Response to reviewer 2:

We would like to thank the reviewer for the useful comments and suggestions. The constructive criticism provided and the changes we have made in response have hopefully improved the paper. We have carefully considered the reviewer's suggestion regarding redirecting the paper to AMT. After much thought we felt that we would like to keep the paper in ACP(D), largely because although the paper is quite technical, it does not in itself describe a specific new measurement technique or a comparison of measurement techniques. Rather, the aim of the paper is to compare several new datasets extending the ozone profile record, especially within the context of longer-term trends. We have tried our best to address all of the reviewer's comments, particularly in terms of providing more discussion of relevant existing publications. We have, however, to some extent limited this so as to avoid too much overlap with results to be presented in the third overview paper of the SI2N special issue (Harris et al., to be submitted to ACPD). Reviewer comments are shown in italics, while the author responses are shown in grey and changes made to the manuscript in bold blue and red (changes highlighted with *latexdiff*).

p25693 l8: Here it is not distinguished between vertical resolution and vertical grid, although these are not always the same. As a consequence of this, the subsequent statements remain vague.

Thank you for highlighting this. We have clarified the statements to refer only to the vertical resolution.

The different vertical and horizontal ~~grids~~ resolutions of each data set also need to be taken into account; adding data ~~from a relatively low-resolution grid/profile~~ with a relatively low resolution to a high-resolution ~~grid/profile~~ data set involves either a degradation of the high-resolution measurement or the need for additional information to justify interpolating the low-resolution product to a higher resolution ~~grid~~.

p25693 l17: For a technical comparison paper focused on data characterization it may be justified to restrict the comparison to data sets which have a common technical characteristic, e.g. the same temporal coverage. For a scientific paper, however, the results have to be discussed in the context of all relevant existing work. E.g. for linear trends the shorter temporal coverage of a data set is not a good excuse to ignore it, because it is inherent to the assumption of a linear trend that the trend does not vary with time and thus trends determined from shorter data sets should be comparable to those inferred from longer data sets. The same applies to annual variability: The assumption that the amplitudes of the annual variation and its overtones is assumed time-independent implies that comparison to annual cycles of shorter data sets should still be meaningful.

We very much agree that results need to be discussed within the context of all relevant work. To this end we have included more discussion throughout the text (see also further comments below) to provide more context. However, the main aim of this paper is to compare long-term ozone profile data sets that have recently been produced to extend existing satellite measurements; to date, this has not been done in a comprehensive way. The comparison of a large number of individual ozone profile satellite measurements was thoroughly completed by Tegtmeier et al. (2013), so we did not feel this needed to be repeated here (although we have now made more reference to the results from this work). Using the multiple linear regression method we have applied here (which has, to date, commonly been used in the literature), one needs to be quite careful about the choice of start and end dates of the trend periods considered, even though, in theory, looking at different portions of the same trends should give the same results. This will be further discussed in a paper soon to be submitted to ACPD (Harris et al., 2015, the third Si2N overview paper); essentially, this paper shows that shifting the start/end point of the piece-wise linear regression can affect the calculated trends. The Harris et al. paper also considers trends from all measurement systems (including shorter-length satellite as well as ground-based observations) providing a more complete overview which we did not wish to repeat here.

p25694 l7: I do not understand the dichotomy ("either or"). I think that hybrid approaches are possible and thus suggest to delete "either".

We did not intend to suggest that only two ways of merging data are possible; indeed there is a wide range of approaches. We have removed the 'either' to make sure that this is clearer.

In principle, data sets can be merged by ~~either~~ combining data from a series of instruments of the same type, or by merging data from instruments of different types.

p25695 l22: The resolution increases to 15km. It is the resolving power which decreases. Since the correct technical wording is admittedly counter-intuitive I suggest to use an unambiguous verb here, e.g. "degrades". The same applies to line 26.

Thank you for pointing this out. We have changed 'decreasing' to read 'degrading' in both cases.

The SBUV instrument has a resolution of 6-7 km near 3 hPa, ~~decreasing~~ degrading to 15 km in the troposphere (Bhartia et al., 2012). Thus SBUV reliably measures the partial column of ozone from the ground to the lower stratosphere, but must use a priori information, which does not include a trend component, to resolve the signal within that range. The SBUV resolution is also somewhat reduced in the upper stratosphere, ~~decreasing~~ degrading to ~10 km above 1 hPa.

p25696 l2: Here broad averaging kernels are mentioned. Averaging kernels do not only carry information on the vertical resolution of a profile but also on the content of prior information in the data set. I consider this kind of information important: Depending on the choice of the prior information of the retrieval (constant or time-dependent) annual cycles, trends, etc. can be damped. Thus knowledge of these issues might be crucial to understand related differences discussed in the later sections.

We have added details on the SBUV a priori to the paragraph above, where without introducing the term 'averaging kernel' we describe the blending of measurement and a priori which is characterized by the averaging kernel.

Thus SBUV reliably measures the partial column of ozone from the ground to the lower stratosphere, but must use a priori information~~, which does not include a trend component,~~ to resolve the signal within that range. The V8.6 SBUV a priori derives from an ozone climatology constructed from AURA MLS and ozonesonde data which varies seasonally (monthly) but has no trend or interannual variability component [McPeters and Labow, 2012; Bhartia et al., 2012].

We limit the vertical range of our analysis of the SBUV data sets to pressures <20 hPa in the tropics (20°N-20°S) and to pressures <30 hPa outside the tropics to exclude regions where the ~~application of the broad SBUV averaging kernel~~ low vertical resolution of SBUV may affect derived trends.

p25701 l4: With the multiple occurrence of SAGE the MDM is certainly not an "unbiased estimate".

We agree. Given that SAGE-II is used in five of the seven datasets, there will be a 'bias' towards SAGE-II in the MDM. The wording in this sentence has been changed, no longer calling the MDM an 'unbiased estimate'. We did feel, however, that attempting to produce an 'unbiased' MDM was beyond the scope of this paper, given the uncertainty in independence of all of the data sets (individual and merged).

A multi-data set mean (MDM) is constructed to provide a common point of reference to compare the data sets. The MDM is by no measure the best representation of ozone but provides ~~an unbiased~~ a simple average of all available data, rather than favouring one particular merged data set. It is calculated by ~~simply~~ averaging all available data for each time step and latitude/pressure bin, with no weighting applied. A weakness of the MDM is

that five of the seven data sets averaged are based on the SAGE-II record, therefore despite some of the data sets either using another instrument as reference or including other observations during the SAGE-II period, the MDM is largely dominated by the SAGE-II signal from 1984-2005.

p25702 l13: Is it adequate to talk of Fourier terms in the context of all coefficients A-H? Shouldn't the attribute "Fourier" be limited to the harmonic components? The treatment of seasonality is not quite clear. I might have missed the point here but a clearer and more detailed description of the model would be helpful.

We agree that it is not adequate to talk of Fourier terms for all coefficients since they are only used for coefficients A-E. We have adapted the equation to no longer include these subscripts and have changed the text to better describe this point. We have also tried to clarify the text describing the model.

Fioletov (2008) provides further details and a general overview of how each of the processes described by these proxies affect ozone variability. Equation 1 presents the simplest form of ~~the model~~ how the model is applied: …

… where Ω_t is the ~~model calculated~~ ozone for a particular month *t* for a particular data set; A-H are the model coefficients corresponding to ~~the annual cycle offset term~~ an offset term (to account for the annual average ozone amount), linear trend, and other basis functions used; while *R(t)* represents the residuals (difference between the measured and statistically modelled ozone values). The subscript on each term A-H indicates how many Fourier pairs the term was expanded into to account for ~~seasonality~~ the seasonal dependence of the basis functions on the ozone anomalies (Bodeker et al., 1998); for example NB=2 indicates two Fourier pairs (two sine, two cosine). An autoregressive model is applied to the residuals *R(t)* following equation 2: …

Eq 1: I do not understand why R(t) appears in the equation which is supposed the MODELED data. Is the residual deterministic such that it can be predicted by the model? I think the residual is the difference between the measured and the modelled data and thus should NOT appear in the equation defining the model.

Equation 1 describes how the multiple linear regression model is applied rather than the statistically modelled ozone value. In this case the form of the equation, including the residuals (R(t)), is correct. The text has been adjusted to reflect this.

(See the changes shown in the comment above).

p25703 l17: I do not understand this sentence, particularly what does "not coefficient A" mean?

The intention was to indicate that the seasonal cycle is calculated directly from the data and not the seasonal cycle derived from the regression. The text has been changed to state this more clearly. (See also the reply to reviewer 1).

Figure 3 presents the annual cycle averaged over 1984-2011 for three selected levels in the three latitude regions~~,~~ (i.e. the climatological average, not ~~coefficient A~~ the annual cycle derived from the multiple regression model).

p25704 l9: The statement explaining the similarities of the annual cycles is quite vague ("This is perhaps due to ..."). The issue of having merged data sets relying partly on the same parent data sets needs a more thorough discussion. A more quantitative estimate on the reduction of differences due to this issue is needed. The statement as made in the current version, viz. that the occurrence of the same parent data sets in multiple merged data sets reduces the differences appears quite trivial to me.

The section on annual cycles was rewritten to both provide a more scientific context to the results presented (as suggested by the reviewer – see comments above and below) as well as to address the issue of providing a more quantitative estimate of the similarity between data sets.

Figure 3(a) shows the annual cycle of ozone in the northern mid-latitude upper stratosphere, with peak values during the winter months (from November through February) and minimum values in the summer (May through August). The annual cycle in this region is largely determined by catalytic ozone destruction, which peaks during the summer months, resulting in maximum values occurring in winter (Brasseur and Solomon, 2005; Perliski et al., 2013). ~~**In fact, in all regions and at all levels the merged**~~ **1989). All seven data sets show** ~~**smaller differences**~~ **similar annual cycles, both in terms of** ~~**annual cycles than seen on an instrument-by-instrument basis by Tegtmeier et al. (2013). This is perhaps due to the effect of having averaged multiple data setsto produce each merged data set.**~~ **phase and amplitude, and mostly lie within the ±1 standard deviation range shown by the grey shading (the standard deviation across all data sets gives an indication of the spread between data sets). However, in terms of absolute values, there is quite a spread between the data sets. The SBUV Merged Cohesive data set shows consistently lower values for all months of the year while SAGE-OSIRIS shows the opposite tendency, with mostly consistently higher values, although for only five of the twelve months the ±2 standard error bars do not overlap with the standard deviation range of all data sets. SWOOSH, GOZCARDS, and SAGE-GOMOS2 show remarkably similar annual cycles, with mean values not significantly different from each other in all months.** ~~**In general, mean annual cycles agree best in the southern mid-latitude middle stratosphere (Figure 3(c)), while there are larger differences between data sets in the equatorial lower stratosphere**~~

In the tropical lower stratosphere, where the ozone seasonal cycle is essentially determined by vertical transport associated with tropical upwelling, the peak ozone values from July through October (Figure 3(b)) ~~**and the northern mid-latitude upper stratosphere (Figure 3(a)). Although, in nearly all cases, the one-standard deviation error bars, indicating variability within each data set, overlap between**~~ **correspond to the months when upward transport of ozone-poor air from the troposphere is at a minimum (Randel et al., 2007). As for the upper stratosphere, differences between data sets are significant in terms of absolute values, as seen in the large ±1σ range (grey shading), which ranges up to 0.2ppmv (~15 %). SWOOSH and GOZCARDS are again most similar to each other for most of the year, and show consistently higher values than the other data sets.** ~~**SWOOSH and GOZCARDS show almost exactly the same annual cycles in all three regions, and the same can be said of the two SBUV data sets in the northern and southern mid-latitudes (as**~~ **SAGE-GOMOS1 and SAGE-OSIRIS show lowest values, with the latter having mean values that fall outside of the ±1σ range in nearly all months. The three data sets that extend the SAGE-II record with just one data set (GOMOS or OSIRIS) have considerably more missing data in this region of the stratosphere, with less than half of the data available for certain months of the year (where no data are shown, see caption Figure 3). This is at least in part because of data being filtered out after the eruption of Mount Pinatubo. As mentioned above, the SBUV data sets are not shown** ~~**at the 50 hPa level in the tropics,**~~ **at this pressure level.**

In the southern mid-latitude middle stratosphere nearly all data sets agree remarkably well with each other throughout the year (Figure 3(~~**b**~~**c)).** ~~**SAGE-GOMOS1 indicates slightly higher values than SAGE-GOMOS2 in both the northern mid-latitude upper stratosphere and tropical lower stratosphere, although for all months the error bars between data sets**~~

~~overlap. The differences between the two SAGE-GOMOS data sets are likely related to two factors:~~ In this region of the atmosphere, the annual cycle is opposite in phase to that in the upper stratosphere ~~the different treatment of the diurnal cycle of ozone likely plays a role, since SAGE-GOMOS1 and SAGE-OSIRIS tend to be most similar despite SAGE-GOMOS1 using GOMOS as reference and SAGE-OSIRIS using SAGE-II as reference; in the tropical~~ lower stratosphere, ~~on the other hand, it is more likely the different choice of reference instrument that plays a role (SAGE-GOMOS1 using GOMOS and SAGE-GOMOS2 using SAGE-II). SAGE-OSIRIS is the~~ (Fioletov, 2008; Perliski et al., 1989), with peak ozone values in the summer (October through February) resulting from photochemical ozone production during this season (Perliski et al., 1989). The only data set that ~~does not show a consistent offset compared to the other data sets, but rather differences that vary from month to month. In the southern mid-latitude winter months, the error bars do not overlap from July to September, when the~~ shows significantly different ozone values for much of the year is SAGE-OSIRIS ~~values are up to 0.8 ppmv~~ , which has values up to 1ppmv (~15 %) higher ~~than all other data sets~~ in the austral winter season and slightly lower values in February and March. This feature is also evident in the northern mid-latitudes during winter (not shown), although to a lesser extent, and is likely due to the reduced sampling of ~~OSIRIS~~ the OSIRIS instrument in the winter hemisphere (see Figure 2), which ~~may~~ seems to quite strongly affect the mid-latitude zonal mean values~~. In the tropics at 50 hPa, the~~ in the merged data set. Overall, in the three regions considered, the seven merged data sets show similar annual cycles, particularly in terms of phase and amplitude. Furthermore, with the exception of the SAGE-OSIRIS ~~error bars are large and overlap in all months with those from all other data sets.~~ data set, the biases between data sets are largely consistent throughout the year. Agreement is best in the mid-latitude middle stratosphere, while there are larger differences between data sets in the tropical lower stratosphere and in the upper stratosphere globally; results which are very similar to that shown on an instrument-by-instrument basis by Tegtmeier et al. (2013).

p25707 l14: I disagree with the use of the standard deviations: The standard deviation is a measure of the expected deviation of, e.g., one particular February from the mean of all Februaries, i.e. it is a measure of the variability of the February-value. For comparison of multiple averages the adequate diagnostic is the standard error of the mean (for uncorrelated data it is the standard deviation divided by the square root of the sample size). To judge how well the mean annual cycles agree, the latter would be the applicable quantity (with the caveat that the multiple use of parent data sets in some merged data sets adds further complication). I do not understand what can be concluded from the fact that the standard deviations overlap.

The intention of showing the standard deviations was to present how variable ozone was for a particular month. We agree that we should not have over-interpreted this to judge how well the datasets agree. We have changed figure 3 to show both the standard error of the mean for each data set as well as the standard deviation across all data sets to show how variable ozone is for a particular month. The text has been changed accordingly (see below and also above for the previous comment).

The vertical error bars indicate ~~the~~ ± two standard error of the mean for each individual data set, while the grey shaded region indicates the one standard deviation range (mean value of all data sets ± the mean of the standard deviations, which are calculated for each data set individually before averaging). ~~Averages~~ The standard error of the mean provides a useful approximate measure of the uncertainty of the mean, although it may not represent the true uncertainty since individual samples of the population may exhibit autocorrelations (Toohey and von Clarmann, 2013). The standard deviation represents the ozone variability for each month, averaged across all systems. Averages, standard error,

and standard deviations were only calculated for months that had data for more than ~~20~~ 14 of the 28 years available for analysis.

Summary Section 3.1: This section is restricted to the annual cycle as a technical diagnostic of the merged data sets. Is there nothing to say about the annual cycle in terms of atmospheric sciences? Are the annual cycles as represented by the merged data sets in agreement with our expectations and with other analyses found in the literature? In other words, can the annual cycles be explained with our current knowledge on atmospheric processes? Publication in ACP requires discussion of such issues beyond the pure technical description.

In response to this comment we have rewritten section 3.1 (see above two comments).

Summary Sections 3.2 and 3.3: Here the same applies as for Section 3.1. The discussion is limited to the technical comparison. For an ACP paper the anomalies should be identified and as far as possible attributed to certain events (e.g. Pinatubo). A statement is needed which anomalies can be explained with current knowledge and if/where any unexplained issues are detected. This does not mean that the authors have to carry out quantitative analyses or model calculations themselves, but at least the data sets under investigation (and the events recognized in them) shall be put into the context of the existing literature. Although dealing mostly with column ozone, Shepherd et al (Nature Geoscience 443-449, 2014) might be useful in this context.

We appreciate the reviewer's point of view and have made a large number of changes throughout the paper to provide more scientific context. In this respect, we also made changes where we felt relevant to section 3.2 and 3.3 [we have not copied both entire sections here for the sake of space].

p25711 l14: How is significance defined in this context, and how is significance evaluated, facing the multiple occurrence of certain parent data sets in the merged data sets which are then averaged?

The trends are considered if their 2-sigma uncertainties exclude zero. This is specified in the text (see below) and in the caption of Figure 8. Evaluating the significance of the MDM trends given that the datasets are not independent from another is challenging; it is difficult to state quantitatively how much this 'non-independence' affects the significance of the trends derived from the MDM. We have changed the wording of the text to further highlight that the trends derived from the MDM should be taken only as a simple trend derived from the mean data set. The issue of addition of uncertainties from multiple datasets is further discussed in Harris et al. (to be submitted, as mentioned above). We have also included a reference to this work in this section.

We also calculate the uncertainty associated with each trend estimate based on the variance in the residual time series and present the 2σ uncertainties on the trends throughout this paper.

Summary Section 4: In the last paragraph of this section, the authors put their data in the context of independent work. This is certainly a step into the right direction but this discussion needs to be extended to make the manuscript suitable for ACP. More detailed suggestions follow below:

The manuscript has been reworked to include more scientific context and to extend the discussion to cover more results from other studies.

p25713: I disagree with the explanation that the discrepancy between the trends found by Gebhardt et al. and those assessed here can be attributed to their shorter data set. As said above, the assumption of a linear trend implies that consistence with trends inferred from subsets of the period should be expected (when represented in Delta vmr/Delta t instead of %/Delta t). Further, Eckert et al. (ACP 14, 2571-2589, 2014) find trends in a similar short period, which fit much better to those trends discussed in the paper. This suggests that the discrepancy w.r.t. Gebhardt et al. is not to be explained by the shorter

time period but that it is a particular characteristic of this data set. The work by Eckert et al. should be included here, not only because it belongs into this context per se, but also because it helps to solve the problem with the comparison to the Gebhardt trends.

We thank the reviewer for this comment. We have included discussion regarding the work of Eckert et al. (2014) in the context of the equatorial trends, in particular in terms of the differences with respect to the trends estimated by Gebhardt et al. (2014).

Gebhardt et al. (2014) also found large negative trends in this region of the middle stratosphere (~~up to~~ as large as -10 %/decade) ~~at altitudes from about 32-38 km, however, they consider a shorter period of observations from SCIAMACHY covering~~ for the 2002-2012 ~~so direct comparison is difficult. In~~ period in SCIAMACHY observations. However, in a comparison for just the 2004-2012 period they show that the SCIAMACHY trends are considerably more negative than either Aura MLS or OSIRIS. Using MIPAS data for the 2002-2012 period, Eckert et al. (2014) show negative trends of similar magnitude to Kyrölä et al. (2014) of approximately -5 %/decade, although somewhat lower in the middle stratosphere. While there is a clear negative trend in many data sets, the range of trend estimates remains large. Finally, in the tropical lower tropical stratosphere, trends are insignificant for nearly all ~~data sets, as for the other regions, because~~ the data sets shown here, largely as a result of the large uncertainties in this part of the atmosphere.

The main results of the additional discussions to be included in Sections 3 and 4 should be summarized in the Conclusions and the Abstract.

The conclusions have been adapted to reflect the changes made throughout the text.

The figures are too small (particularly, in Figure 1 the structures can hardly be seen, in Figure 4 and 8 the lines are hardly discernible. All labels are very small.

All figures have been modified to make features of interest more visible and to make labels clearer. In regard to Figure 4, the original figure already has the datasets separated in each sub-plot so as to make them more visible. We felt that further dividing the datasets would make it more difficult to compare the datasets directly. The more the lines overlap the better the datasets agree, and unfortunately this means that it is difficult to tell each dataset apart. The same is true for Figure 8.

My critical review results from the fact that the current content of the paper is much more suitable for AMT than ACP. After consideration of the issues discussed above, the paper to my judgement still has a fair chance to meet the criteria for publication also in ACP. I recommend publication in ACP after major revision. Redirection to AMT (which is also linked to the special issue on SI2N) would be the more straightforward option but I guess the decision in favour of ACP has already been made.

Please see the overall response to the reviewer at the beginning as well as responses to the comments above.