Reply to anonymous reviewer #2

First of all, we would like to thank the referee for his/her thorough review and helpful suggestions. The following text includes a point-by-point answer (in black) to each of the reviewer's comments (in blue).

The results seem good enough to conclude that this new dataset will be a useful merged global record. My main comments have to do with how much better or worse the new dataset really is, compared to the older merged product (which is mentioned somewhat too quickly), or versus other independent data (also briefly mentioned). Maybe this means providing somewhat more quantitative references to other published work (already mentioned in this manuscript), without repeating all the details. This would help place the work in context for the reader, even if this manuscript is mostly supposed to present the dataset creation and validation results. Also, the Lerot et al. (JGR, 2014) paper describes in a fair amount of detail the procedures used to create the dataset mentioned in this manuscript, and there are some overlapping types of Figures between these two papers. This results in duplication of some of the descriptions and comparisons, and one might expect, therefore, that this manuscript could be even more succinct regarding some of these aspects, while emphasizing the validation/comparisons even more. This is not a huge problem, but nevertheless, some additional thought should be given to this issue, and as this paper is the "third article on the ESA-CCI total ozone ECV", quoting the authors. Having an international team collaboration on such a project is great, but keeping the number of papers describing approach and results distinct enough is also useful, in this world of busy researchers seeking succinct, timely, and clear messages from added publications.

> We agree with the reviewer and have put more emphasis on the validation/comparison results themselves. We have included more quantitative results, in particular related to the long-term stability of the GTO-ECV level 3 product. Furthermore the differences between the old and the new GTO-ECV product have been analyzed in more detail.

Specific Comments:

- Introduction: Since WMO (2014) appeared in December, 2014, the WMO Report (2011) reference could be enhanced by pointing to this newer reference (or mentioning both). The same holds for Section 4, where the "last WMO report" is mentioned (but is no longer the last one).

> We added the latest WMO report (2014) and changed the wording in Section 4.

- pg. 4611, line 12, I suggest changing "an homogeneous" to "a homogeneous".

> Done.

- pg. 4612, line 9: "GOME lost its global coverage". Could you provide more specifics regarding what this means for the subsequent time period? Did the coverage decrease continuously and how many days are needed in the end to obtain global coverage? Or point to a later part of the manuscript if this is better described later on. Not every reader is as familiar with the details as the authors are.

> We added more information (in Sec.1) related to the ERS-2 tape recorder failure and the subsequent loss of global coverage.

- The merging procedure describes offsets that vary with latitude and time, but then mentions that there is no temporal drift in the corrections. Does this mean that using a time-dependence in the corrections is not really needed? Or is what is most relevant in the time dependence short-term in nature, so the long-term drift can be small or zero even with substantial short-term "drifts"? This also seems to assume that GOME does not drift in and of itself (versus "truth"), as it is used as a reference (and is this true?). It is is somewhat difficult to understand from the description provided in this manuscript alone (without maybe reading other related publications) why this procedure was chosen rather than an average of SCIA and GOME during their overlap period. Some more clarifications and motivation would have been useful, besides what the merging algorithm itself does, as there could be more than one approach (and this is indeed a change from the previous method/results).

> The main reason for including a time dependence in the correction factors was to account for short-term fluctuations in the differences (e.g., for GOME-2 induced by the second throughput test in late 2009, see Figure 1). The companion papers by Lerot et al. (2014) and Koukouli et al. (2015) have shown that all sensors behave similarly, i.e. inter-sensor differences are small, and that the decadal stability falls within the 1-3% target requirement of the Global Climate Observing System (Koukouli et al., 2015, their Table IV). We decided to use GOME as a reference data record as it provides the longest overlap periods with the other two sensors. Furthermore, GOME was found to be the most stable instrument over its lifetime before the application of the soft calibration scheme (Lerot et al., 2014).
> We have added some more information.

- pg. 4613, line 13: what is meant exactly by "low time resolution" (please specify the resolution)?

> The time resolution is one year. This information has been added to the text.

- pg. 4616, line 12: please describe more specifically how this standard deviation is obtained? Are the combined (adjusted) datasets used to obtain their combined variability (for what grid size and time step, 1 by 1 degrees and one month?

> The standard deviation is the standard deviation of the monthly mean that has been obtained from the daily gridded data. It is provided for the same grid and time step as the total ozone column. It is not needed to obtain a combined variability because only one sensor is used at a time in the merged product. The explanation has been included in the text.

- It is interesting that the standard error is not obtained from something like the standard deviation divided by the square root of the number of included input points, and a simulation is used for this result. I understand that there is a sampling component that one can try to quantify from such a simulation, but if the measurement errors (noise) is excluded (and maybe it is not), is there not an underestimate of this standard error (or is this a negligible effect)? This also assumes that the model variability mimics the actual variability well enough...

> The “standard error” introduced in Sec. 2.3 has been calculated as the standard deviation divided by the square root of the number of included measurements and multiplied by a factor “r” accounting for spatial and temporal sampling issues of the sensors on a grid cell basis. Only this factor “r” has been obtained from the OSSE. Thus the measurement error is not excluded as it is part of the standard deviation. We agree with the reviewer regarding his concern about the model variability.

- pg. 4619, line 7: not sure what "a limit" means here. Is this a screening criterion? "Optimization reasons" is also quite nebulous. Probably a detail, but if you wish to start to explain this, maybe you should pursue it more fully (or ignore this if it is really not worth the effort, or makes little difference in the end).

“Limit” is a screening criterion (threshold) in this case, i.e. a minimum number of ten (ground-based) measurements per month must be available which are then used to compute the monthly mean. We have deleted the following sentence “The monthly mean … for optimization reasons” as it is not necessarily needed.

- pg. 4620: Statistics for the Northern Hemisphere, you give a bias for the Dobson comparison but no such number for the Brewer comparison. Is that (mean) bias zero (within error bars)?

The bias for the NH Brewer comparison is zero within error bars (0.16 ± 0.66 %). These numbers are presented in Table 4.

- pg. 4620, line 7: you could delete the "and" after "(from top to bottom)".

Done.

- pg. 4620, line 28: is there a significant trend in the differences between the L3 and L2 comparisons (end of parag., "very satisfactory" is vague language)? Overall, should one try to delete outliers in the L3 results to obtain the same trend as the L2 results? It would be nice to see somewhat more quantitative results or descriptions of the agreements (same for the next page on SH results which are in "near-perfect" agreement). Trends will matter to people looking to such datasets for robust longer-term analyses (and you have other references on this topic), so I wish you could expand or refer more quantitatively to another reference regarding this topic. The qualitative language and the (admittedly nice looking) plots can only go so far in terms of convincing the reader as to how useful this new dataset really can be. I am not necessarily suggesting a complete trend analysis, but a slightly more quantitative result would be more useful. Table 4 has some results, but not necessarily comparing L3 to L2 (?).

We agree with the reviewer and have included more quantitative results. Regarding trends in the differences we provide the numbers for the level 3 comparison and compare them to the trends from the level 2 comparison presented in Koukouli et al. (2015).

- pg. 4622, line 15: change "meets easily" to "easily meets".

Done.

- pg. 4627, line 10/11: "CCMs are three-dimensional..." I would delete this well known description.

Done.

- Fig. 3: The caption mentions March as the change-over date, whereas Fig. 2 shows April (please correct one or the other for consistency). If Fig. 2 is correct, would one not state that April is the changeover date (to SCIA data for example)?

Figure 2 is correct; April is the change-over date. We corrected the caption of Fig.3.

- Figs. 4/5: Can you specify what GOME 2A means?

We have added the explanation.

- Fig. 8: there is a lot of white space here, the y axis ranges could be shrunk to help see the results better.

We agree with the reviewer and have altered the Figures accordingly.

- Fig. 9: Not clear that the outliers have been adequately explained in the text - but I gather this is mainly a satellite sampling/gridding issue; would you generally trust the L2 results more (this was maybe not spelled out well enough)?

We think that the origin of the outliers is adequately explained on p. 4620, ll. 14-22. In these cases different sets of days form the basis for the level 3 monthly averages from ground-based and satellite-based data, whereas the level 2 comparison is based on coincident data.
The intention of this paper is not to show which data record (level 2 or level 3) is better, but to show that the transition from individual level 2 to merged level 3 data does not introduce artifacts (induced either by the level 3 algorithm or the merging approach). Figs. 8-11 clearly show that the level 3 comparison closely follows that for level 2, except for the aforementioned outliers. Therefore we conclude that our level 3 data record fulfills the requirements defined for climate applications such as global long-term studies. For these studies gridded level 3 data rather than level 2 data are needed.

- Fig. 10/11: again, I find this hard to read because of all the white space but it appears the authors want to keep one scale throughout; at least the scale could start at -10 rather than -15, it would seem...

We agree with the reviewer and have altered the Figures as suggested.

- I like Fig. 12, but/so something similar for the Brewer results might be worth considering as well (?).

We do have the same plot for the Brewer results, but since Brewer ground stations are available only in the Northern Hemisphere (furthermore there is no station in the latitude belt 10°-20°N), the corresponding contour plot is not that illustrative. Furthermore, no seasonal features were observed for the Brewer comparison (0.16 ± 0.30 %, see Table 4). Therefore we would prefer to show only the Dobson comparison in this case.

- Fig. 13: The "red open circles" are quite difficult to see. Please try using solid black or blue dots or another clearer choice.

The figure has been adjusted for better visibility.

- Fig. 15: please specify the years used in the caption (as done in Fig. 14).

Done.

- Since the patterns between Figs. 14 and 15 are pretty similar (but hard to tell unless one has the numbers or overplots), with dips at high latitudes, should one not conclude that the GDP product was/is closer to the SBUV-MOD V8.6 result and the new CCI results are in poorer agreement with SBUV? Please quantify or address this.

We agree with the reviewer. The mean differences between GTO-ECV GDP and SBUV are a bit smaller than the mean differences between GTO-ECV CCI and SBUV, which are slightly negative in high latitudes and slightly positive in the tropics as indicated in Figure 15. This does not necessarily indicate a poorer quality of the CCI product. The standard deviation of the differences is smaller for the CCI data record than for the GDP data record for all 5° latitude belts. The work by Chiou et al. (2014), indicated that the mean differences, the standard deviations as well as the ranges of the differences between SBUV and ground-based data and between GTO-ECV CCI and ground-based data, respectively, are of the same size.

- It would have also been nice to see how the comparisons versus ground-based data have changed in the new dataset versus the old one (e.g., showing Table 4-type results for the older (GDP) product?). Do such results not exist for both datasets?

For the GTO-ECV GDP level 3 merged product a detailed validation with ground-based data (as provided with this study for the CCI data product) is not available. But the GDP product has been included in the previous WMO Assessment of Ozone Depletion (2010) and is in good agreement with other data records such as ground-based and satellite data (Section 2 of WMO, 2010).

- Table 4: What does the "Latitude" row represent? Are the error bars 1-sigma?

The “Latitude” row denotes the mean values extracted from Figure 8, the latitude dependency plot. All error bars are the 1-sigma standard deviation. As suggested by the other reviewer this row has been renamed to “Latitudinal variation of biases”

References:

Chiou et al., Comparison of profile total ozone from SBUV (v8.6) with GOME-type and ground-based total ozone for a 16-year period (1996 to 2011), Atmos. Meas. Tech., 7, 1681-1692, doi:10.5194/amt-7-1681-2014, 2014.

Koukouli et al., Evaluating a new homogeneous total ozone climate data record from GOME/ERS-2, SCIAMACHY/Envisat, and GOME-2/MetOp-A, under review for J. Geophys. Res. Atmos., 2015.

Lerot et al., Homogenized total ozone data records from the European sensors GOME/ERS-2, SCIAMACHY/Envisat, and GOME-2/MetOp-A, J. Geophys. Res. Atmos., 119, 1639-1662, doi:10.1002/2013JD020831, 2014.