Reply to Anonymous Referee #2 review of manuscript acp-2017-1157

Symptoms of total ozone recovery inside the Antarctic vortex during Austral spring

Andrea Pazmino on behalf of all co-authors

We thank Anonymous Referee #2 for the time devoted to evaluate our work. Your valuable comments have helped us to improve our manuscript. Since MSR-2 total ozone data have become available until the end of October 2017, we decided to extend our study to the year 2017 using SAT and MSR-2 data. Due to this extension, all figures of the manuscript have been revised, except Figure 1 where white cross marks where added to highlight the region considered inside the vortex by the 400 K-600 K classification range. In addition we have noticed that the figure 12 of the original manuscript about the time shift of low values was not very clear. A new figure, Figure 13, has been produced in order to better illustrate the time shift in appearance of low total ozone values within the vortex. Similar conclusions as in the original version of the manuscript were provided. Furthermore the word "Multiple" was added to the title to highlights that different signs of recovery were obtained in this work, , e.g. (1) Significant positive trends of total ozone since 2001 in September and for the first time in the period of maximum ozone depletion (15Sept-15Oct) using MLR analysis on average ozone inside the vortex and Ozone Mass Deficit, (2) Decrease of occurrences of very low ozone values within the vortex and (3) increased delay of occurrence of low total ozone levels in the September 1st – October 15th period.

Please find our answers to your comments (in red):

I concur with much of what the other reviewer articulated, in particular these points from DeLaat's review:
1. "The presence of this exhaustive list of issues and questions would be less of a problem if the paper introduced new concepts or new ideas, but the paper mostly builds on previous work and confirms what other papers have also concluded."

2. "This paper does not address these issues, nor are results put in the context of this work."

3. "The few time series that are looked at are then seen as the truth, every wiggle becomes meaningful, and too much attention is given to the formal statistical significances, whereas structural uncertainties are important as well. For example, we have shown that rather arbitrary choices with regard to the proxies used in the regression have a strong impact on the formal statistical trend errors. We therefore argued that structural uncertainties are much larger than the formal statistical trend errors, which is important for confident statements about whether recovery has started or not."

I especially agree with DeLaat's concerns about the 'structural uncertainties' in this regression analysis, so please address all issues described in his review. In addition, there are other issues below related to ozone data sets that need to be addressed in a revised manuscript. If revisions are made that address both DeLaat's and my review, this paper could be published in ACP.

Please see our answer to de Laat's review to the different points specified above.

Specific topics of Concern

The composite satellite total ozone time series, referred to as SAT. The merging of satellite data sets into a single record is something to be done very carefully. Instrument measurements have bias and drift, and combining data sets in order to extract small trends (i.e., ozone recovery) requires a great deal of care and a good deal of knowledge about each instrument's characteristics and sampling pattern (i.e., coverage). I see no evidence here that any such considerations were used when combining the data sets. In fact in Figure 4, the difference between the assimilated ozone time series (MSR) and the SAT shows big jumps! There is a large trend from 1990-2005. Does this represent an unphysical trend (i.e., changes in the observing system) in the assimilation, or is this coming from how the individual data sets in the SAT were merged? Have you tried your trend analyses on the 5 merged ozone data sets referenced in Weber et al. [2017]? Without any discussion or justification of how the data sets were merged in this study, I don't see how the trend results presented here (and especially their uncertainties!) can be taken seriously.

Since MSR-2 data are based on the assimilated satellite ozone time series already corrected from offset, trends and variations of solar zenith angle and temperature in the stratosphere, we would like to consider in addition in our work a satellite datasets commonly used in ozone studies (including recovery) with similar kind of instrument (TOMS and OMI) and similar retrieval; without applying any correction. Since SBUV data are sparse, we have decided in the revised version to fill the 1993 – 1995 gap years with MSR-2 data, but taking into account the same spatial coverage as that of TOMS and OMI instruments (see new Fig. 4). Finally, due to the important differences observed in September, particularly the unexplained trend in the 1990-2005 period that you mention, we decided not to include SAT datasets for trend retrieval in that month. Discussions about this difference as well as the issue of spatial coverage of the SAT data is discussed in the Sect. 4 of the marked-up manuscript. The following paragraph was added in page 6, line 1 to line 19

"MSR-2 total ozone data series inside the vortex are compared to SAT series as shown in Fig. 4, which displays the relative difference between MSR-2 and SAT for the 400K-600K range classification. Differences of about ±0.5% are observed in the 1980s. Small differences are expected during this period since only TOMS data are used in both data sets until 1993. In the 1993-1995 period discrepancies between both curves are only due to the differences in the selection of MSR-2 data for the SAT record in order to have similar spatial coverage as the data from the other instruments incorporated in the SAT time series. These differences varying between -1 and 0.5 % represent an estimation of the impact of reduced spatial coverage in SAT dataset on the averaged total ozone level in September. The 15Sept-15Oct period presents negligible differences. The addition of GOME (1996-2005) in MSR-2 assimilation could explain the discrepancies with the SAT dataset that considers only TOMS-EP. From 2001, differences are larger and generally positive, reaching ~5% in September and ~3% in 15Sept-15Oct. period. These increased differences are especially visible during the period where data from instruments on board the ENVISAT platform (e.g. SCIAMACHY) are assimilated in the MSR-2 record. Overall, values in September present a mean bias of 1.3 % (dash blue line in Fig. 4), and in 15Sept-15Oct a smaller bias value of 0.5 % (dash red line in Fig. 4). Temporal evolution of the differences, e.g. negative trend in the 1980s and positive trend in the 2000s, can have an impact on the long-term ozone trends retrieved from both records. In general, differences between SAT and MSR-2 records are caused by MSR-2 starting to use multiple satellite total ozone columns records after 1996, the procedures in MSR-2 to account for inter-instrument differences, and the data assimilation methodology that allows for filling gaps (van der A et al., 2015)."

Regarding the 5 merged ozone data sets referenced in Weber et al. [2017], they correspond to zonal averages and cannot be used for total ozone classification as a function of equivalent latitude as it is done in our study.

The 'range method' is not clearly explained. I understand that you are using it to see the sensitivity of the calculated trends to the definition used for the area of depletion, and I get that you calculate different areas depending on which isentropic level is used, but exactly how are you deciding which levels to use? Are you averaging over all the 400-600K level results? Only some of them? Do you choose the same range for each year? The details of this methodology were not made clear. It's interesting that in the end you conclude that the 475K results are as good as the other definitions. Is this because this is an altitude where there is some of the most severe depletion? An explanation for this result should be offered.

The range method was better explained in Sect. 3.2 (Methodology for classification), in Page 5, line 4 to line 8.

"*The total ozone column may thus not represent the ozone behaviour inside the vortex. In order to consider possible vortex baroclinicity, another approach is used, where vortex classification at different isentropic levels is considered at the same time. For this second approach, the range of selected isentropic levels is chosen in the altitude region of maximum ozone depletion: from 400 K to 600 K with a step of 25 K. The same 9 isentropic levels considered for 400 K-600 K range classification are applied each year.*"

The range classification considers selected isentropic levels between 400 K and 600 K with a step of 25K. Then the same 9 isentropic levels are used each year for the classification. While this new classification provide a better constraint of low ozone values within the vortex, differences in trend results are not significant at 2 sigma levels. We suggest in the revised version that the reason could be the good correlation between the different data sets (R>0.98) using the different methods. Sentences in the Conclusions (page 13, line 15 to line 24) were changed as follow:

"For the classification of total ozone measurements inside the vortex, the classical Nash et al. (1996) method is used. In order to evaluate the impact of vortex baroclinicity on trend analysis, classifications using a single isentropic levels (475 K, 550 K) and a range of levels (400 K – 600 K) are tested. Systematic differences are found between the various total ozone time series. However the inter-annual variability is similar with correlation coefficients ranging from 0.98 to 0.99 in both studied periods. While larger trend values are generally found with the 475 K classification, the differences with trends related to the 400 K – 600 K range classification are not significant at 2σ level."

The satellite instruments used (all UV sensors) do not see to the south pole in early September. The analysis calculated results for the polar region for the entire month of September, but measurements cannot be made at the highest latitudes in early September. Thus the 'September average' will be more strongly weighted by lower latitudes and later September dates. Please describe how the satellites' sampling of the polar area varies over September and what this does to the 'September averaged' quantity. This may impact the meaning of the trend results as they will include more of the late September, higher dynamical variability measurements.

We agree with your arguments. We have excluded SAT datasets for September also for this reason. In addition we have included a sentence on UV sensors sampling in September where measurements are not available for regions poleward of 77°S in the beginning of September, 82°S mid-September and 89°S at the end of the month in the Sect. 2.1 of the marked-up manuscript (page 3, line 31 to line 33)

"*Since TOMS and OMI UV sensors do not receive enough UV light in early September, originating from regions not illuminated by the Sun (from 77°S to 82.5°S up to mid-September), these regions were not considered to compute the total ozone mean value in MSR-2 data.*"

The impact of spatial coverage differences between SAT and MSR-2 was discuseed in Sect. 4 (page 6, line 6 to line 10).

"*In the 1993-1995 period discrepancies between both curves are only due to the differences in the selection of MSR-2 data for the SAT record in order to have similar spatial coverage as the data from the other instruments incorporated in the SAT time series. These differences varying between -1 and 0.5 % represent an estimation of the impact of reduced spatial coverage in SAT dataset on the averaged total ozone level in September. The 15Sept-15Oct period presents negligible differences.*"