Response to RC2

Thank you for your careful review—please see our response in blue in the text below.

Summary:

This paper reports on the effect of averaging kernels from nadir-type merged ozone timeseries, here SBUV Mod, on the detectability of long-term ozone trends. Using CCM model data, the uncertainty in trend detection due to internal variability as a function of the record length was investigated. As stated in this study, the uncertainties in the representativeness of modeled internal variability for observations is rather large so that the results from this study should be considered with some caution.

The main issue with the averaging kernels from nadir observations, is that they are quite broad (low vertical resolution) and in the lower stratosphere asymmetric, so that trends at a given altitude have, in some cases major, contributions, from other altitudes.

Overall the paper is well written. My major concern is that neither the abstract nor the introduction clearly state what the scope of this study is and that the results are limited only to a particular subset of satellite ozone profile measurements. One should not use the term retrieval as a synonym for averaging kernels (AK). AKs are used in the retrievals but they are inherent to nadir observation geometry in the UV. So regardless of the algorithm used, all AKs are similar for the same observation geometry. Also a clear distinction should be made early on that this study focuses on nadir-derived ozone timeseries and that other datasets based upon limb observations have narrower AKs and are not investigated here. We've added that distinction in the data Section (2.4), and we've replaced references to the 'SBUV retrieval process' with 'SBUV kernels'.

In many figures panels are missing titles. This would make the figures more readable even without checking the figure caption. This applies to Figs. 1. 3, 4, 5, 6, and 10. We've added panels titles in Figs. 1, 3, 4, 5, 6, and 10, as suggested. Where panels were labeled (a, b, c, etc) we've moved the labels to the panel titles to reduce visual clutter.

Specifics:

Paper title: The paper title is quite unspecific. I also think that it is not clear what is meant with "forced" here. I strongly suggest to add "nadir-derived", like: "Detectability of trends in nadir-derived stratospheric ozone". We appreciate the criticism, and understand that the term 'forced' could cause some confusion, especially because the forcing in this case corresponds to the *decrease* in ODS abundances. We could clarify that the analysis is done "in the context of climate variability," but for the sake of keeping the title short, we propose: "Satellite nadir-viewing geometry affects the magnitude and detectability of long-term trends in stratospheric ozone."

l.2: In this paper the focus is on the impact of averaging kernels, which are inherent to the observation type rather than the retrieval algorithm (see comments above). Retrieval algorithms can have different settings that may also impact trends (e.g. uncertainties due to temperature dependence of ozone cross-sections). Do not use retrieval algorithm as a synonym for averaging kernels (check also thee entire paper) Thank you for pointing this out; as noted above, we have replaced “retrievals” with “kernels” throughout the paper.

l.4: mention here the use of SBUV MOD and that the data is based upon satellite nadir measurements. We prefer to introduce the specific data sets later on to avoid defining acronyms in the abstract. We instead added mention of the use of broad averaging kernels.

l.22: "to decrease lower stratospheric ozone abundances". This is limited to the tropics, a corresponding increase in extratropical ozone is expected. Done.

l.42: "disagreement". Probably better to say that “models, observations, and trends are highly uncertain” Done.

l.42: see also WMO (2022, p. 165) which recommends to be very cautious with trends derived from reanalyses. Done.

l.45: A bit more explanation on the role of internal variability is needed here (see for instance, doi:10.1038/s41612-023-00389-0). Does the determination of internal variability not require an ensemble of model runs rather than using single (or two) model runs as done in this paper. Please discuss. Quantifying internal variability does require a large sample size, as discussed in the reference you provide. In the case of model simulations with time-dependent external forcings (e.g., for historical time periods), exploring the diversity of initial conditions is crucial as external forcings interact with the internal variability, and small differences in initial conditions can yield different future evolutions. The same is not true of the pre-industrial simulation we are using as control (i.e., without external forcings): that simulation captures a stationary system in which the memory of initial conditions decays due to chaotic processes, and the timescale of the simulation is much longer than the dominant variability modes. In other words, the use of multiple model realizations is important when quantifying internal variability during a particular time period, but that is not our case: we focus on the role of internal variability over *any* time period of chosen length. For this reason, we follow the standard practice of using a single, long control simulation (500 years in our case) as reference for internal variability.
We added a short discussion in the first and second paragraphs of section 2.1 to clarify this.

l.46: the first sentence in this paragraph is misleading as it suggests that all these aspects are considered in this study (see earlier comments). We have modified this sentence in accordance with your comment and that of RC1.

l. 108: add some references to the SBUV MOD data, e.g doi:10.5194/acp-17-14695-2017. We added both this reference as well as the earlier Frith et al. 2014 reference (https://doi.org/10.1002/2014JD021889)

l. 133: one should mention here that all LOTUS datasets except for SBUV are derived from limb observations with narrower and less asymmetric averaging kernels (see earlier comment) We have modified this sentence in accordance with your comment and that of RC1.

l.150: What is the vertical sampling of ESM4.1. State it here or earlier when describing model data. Also it may be helpful to provide some typical numbers for the vertical resolution of nadir-type ozone profiles (see, for instance, 10.5194/amt-14-6057-2021), here or before when describing SBUV MOD. The model features 2-3 km resolution in the lower stratosphere; we've added that to the model section. The range of vertical resolution for nadir sounders is relatively wide and spatially dependent; to avoid lengthening this section, we simply state that the SBUV vertical resolution is comparable to that of other nadir records.

l. 169: this has not to do with retrieval quality as this is an inherent physical property from the retrieval using nadir observations (see comments above). Maybe "limits" is better. We have rephrased the sentence to eliminate the term 'quality,' as suggested.

l. 227 (Eq.) earlier you mentioned that the LOTUS regression on the six merged datasets includes AO/AAO, and/or NAO. Not important here? The LOTUS report states that the AO, AAO, and NAO proxies have a negligible impact on trend estimates. We now mention this in Section 3.1.3 where the removal of known modes of variability is discussed.

l. 253: replace "retrieval" by "data". Done

l. 263. Regarding ozone the following paper paper is also relevant here doi:10.1029/98JD00995. We've added both this reference and that to:

Tiao, G.C., Reinsel, G.C., Xu, D., Pedrick, J.H., Zhu, X., Miller, A.J., DeLuisi, J.J., Mateer, C.L. and Wuebbles, D.J., 1990. Effects of autocorrelation and temporal sampling schemes on estimates of trend and spatial correlation. *Journal of Geophysical Research: Atmospheres*, *95*(D12), pp.20507-20517.

in the section.

l.275: The distribution with s and k should be shown in Fig. 4. We show normal distributions (s=0, k-3=0 by definition) to highlight the non-Gaussian behavior in the ozone residuals, since the assumption that residuals are normally distributed is required in other methods to calculate the time of emergence, unlike the one we use here.

l.345: AK-adjusted is better than "retrieval-adjusted" We use 'kernel-adjusted' to avoid defining additional acronyms.

l.356: errors cannot be negative, but trends can be. The standard deviation is not unexpected to peak near the ozone peak. In this context, the error is the relative difference between two variables, which is at times negative. We now make a mention of this when discussing negative errors, to avoid confusion. Regarding standard deviation: relative errors on the time of emergence are independent of it, see a detailed discussion in the paper which describes the ToE method we use: https://doi.org/10.1029/2024GL109638.

l.381: "... this analysis shows why trends should be analyzed as vertical profiles rather than at individual vertical levels." I do not understand what is meant here. Do you mean that trends should not be evaluated at a single altitude level, but for all levels? Does this make sense? Please clarify. RC3 had a similar comment. Our statement was meant to echo the result that SBUV kernels can distort the trend profile; we've replaced it with a more general comment about the importance of accounting for averaging kernels when analyzing trends.

l.410: Of the Limb datasets only two use MLS (GOZCARDS, SWOOSH), rather say hat the LOTUS mean trends are heavily weighted by trends from the higher vertically resolved merged limb datasets. Done

Figure 9: Probably legend is wrong (no shading for “undetectable"). It seems that the "undectable" and "undectable according to SBUV" are very similar and differ only by a few tenths of a percent/dec for most altitudes. I think this should be mentioned. Thank you for spotting this error; we modified the figure to ensure that legend and shading match. We have also indicated in the description of Figure 9 that the absolute changes in the detectability envelopes are small.

l.435: Regarding Antarctic column ozone recovery, add some references here. We added a reference to Chapter 4 of the WMO 2022 Ozone Assessment Report.

l. 444: emphasize in item 1, that this is only true for nadir-derived ozone profiles. Done

l. 457: "large in recent years". Are you referring to Hunga-Tonga. Please specify the large events. Following a similar comment by RC1, we changed this to ‘which can be large’.

l. 462: add some references dealing with size and timing of the ozone hole. We added a reference to Chapter 4, Section 4.4.2.1 of the WMO 2022 Ozone Assessment Report.

Fig. 4: Both pre-industrial and 500-year runs are labeled ESM4 in the panels. Use different abbreviations for each run. The pre-industrial run is 500 years long (see Section 2.1). We added ‘pre-industrial’ again after ‘500 years’ in the legend to avoid any confusion.

Figure 5a: More common unit for profiles is DU/km (equivalent to number density). Clarify here. Done.

Fig. 6: add the corresponding SBUV ozone profile to the right panel. We are unsure which panel you are referring to, and have therefore made no changes in response to this comment.

Fig. 7. What are the different conditions between the three panels? different zonal bands like in Fig. 8?, but averaging kernels from 42.5 degs only used for the synthetic data? Note that averaging kernels are solar zenith angle (SZA) dependent and SZAs of SBUV measurements are different in the tropics and higher latitudes. We clarified that the three panels show three different, hypothetical trend profiles in the mid to upper stratosphere, as would be seen by SBUV at 42.5N.