV retrievals are averaged zonally and monthly across a large number of profiles (~1000, see Section 3.1.2), and we assume that accounting for specific times and locations of sampling would have a limited effect on the results (see Section 3.1.2);
3. Applying a SBUV-like sampling in the literal sense to the 500-year-long pre-industrial simulation is difficult because the SBUV record is much shorter. Even using the longest continuous portion of the SBUV record (~13 years) as sampling template could introduce a ~13-year cycle in the 500-year-long synthetic observations. This is also why we use one chosen year of kernels to perform the sampling (see Section 3.1.1).

Section 4.3: What you have done is applied averaging kernels and then done trends (avk -> trend). An interesting question to me is whether doing trends first and then applying the averaging kernels to the trend profile (trend -> avk) would give the same result. For the mean this should be the case, because both averging kernels and trend derivation are linear operations on the underlying data. Not sure what it means for the uncertainties though.
This is an interesting consideration; perhaps kernel operators could be applied to trend profiles in DU/decade (rather than %/decade). However, this would require determining an *a priori* trend profile, that is, a representative trend profile for the record. Determining such a profile is in essence the very goal of the scientific community working with merged satellite records. In addition, it is unclear to us what real world process this would represent—trends derived from O3 retrievals already incorporate the effects of kernels after all. Thus, we have not explored this possibility

Figure 7: not sure what the difference between these three panels is. Are you just assuming three different trend profiles? What is the difference between the left panel and the middle panel? Please explain.
Thank you for pointing this out. The panels differ indeed in the idealized trend profile shown (the location and sharpness of the trend maximum). We clarify this in the legend of the figure and in the text.

Figure 10: Please put a label / title on each of the three panels. Top panel is 1.6 to 1 hPa, middle panel 25 to 16 hPa, bottom panel is total column. Done.

Line 438: replace "the total column" by "some ozone column metrics”? Done.

Line 442: "modelled climate" instead of “climate" Agreed that strictly speaking, the reference is to the modelled climate system, however, we feel that in this more general context it is more appropriate to recall the problem of climate internal variability at large.

Line 443: "large" How large? Give numbers. Overall, the changes in uncertainty / significance don't seem to be very large for SBUV (maybe 0.1 or 0.2 % per decade for trend uncertainty according to Fig. 9, a few years according to Fig. 10). They should be smaller to negligible for the LIMB satellites which have much better altitude resolution. Also in lines 6 to 8 in the abstract, it would be good to give some numbers. Fig 9 is not directly relevant to trend uncertainty; rather, Fig 9 quantifies errors in the smallest detectable trend, as a metric for detectability. Fig 7 does quantify trend uncertainties for hypothetical scenarios, and they can be large (as large as 1 %/dec, i.e., about 100% errors near local maxima in the 'true' trend profile). Regarding errors on significance, Fig 10 shows that up to 13 years may be needed, and Figs 5 and 6 together show that several decades may be needed in some locations (near 25 hPa in the tropics for instance). We've amended our conclusion section and the abstract to be more quantitative, as suggested. Since our trend error estimates come from hypothetical ozone recovery scenarios, we specify this in the new statement.

Line 450, and also discussion of Fig. 7: You might want to refer to Fig. 3-10 of WMO 2022, which shows the latitude altitude distribution of ozone trends from various satellite records. SBUV-MOD is shown in the top left panel of that Figure. You can very clearly see that the peak of upper stratospheric trends is shifted downwards to about 10 hPa in the SBUV-MOD record, and that SBUV-MOD trends are reduced in the 2 to 3 hPa region. Thank you for the suggestion; we now point out this reference here and in the discussion of Fig. 7.

Line 457: "which has been large in recent years". I would say "which can be large". Compared to Pinatubo in 1991, or El-Chichon in 1982/83, most recent volcanic eruptions, even Hunga Tonga, have only had a small influence on stratospheric ozone. Done.