Author reply to Referee #4

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.

Thank you for your commitment to go through our paper, we highly appreciated your review. Our response to your questions, comments or suggestions can be found below, with different text formatting for referee comments and author replies.

The manuscript provides a consistent evaluation of fourteen different satellite limb and occultation instrument data records, although the results are varied in significance because of the short coverage or sparse sampling of some of the records. The results are well-structured and well-referenced and apply good statistical analysis. The paper is well-written and logically ordered. It is very long as can be expected given the large number of records presented. Some specific strengths and weaknesses are discussed below.

The comparisons were made to a selection of ground-based lidar and in situ ozonesonde records. It would have been good if the authors had included groundbased microwave records as well, especially Lauder and Mauna Loa where there are both Lidar and microwave instruments with long records. It is probably not practical to go back and include those stations for this paper but estimates of the Lidar record consistency with the microwave one when both have long records at the same station are probably accessible to a co-author familiar with a given Lidar instrument. (E.g., are the upper stratospheric biases mentioned on Page 6708 line 17 seen in Lidar versus ground-based mircrowave?)

We will include a discussion of the agreement between lidar and microwave instruments .

As noted on Page 6684, line 24 et seq., selecting a matchup criteria is always a tradeoff between increasing the number of data points and decreasing their quality or increasing the matchup noise. Since the east/west gradients are usually smaller or more random (less persistent) than the north/south gradients, one should consider a looser east/west distance criteria than for the north/south one. That is, 500 km of N/S mismatch is not the same as 500 km of E/W mismatch. Given the seasonal pattern of some occultation measurements, there can be corresponding patterns of latitudinal bias when compared with fixed location ground-based records.

One measure that can catch problematic matchups is to use the daily total ozone from a mapping instrument to check for strong gradients in the ozone field (atmospheric inhomogeneity). That is, a quantity of matchup merit can be the difference between the total column ozone values at the station and those at the satellite measurement location. Large values of this statistic can indicate that the two measurements were made for different air masses and give an independent quantity for filtering the matchups. The statistics of matchup differences as a function of distances and total ozone differences can be examined to check the consistency.

This is another interesting suggestion to improve the co-location criteria in future work (see report Referee #2). However, it was outside the scope of the paper to study spatio-temporal mismatch uncertainty in great detail for all instruments. This is an analysis in its own right, and was done e.g. by Verhoelst et al. (AMTD, 2015) in the context of the validation of total ozone columns.

The authors include numerous comparisons among satellite products and between satellite products and ground-based (Sub-orbital) systems. There are some good references to the current results as agreeing or contrasting with previous ones but I had some difficulty connecting all the information. In particular, if two coincident satellite records have significant trends or biases with respect to the ground-based records, one can ask if this is in agreement with direct comparisons of the two records. That is, close the circle. It is usually the case that the number of matchups between two satellite records is much greater than with ground-stations. One can also create and compare zonal means for well-sampled satellite records. Are the inter-record biases estimates (between pairs of satellite records) used to create the merged data sets in Table 6 consistent with the relative differences in the pairs results versus ground-based records given in Table 5?

An extensive evaluation of the consistency between all available validation and intercomparison results was outside the scope of this paper. Such summary of the state-of-the-art of ozone profile comparisons is foreseen for the second overview paper of the SI2N initiative (Lambert et al., AMT 2016). We will refer to the reader to this paper instead.

Unfortunately it is impossible to answer the latter question since the different merging teams have so far not published quantitative information on the (bias) corrections found/used in the merging process. What we can say is that most merging schemes allow for the correction of the types of patterns identified in our analyses, except for the long-term temporal component (i.e. drift).

Make sure that units are given for the entries in all tables. For example, what are the units for Table 6 – %/year, %/decade?

Thanks for the catch. We verified that entries in tables and figures are correctly labelled.

There are too many lines in Figure 12. It would be better to have separate summary figures for the pre and post 2006 instruments. Once that is done, the instrument types could be identified by linestyles and the specific instrument identified by color.

As mentioned in our reply to Referee #3, the take-home message of Figs. 10-12 is the ensemble, not the individual satellite results. Quantitative drift results per instrument appear in Fig. 5.

Why were bootstrap methods (Page 6689, line 1 and Figure 2) used to estimate the trend uncertainties? Do the residuals from the have serial correlation? If so, how much, and were the residuals resampled in sequences of more than one value to capture it? The text says that the time series were sampled not that the residuals were sampled after the fit by a trend line. Please clarify the resampling strategy. One would expect a random resampling of the time ordered time components series to give similar values as the standard analysis even if the data had serial correlation as it is would not be captured in a single element resampling.

The comparison time series have at best daily sampling, in general with many gaps (see Sect. 4.1.1). As far as we know, there is no rigorous way of estimating auto-correlation in the presence of gaps. Auto-correlation is one motivation for using a data-driven method to cross-check the analytical approximation of regression uncertainty. Another is the influence of outliers, especially at the extremities of the time series. The bootstrap analysis merely served as a cross-check to build confidence in the analytical estimates.

The resampling was done on single elements in the original comparison time series (so not the residuals after fit), to obtain an ensemble of outcomes for each regression parameter. This is now clarified in the manuscript.

It is often better to use relative differences of the form 2*(x-y)/(x+y) instead of (x-y)/y when comparing two noisy times series to avoid biases at y extrema. Since you have not looked at the dependence of differences on the estimate size, it is not critical. While the difference plots do not show obvious seasonality, for the longer, well-populated time series it could be useful to see if the trend uncertainties were reduced when an annual cycle is included in the regression model.

We have tried the inclusion of an annual cycle term in the regression, as explained at the end of Sect. 4.1 and illustrated by the orange line in Fig. 2. In general, we did not find a clear reduction in the drift uncertainty with the more complex model.