Reply to Anonymous Reviewer #1

The paper is dedicated to the important issue: uncertainty estimates for the merged SBUV MOD v8.6 ozone profile dataset. This paper contributes to understanding the differences between two SBUV datasets, and dominating sources of uncertainties in the merged dataset. The implications for the evaluation of ozone trends are discussed. The paper is interesting and well-written. Please find below my comments and suggestions.

Thank you for taking the time to review the manuscript. Our responses are in red below.

MAIN COMMENTS

1. Please clarify whether uncertainties of NOAA-9 data (which are not used in the MOD dataset) are used for the illustration in Figure 7. If yes, I suggest including one more panel, which will illustrate only uncertainties associated with the data that are actually used for creating the MOD dataset.

No, these illustrations do not include NOAA-9 data, but we see the possible confusion in the text when noting that NOAA-9, while not included, has been calibrated to NOAA-11 in the Version 8.6 processing. To clarify we made the following changes to the manuscript text and Figure 7 caption:
P7 L17 Changed text to “Nimbus 7 is calibrated to NOAA 11_pm, while all later instruments are calibrated using NOAA 17 as the reference.” and added text to Figure 7 caption, “In these examples only data included in MOD are considered.”

2. In Sect. 3.4, please provide more details (briefly) on the NOAA Cohesive dataset. It seems to be different from that used in (Tummon et al., 2015) – true? What are the differences?

Yes, it has been changed, thank you for highlighting this point. We add the following as the first paragraph of section 3.4:

The NOAA Cohesive data set is an independently constructed merged ozone record based on the SBUV series of instruments (Wild et al., 2017). NOAA Cohesive takes an alternate approach to account for the offsets between SBUV instruments. Only one instrument is included at any given time, and additional external offsets are applied to improve consistency over the record. This version of the NOAA Cohesive record is an update from that reported in the SPARC/IO$_3$C/IGACO-O$_3$/NDACC (SI^{2}N) Past Changes in the Vertical Distribution of Ozone Initiative (Hassler et al., 2014; Harris et al., 2015). Comparisons done within SI^{2}N indicated the potential for unphysical trends when a successive head to tail adjustment scheme was applied, as a result of the lower quality NOAA-9 (NOAA Cohesive uses NOAA-9 rather than NOAA-14 in its construction) and NOAA-11 data (Tummon et al., 2015; Wild et al., 2017; also Figure 7a). NOAA Cohesive was revised to not include offsets to Nimbus-7 and NOAA-11 data in the early portion of the record. The long overlap periods from NOAA-16 to NOAA-19 are used to make adjustments in the latter period of the record using NOAA-18 as the reference instrument, but these adjustments are not linked back to the beginning of the record. The resulting data set no longer shows unrealistic trends (Steinbrecht et al., 2017; Wild et al., 2017).

3. I think, in Figure 1 it is worth to include all NOAA-SBUV satellites, also before 2000.
We will update this figure as suggested

4. Figure 3 and the related text: what will be the changes in error bars if autocorrelation is taken into account?
We have updated Figure 3 to include autocorrelation, and made related changes to the text.

DETAILED COMMENTS (small clarification, technical corrections)
P.1 L. 21 “dominate” -> dominant?
Corrected

P.2. L. 24 “Damedeo” -> “Damadeo”
Corrected

P.4, L. 31 “In the 6.4 to 4 hPa layer we see a smaller decrease” in the 1980-2000 period?
Yes, this is now clarified in the text.

P.8 L. 5 “though” -> “through”
Corrected

P.8 L. 10 “very little difference” – How small? Please quantify
These differences are negligible for all fits, and the text is altered to state this.

Reply to Anonymous Reviewer #2

This manuscript provides a timely, and important update of the SBUV MOD merged dataset, and a natural follow on from Frith et al., 2014's work on total column ozone that use the same instruments and a similar uncertainty estimate approach. The manuscript is well written, with clearly presented analysis, and careful consideration of uncertainties. I commend the authors for their effort in these regards. Given that the main problems in the trend estimates are due to how artefacts (drifts and offsets) are accounted for in the merging procedure, and this is the most prohibitive component in estimating decadal trends in merged records (e.g. Harris et al., 2015), I have put forward some suggestions below of things authors should perhaps consider to further improve/reduce-uncertainties on their estimates. Such considerations may point towards better approaches to merging the data in future iterations.

Following consideration of the points below, I recommend publication.

Thank you for taking the time to review the manuscript. Our responses are in red below.

Main comments: P3, L27-29: Does a period of 24 months, instead of 12, when volcanic aerosols likely persisted (at least in the lower stratosphere), make a difference to your uncertainty estimates? Have you performed any sensitivity tests, or does it simply make very little difference once you get beyond 12 months. Some ozone records continue to have artefacts clear in the data even after 12 months.

In constructing the merged record we attempted to remove the portion of the SBUV data that was clearly affected by interference of volcanic aerosols within the algorithm, based on both internal algorithm flags and external comparisons. We believe the remaining artefacts are largely real ozone responses to the aerosol rather than algorithm problems, though some minor effects may linger. For El Chichon we remove ~ 1 year of data based primarily on internal algorithm flags and comparison of total ozone with TOMS data, which are not strongly affected by aerosols. For the Pinatubo period we remove ~ 18 months of data based on internal flags and comparisons with UARS MLS data. But in either case, the Monte Carlo uncertainties will not be affected because both volcanoes occur when only one instrument is included in MOD, while the MC parameters are based only on overlap pairs. However, the remaining volcanic signal and potential continued algorithm artefacts would affect the uncertainty of a regression fit. We have done some limited sensitivity tests of our volcanic aerosol fit within the regression, but for the purposes of this paper we simply note that the period after the volcanic eruptions should be treated with caution, and add some clarification to the periods of data removed.

We changed the text as follows:
Additionally for the profile data set, measurements are removed for a year after the El Chichon volcanic eruption and for 18 months after the eruption of Mt. Pinatubo to avoid periods when volcanic aerosols likely interfered with the algorithm (Bhartia et al., 2013). We identified these periods using internal algorithm parameters and external data comparisons, but caution that small volcanic effects may persist beyond the period of missing data. No NOAA 9 profile data are used in the profile MOD (limited NOAA 9 data were used in the total ozone MOD to fill data gaps).

P4, L12: Given the much higher sampling (by time and latitude) of MLS, and given the agreement with the diurnal variation as observed from Mauna Loa ground based station, is MLS therefore providing a more ‘daily-average’-like ozone estimate? If this is correct, then it would further strengthen your argument to state this so that your biases are clearly a result of the diurnal cycle.

AURA MLS makes daytime measurements at the same local time each day, ~ 1:30pm, so it does not represent a daily average or average over time of day. We believe a portion of the differences are diurnal, but without a model it is very difficult to do more than suggest this relationship.

P4, L22: I think it may also be possible that the higher uncertainty is a result of a ‘non-linear’ divergence of instrument records, given that all the other SBUV datasets presented retain a relatively stable equator-crossing time. NOAA 16 has the most rapid changes, and I wonder if you compare with earlier instruments (in terms of their uncertainty and change in equator-crossing time) if this is the case. If true, then a non-linear fit (rather than a linear) might account for much of this uncertainty, and is a valid model given you know that the change in equator crossing times become more rapid as you approach/leave the terminator (e.g. this might improve estimates discussed on P6, L33 regarding time-dependent drifts), and therefore tighten your uncertainties both in the merged dataset, and in your analysis.

We have looked at the drifting orbits in the earlier records, but have not been able to establish a relationship, linear or otherwise, to orbit drift beyond noting that the measurement noise and uncertainty increases as the orbit nears the terminator. We now include the orbit drift for all instruments for context in Figure 1. We have also updated Figure 3 to show the drifts with autocorrelation estimated in the error bars. At 35-50N the trends are now not significant, but at 35-50S the N16 trends remain significant. We verified this against N17 data over the longer overlap, and update the text to reflect this. We suspect orbit drift is a primary cause of larger drifts in the N9-N14 instruments, but the instruments drift in different ways over a similar change in equator crossing time. So while it is possible that the faster N16 drift contributes to the greater drift against AURA MLS, we cannot confirm this.

P4, L22-24: Following the argument in the previous point, your comment about a short overlap influencing the uncertainty is a valid point. But, the stabilisation of the drift as reported by Kramarova et al., 2013b taking 70 months to becomes stable with longer periods, also may hint at a non-linear fit being better than a linear one to account for the drifts. My interpretation of Figs 1 and S19 in Kramarova et al, 2013b, look like there may be a non-linearity in the drift estimate, in line with the turn-around in the latitude change of the orbital drift (which would then cancel out a drift in one direction later on as we see in Fig. S19); but this is eyeballing and perhaps I have misinterpreted Fig S19. Have you considered such an approach, or otherwise why would such an approach be inappropriate?

We have not considered such an approach, primarily for the reason stated above, that we cannot isolate a relationship that holds over multiple instruments even when the orbits drifted in a similar fashion. However we definitely take the reviewers point that there are variations in the data that are not linear. We know for example that there are seasonal and QBO-scale variations (particularly relative to independent data sources), and occasionally a drift develops later in the record but is not consistent throughout. But the lack of consistency limits our ability to model these differences.

Fig. 4 & P4, L34/35: Could this difference in seasonal cycle also be due to slightly different times of day being observed, and slightly different observational properties of the instruments (rather than being due to chlorine)?

Yes, we expect that at least some of the differences are due to algorithm issues, as there are seasonal differences between the instruments in the same time period (though not as large as the seasonal cycle between the early and later periods shown here). We do not believe it is due to time differences, though they exist. Nimbus 7 is measuring near noon, while the later instruments are measuring near 2pm. However, the AM instruments, including N17 do not show as large a variation even though they are measuring closer to 10am.

We altered the text to read:
These seasonal differences may in part be a real response to long-term chlorine changes (e.g., Stolarski et al., 2012 and references therein), though differences in the amplitude of the seasonal variations among instruments over the same time period also suggest instrument issues account for some of the seasonality.

P6, L4-5: "We scale ...pairs." Do you mean that you compute (and redistribute) the uncertainty between pairs (and only pairs) of instruments (i.e. you only consider pairs, not triplets etc when more than two exist)?

Yes, we compute, collect and redistribute uncertainties between all available pairs of instruments. There is some degree of double counting when more than two exist. For example, we include N17-N18 (72 months), N17-N19 (31 months), and N18-N19 (45 months). However each is computed over a different time frame, indicated by the number of months in parentheses. In this way we do sample, though not directly, possible non-linear effects. If N18 for example develops a drift later in its record, this drift will show up in the N18-N19 difference, but not in the N17-N18 difference.

P6, L13: Why is 24 months required for an overlap? Don't shorter periods simply lead to higher uncertainties, and therefore lower weights?

There can be a difference in the seasonal amplitude between SBUV instruments (also noted in Kramarova et al., 2013b) that can interfere with the drift computation. We remove the seasonal cycle in the instrument differences before computing the linear drift. Requiring at least two years of data helps avoid problems of a partial seasonal cycle aliasing as a linear drift. Though we hope the autocorrelation would also do this, the correction for autocorrelation is an estimate and we could clearly see the seasonal cycle of the shorter overlap records was aliasing into our results.

We changed the text as follows:
To avoid aliasing of seasonal differences into our drift computation, we first remove any seasonal signal in the differences. We thus require at least 24 months of overlap to account for seasonal variability and compute drift, leaving only two Tier 2 instrument pairs with sufficient

overlap, NOAA 11_am/NOAA 14_pm and NOAA 16_am/NOAA 19 (in our total ozone analysis only 1 Tier 2 instrument pair was used).

P7, L1-12: The procedure discussed here is reasonable to first order. However, I have strong concerns that the assumption of Gaussianity to represent drift and offset uncertainties (given the examples presented in Fig 5 and 6) is clearly inadequate. It would be better to sample directly from the distribution itself or, further, the joint-distribution between drift and offset, since these are not-independent quantities; the additional use of a Markov Chain would correctly sample the joint distribution and provide better uncertainties. While perhaps infeasible to address at this point, this would perhaps be something at least worth mentioning and considering/proposing for future improvements in the merging procedure (and perhaps in the conclusions; see below). As it stands, the current approach considered perhaps leads to rather conservative uncertainty estimates that might be reduced with a procedure more akin to that mentioned here; such conservative uncertainties propagate into the linear regression approach taken (e.g. P8, L20–21), as made clear in the analysis section.

We agree with the reviewers that the offsets and drifts are clearly not Gaussian, and we have begun work trying to address these issues. We have purposely erred on the conservative side, and do hope to refine the error bars in future efforts. We have added a paragraph as suggested in the conclusions noting these issues and solutions we have considered. We do note however that the range of differences between the individual SBUV records and the NOAA merged product are generally in keeping with the size of the uncertainty.

P7, L1-12: This procedure would benefit from a brief schematic showing the steps taken (e.g. as in Fig. 11 of Laine et al., 2014); but this is simply a suggestion for clarity.

Rather than add a figure or schematic, we suggest the following text, which refers to the schematic in the Frith et al., 2004 paper and more thoroughly describes the small differences from that approach in a step by step fashion.

We follow closely, though not exactly, the approach illustrated in Figure 7 of Frith et al. [2014]. The difference here is we treat the drift and offset separately, rather than applying a drift and offset simultaneously, as was done in Frith et al. [2014], Figure 7a.

Specifically, we went through the following consecutive steps:

- Step 1: apply a drift uncertainty to each instrument. This step results in traces similar to Fig. 7a of Frith et al. [2014] but each trace starts on the zero line;
- Step 2: inter-calibrate the drifting records from individual instruments using two reference instruments - NOAA 11_pm and NOAA 17. In this step we repeat the instrument calibration process used in the algorithm (DeLand et al., 2012; Frith et al., 2014, Figure 7b). Through this adjustment we induce the time dependence we expect from the internal calibration process, but remove instrument to instrument offset that is solely due to drift in one or more

instruments, because this offset is also a component of the offset distribution and we want to avoid double counting uncertainties;

- Step 3: add a bias uncertainty to each instrument. By adding the bias uncertainty after the calibration (Step 2) we avoid removing the offset through calibration and thus more realistically reflect the offsets that exist after the internal calibration process;
- Step 4: average time series from individual instruments into a single simulated time series.

The steps described above are repeated 10,000 times to form the distribution of potential error due to the merging process. We note that in our total ozone analysis we used Dobson ground-based data to estimate the precision of the calibration between the beginning and the end of the record (Frith et al. [2014], Figure 7d), but we eliminate this step in the profile analysis because we do not have a comparable correlative profile ozone record that we believe is stable over the full time period.

Fig. 7: It would be instructive to mark the satellite periods (perhaps with horizontal lines/bars) to indicate satellite periods and am/pm drifts, making clearer to the reader the source of drifts/jumps.

We have added notations as suggested.

P8, L29-30: “Smaller ... (1999).” So what does this physically imply? Is it that chlorine related recovery is not strong here, or just that the relative increase is smaller due to higher background levels? (i.e. are absolute changes are similar?).

It is a result of the smaller degree of ozone loss in the early period such that the recovery is also less, as shown also in the bottom panel of Figure 4. This vertical structure has been observed in various satellite and ground-based records.

We changed the text as follows:
Smaller trends at this pressure level are consistent with the vertical trend structure observed across different satellite and ground-based systems reflecting smaller ozone losses relative to layers above and below (e.g., Randel et al., 1999; Harris et al., 2015).

P9, L15-16: When both MOD and NOAA datasets use only Nimbus7, why is there a scatter in the data in Fig 10 (i.e. prior to 1989)?

This scatter most likely results from minor differences in how the monthly zonal means are constructed from the individual profiles. There are choices to be made based on several flag values included with each profile, and other quality measures may or may not be used to further identify and flag suspect measurements. Differences in filtering criteria as well as other potential differences in approach such as the exact way we define the latitude bins can all contribute to these small random variations.

P10, L13-14: As you very nicely explain later in the paragraph, fitting EESC is not giving you the trend. Therefore, this sentence sounds slightly misleading that by using an EESC term in the regression analysis, you get a (more) significant trend; but isn't this simply leading to a more significant estimate of the EESC component. Please rephrase this to ensure this isn't misinterpreted.

We have rephrased as follows:
Despite the insignificance of the derived trends from a linear fit to data from 2001, a fit to EESC over the entire record is statistically significant at all latitudes.

P10, L24-26: Perhaps it would be good to address the major point above here, in terms of future approaches to merge data and account for uncertainties.

We agree, and add the following text:
Significant efforts are underway within the ozone trend community to properly characterize the errors associated with merged ozone records. Approaches based on comparing individual data sets directly (Hubert et al., 2016; Harris et al., 2015; this work) indicate larger uncertainties than are suggested based on the spread of the derived trends from multiple merged records (Steinbrecht et al., 2017). While our Monte Carlo approach worked well for total column ozone, the larger differences in the profile warrant investigation of a more complex means of distributing the uncertainties; the use of wide Gaussian distributions to represent the actual distributions shown in Figures 5 and 6 likely lead to overly conservative error estimates. We know for example that the biases and drifts tend to be correlated from layer to layer and largely cancel in the total, putting a constraint on the potential offsets and drifts. We are also testing the sensitivity of the derived uncertainty to treating each instrument separately (as is done here for the two pairs of Tier 2 instruments) rather than combining multiple instrument pairs into a single distribution. Nevertheless the differences with the NOAA Cohesive dataset suggest the uncertainties modeled here are not overly exaggerated, and that we still need more data to resolve these issues.

Minor points (grammar etc.): P10, L7: Nosier -> Noisier
Corrected