Author reply to Referee #3

Ground-based assessment of the bias and long-term stability of fourteen limb and occultation ozone profile data records

Hubert *et al.*, Atmos. Meas. Tech. Disc., 2015, 8, 6661-6757.

We would like to thank you for the effort of digging through our lengthy paper. Our response to your questions, comments or suggestions can be found below, with different text formatting for referee comments and author replies.

The manuscript presents a comprehensive assessment of fourteen limb and occultation satellite ozone sensors by comparison with ground-based ozone sondes and lidars. The main focus of this study is to evaluate the long-term stability of the ozone profile datasets. By using a network of ground-based instruments as a reference, authors estimated relative drifts and the range of uncertainties for reliable detection of drifts in the satellite data records. These results can be used for interpretation of trend estimates from individual and merged satellite records.

The paper is well written and fits to the scope of problems discussed in AMT. The paper indeed is very long and contains detailed descriptions of the datasets and methods, used in the study, as well as a throughout analysis of the obtained results. However, the referee feels that all these details are important and relevant for the presented discussion. The paper is recommended for publications after minor changes.

Specific comments:

P. 6679, l. 17: What is "the nominal mode" here? Is it the same as OR phase?

> This can be confusing, indeed. Resolution mode and measurement mode are not identical. Each measurement mode has another vertical scan pattern, with different vertical range and step sizes between scans. This is now clarified.

P. 6684, l. 20: The negative ozone values were removed from the analysis in this study. Even though the negative values are unphysical, but when the ozone concentration is very low and natural variability are high (e.g. like UTLS region), a retrieval algorithm can produce negative values. Moreover, these negative values are recommended to be included in the bias estimation or any other statistical analysis. By removing negative values you shift the means to higher positive values, and biases will become more positive relative to other instruments as well.

> You are right, and we will change that in our future analyses. However our investigation showed that there is (almost) no impact on our current results, for two main reasons. First of all, there are almost no negative ozone values in the correlative data files, possibly as a result of screening by the data provider. Second, the use of robust statistical indicators reduces the influence of negative outliers on the outcome. And since there are not too many outliers the final results are nearly identical.

P. 6685, l. 2: what was a motivation for selecting v=100km/h? Please, explain.

> This was taken as a rough estimate of horizontal wind speed in the stratosphere. Now explained in the manuscript.

P. 6685, l. 15: Why did you make an exception for MIPAS and used the MIPAS AKs instead of the triangular function? Do MIPAS AKs significantly differ from the triangular shape? Do they change over time?

Our objective was to use the satellite data as recommended by the data providers, which would/should correspond to how users are treating the data in their analyses. There is one exception to this guiding principle: SCIAMACHY. While the Quality Working Group recommends the use of the provided vertical AKs, we found that these AKs degraded the comparison results in an unexpected way (as described in text). The MIPAS AK-smoothed comparisons on the other hand appear very reasonable, and where hence used. These AK are profile-dependent and could indeed change over time. However, this was not studied in detail for this paper.

P. 6690, l. 16: What do you mean here as "average single station drift uncertainties"? Is that a simple mean of uncertainties derived for each station?

Indeed, this is the average drift uncertainty over the ensemble of individual station estimates. We rephrased the manuscript.

P. 6690, l. 18: I don't think that drifts for SAGE II relative to lidars and sondes "are very consistent". There are significant differences between blue and black lines especially between 15 and 22 km.

The sonde and lidar-derived drift estimates agree within the error bar, so there is not statistically significant. We replaced "estimates are very consistent" by "estimates are statistically consistent", to make this clearer.

P. 6690, l. 21-22: I agree that the drift is slightly negative relative to sonde data between 20 and 40 km, but not relative to lidars. Please, specify that in the text.

There are indeed positive excursions for lidar-derived drift, which is why we originally phrased as "drift is *generally* slightly negative". This statement is now toned down in the revised version "average drift is slightly negative except around ~33 km".

P. 6691, l. 21: It is not clear here what you meant by "lower and higher altitudes", please, specify.

Done.

P. 6710, l. 12: Add 'for these instruments' at the end of the sentence "to the detection threshold for these instruments."

Done.

P. 6743, Table 4: For Aura MLS 1-sigma range between 30-45 km is given as 1.5-5%, while in the text on p. 6693 it says 2-3%. I see that most of the other satellite instruments with the comparable length of the record have smaller uncertainties. Please, check if the numbers for Aura MLS given in Table 4 are correct;

This is corrected in the revised manuscript. Thanks for catching the inconsistency.

P. 6746, Figure 1: You show that four satellite instruments extend their measurements down to the surface. However, I doubt that any of these sensors can provide valuable measurements in the troposphere. So my question would be: what is the lowest level you use in the analysis for these instruments? I think it would be appropriate to not show any satellite data below these levels on this figure.

This was also a question by Referee #1. We will remove the grid levels in Fig. 1 that were not used the analysis.

P. 6748, Figure 3: What is "the standard deviation of the single site drifts"? Is that a simple mean from all red and grey error bars? Please, explain.

This is the standard deviation of the ensemble of drift estimates from the ozonesonde network. This is now clarified in the caption.

P. 6749, Figure 4: The grey line for SAGE II is very difficult to see. Please, consider to use another color.

We will darken the tone for the SAGE II result.

P. 6750, Figure 5: This figure is a central piece of the manuscript. On my opinion individual plots on this figure are too small. I suggest to divide this figure on three parts (5a, 5b and 5c) and show no more than 4-5 plots on one page. For example, you can divide instruments based on the length of the record, and show instruments with more than 10 years of the record on one page (SAGE II, Aura MLS, OSIRIS, HALOE and SMR) and instruments with >5 years on another, or find any other way to split these plots. But the bottom line is that it is difficult to see details with 14 plots on one page. And I believe that these plots with the instrumental drifts are very important and deserve readers attention. They need to be enlarged.

The layout formats of AMTD and AMT are very different. All figures were optimised for best readability in the AMT layout printed on A4. We will therefore keep the figure as it is, presenting all drift results on a single page.

P. 6751, Figure 6: I have the same remarks here as for Figure 5: split this figure into 3 and show no more than 4-5 plots per page. The current size of individual plots doesn't allow to see any details, and make it very difficult to find a correspondence between the text in Section 5 and Figure 6.

See our reply to previous remark. The figures are larger and readable in the final AMT layout when printed on A4. We will therefore keep the figure as it is.

P. 6755, Figure 10: This figure is very crowded with 10 plots and 14 lines. Also, I think that the color selection doesn't work very well here. It is very difficult to distinguish lines for ACE and MAESTRO, or HALOE and GOMOS. You might try to split this figure in two, and show 6 instruments that ceased prior 2006 on one page, and all recent instruments on another.

The take-home message of Figs. 10-12 is the ensemble, rather than the individual satellite results. The latter are presented in Fig. 5 for drift. We will add similar, more quantitative figures for bias and comparison spread of single satellite records in appendix.

Minor comments:

P. 6664, l. 10: I think the word "harmonized" doesn't fit very well here. Perhaps, "consistent" would be better.

Indeed, "consistent" may be a better phrasing for many readers.

P. 6670, l. 8: Shouldn't it be "artifacts" instead of "artefacts"?

We adopted British spelling throughout the manuscript, so "artefacts", "-ise", …

P. 6695, l.14: Since you refer readers to the next section it might better say "we will show";

Ok, done.

P. 6703, l. 12: It might be better to replace "a more important" with "a stronger".

We rephrased as "a more pronounced".