The authors would like to thank the reviewer for taking the time to provide feedback on “On the Use of Satellite Observations to Fill Gaps in the Halley Station Total Ozone Record.” The insightful comments we received helped us understand how to improve on and better communicate the ideas presented in the paper. Nearly all suggestions were incorporated, and a line-by-line response can be found below (with author comments in blue):

RC1: 'Interesting, but needs clarification of uncertainty measures', Anonymous Referee #1, 26 Mar 2021

Overall Comments

The manuscript describes the use of satellite data of total ozone to fill gaps in the ground-based Dobson total ozone record at Halley Bay, Antarctica.

As mentioned in the text, Halley Bay has one of the longest and most important total ozone records. This record was the key for the dectction of the Antarctic ozone hole.

It is a good idea, and scientifically sound, to fill gaps in this important record with satellite data, as well as check for consistency. Overall, the paper is well written and merits publication in ACP.

Before publication, however, I suggest a few important clarifications. Thoroughout the manuscript, I get confused about the use of individual total ozone measurements, daily averages and difference, monthly averages and differences, and the corresponding standard deviations. Sometimes standard deaviations appear to be mis-named "averages" as well. The description of the applied method to shift satellite data towards the Dobson data is also quite long-winded. It would benefit from shortening and clarification. There is no need to make a simple average bias correction appear much more complicated than it is.

Specific Comments

line 21, average of 2 Dobson Units: I don't think this is what is meant. My understanding is the each satellite record is shifted by the Δ from Fig. 3, so that it matches the Dobson data on average. Therefore the satellite average should reproduce the Dobson average exactly, by construction. What is probably meant here is "within a standard deviation of 2 Dobson units". Even more information is required here: are the (presumably) 2 DU standard deviation for monthly means or for daily means? Is the given value 1 or 2 standard deviations? Is it +-1 DU or +-2 DU? Is it even correct? In lines 184 and 185 the stated standard deviation of the differences is 6 to 7 DU. This is much larger than 2 and neesd to be checked.

Thank you for highlighting this issue, and we apologize for the lack of clarity. The original phrase, *“Tests suggest that our method reproduces the monthly ground- based Dobson total ozone values to within an average of 2 Dobson,”* was based on the results of testing how well the satellites used to fill in the 2017-2018 gap could reproduce Dobson monthly means using our Δ-adjustment method (tested on 2013-2015). The phrase is indeed misleading. A more accurate description of our results would be that *“Tests suggest that our method reproduces monthly* *ground-based Dobson total ozone values* ***with an average difference of 1.1 ± 6.2 DU for the satellites used to fill in the 2017-2018 gap****,"* and we have made that change in the text.

The reason the Δ-adjusted satellite average does not reproduce the Dobson averages exactly in our tests is because the years being tested were not included in the calculations of the Δs. Rather than calculate Δs for 2013-2015 and subtract them from the 2013-2015 satellite data (which would lead to an exact reproduction), we pretended the 2013-2015 Dobson data was "missing" to see how well our method could fill in Dobson data by using Δs characterized from the rest of the available data—as would be the case for filling in genuine data gaps, as in 2017-2018. Although the Δs would most likely not be exactly the same as those of the years being tested, our results showed that our Δ-adjustment method was still able to improve on the raw satellite average. This allowed us to understand how well we could fill in the 2017-2018 missing data. We have reworked the discussion of our tests on Lines 190-213 to more clearly explain the data and processes involved and more specific issues are addressed below.

From text and Table 1 it appears that the "root mean square difference" (which is the same as the standard deviation!) for daily average data is about 12 DU. So the 2 DU are probably for the monthly average data, but the 12 DU for the daily data should be mentioned here as well. (Assuming a Gaussian distribution, 67% of the data should be within +-1 standard deviation of the mean (which should be zero here by construction), 95% of the data within +-2 standard deviations, ...

Note: Table 1 (now Table 2) has been changed to display average differences with the Dobson as opposed to root mean square differences. This was done to maintain a consistent comparison metric across the paper.

We apologize again for the confusion that the 2 DU refers to a standard deviation. Table 1 (now Table 2) contains data about the raw satellite daily averages and is used as an introductory comparison of the individual instruments with the Dobson. The table also serves to explain our choice of using the satellite average. The caption has been updated to read *"****(raw) daily*** *measurements by GOME-2A, GOME-2B, OMI…"*

While I call this lack of clarity out here for line 21, it exists throughout the text, and needs to be fixed everywhere.

The phrase has been corrected throughout the text, and the discussion of the text has also been reworked.

line 39: Here it says "throughout the year", line 37 said that no data are available for May to July. What is true now?

It is true that data is not available for May to July. For clarity, we have changed the phrase to be: *"throughout the* ***measurement season****."*

line 45: delete "the" before "satellite"?

The word "system" was accidentally left out of the sentence, it now reads *"In the first decades of the satellite observing* ***system****…"*

line 50: replace "well tested" by "in place"?

For clarity and concision, the phrase *"With the advanced multi-satellite observing system now well-tested"* was removed and replaced with *"****Therefore****, we undertook the development…"*

Fig. 1: are those all measurements or daily averages? Please mention. Are the satellite data the original data from all satellites, or the adjusted data matching the Dobson?

The values are daily averages and the satellite data plotted has not yet been adjusted. The caption has been updated to reflect this and now reads: *"****Daily averages for*** *total column ozone measurements by Dobson instruments at Halley station (in black) overlaid on top of* ***all*** *available* ***(raw)*** *satellite daily averages (in red) from 2014-2019."*

Lines 79 to 104: Would be good to also give the size of the satellite ground pixels near Halley Bay for all the satellite instruments. In addition, I think it absolutely necessary to state which data version was used for each satellite, and where / from which URL the satellite data came from. For GOME2, for example, there are data from Uni-Bremen, from DLR / EUMETSAT, from RAL / ESA_CCI, ... A table of URLs and versions would help here.

This is a good suggestion, and now provide the ground pixel size for each instrument in our revised discussion of each satellite instrument. In addition, we have created a table of URLs, versions, and sources for the satellite data (Table 1 in the revised paper).

Line 85: "cross-calibrated" My understanding is that the current SBUV 8.6 version is not cross-calibrated between satellites, but relies on improved calibration at the radiance level for each satellite. Please check. Natalya Kramarova will know.

The calibration of SBUV instruments for v8.6 is described in DeLand et al. (2012), which is now referenced in the revised paper. The calibration process included updates in calibrations and characterizations for each individual SBUV instrument as well as cross-calibrations of overlapping SBUV instruments. The cross-calibration process was an important step for producing consistent SBUV v8.6 ozone record, for more details on cross-calibrations see Sect. 3 in DeLand et al. (2012).

Line 103: should be "polarization effects"

The section on GOME, GOME-2, and SCIAMACHY was reworked to standardize the discussion of the satellite instruments. The line in question was removed during the process.

Line 107: How were overpasses defined? Satellite foot-point within what distance? Same for all satellites?

The criteria were not the same for all instruments.

GOME-2, GOME, and SCIAMACHY: a weighted average from all footprints (pixels) available per orbit and within 100 km (300 km GOME/SCIAMACHY) of the station were defined as satellite overpass value. All instruments are flying in a near-polar orbit so that at many days several orbits fly over Antarctica. The daily mean overpass ozone was calculated from averaging over the overpassing orbits. The weights are given by sqrt(1-distance^2 [km^2]/max_distance^2). A line briefly describing this has been added to the revised paper: "***Daily mean overpasses were calculated by averaging ozone columns from all ground pixels within 100 km (GOME-2) and 300 km (SCIAMACHY, GOME) of the station.***"

SBUV: the method for creating overpasses for SBUV is described by Labow et al. (2013, see Sect. 5 there). The overpass algorithm for the SBUV data has been created to return daily overpass values each day, even if the SBUV measurements are not directly overhead of the ground station. A box of 2^o in latitude and 20^o in longitude (large enough to encompass two orbits), is chosen around the ground station's location. The SBUV ozone measurements first interpolated along the orbital track with 0.5^o step. Then the SBUV value at the station is calculated as a weighted 1/distance average using all points in the box. A line referencing Labow et al. (2013) has been added to the SBUV discussion.

OMI/TOMS/OMPS-NM: For the overpasses each day the single pixel most nearly co-located with the ground station is selected. At high latitudes a given ground location can be viewed from multiple orbits. In that case a pixel with very high optical path will be rejected in favor of one with a slightly poorer spatial coincidence but with a lower optical path.

OMPS-NP: The overpasses are based on the pixel closest to the station.

The following lines have been added to the discussion of the NASA GSFC instruments: ***"Overpasses for the TOMS, OMI, and OMPS-NP instruments are defined by selecting the single pixel most nearly co-located with Halley Station. In the case of there being multiple pixels available, a pixel with high optical path will be rejected in favor of one with slightly poorer spatial coincidence but lower optical path. For the OMPS-NP instrument, the pixel closest to the station is chosen."***

line 127: Would it not be better to have Figure 3 and lines 155 to 160 right here in section 2.3? After all, the Figure shows the Δ-s that are discussed in lines 120 to 127?

We decided to keep Figure 3 under Results because it is the result of the methodology described in section 2.3. The discussion on the figure has also been expanded (L174-179, revised paper).

Line 128, Section 2.4: Would it not be clearer, to have section 2.2 here, after section 2.3. That way, you would have a more logical flow. a.) discuss Δ-s for individual satellites b.) discuss how you use all satellites to fill in, and how that looks for the different months.

We decided to keep the order of sections 2.2 and 2.3 because the calculation of the Δs uses the averaged data described in section 2.2.

Table 1: What is shown here? Differences between monthly averages, or differences between daily averages? From the numbers, around 12 DU, it looks like it was daily averageds. Were the satellite data Δ-adjusted or not? In April, that would make a large difference according to Fig. 3.

Note: Table 1 (now Table 2) has been changed to display average differences with the Dobson as opposed to root mean square differences. This was done to maintain a consistent comparison metric across the paper.

The differences are between the daily averages, which were then averaged across each month and in total. The satellite data has not been adjusted here. Section 2.2 (Data Analysis) now better explains this: "With all measurements and differences in the form of averaged daily values, data were categorized **and then averaged** according to their corresponding month and day of year (DOY)."

The caption for Table 1 (now Table 2) has also been changed for clarity, it now reads: *"**Average absolute** differences in DU between the total column of O3 retrieved from the Halley Dobson instrument and those retrieved from the **(raw) daily** measurements by GOME-2A, GOME-2B, OMI, OMPS-NM, OMPS-NP, SBUV **averaged by month and in total** for the period from 2013-2018."*

Line 143: Would the Δ-adjustment not take care of the Bass-Paur difference as well? Is it necessary to mention systematic biases here, since the filling-in method takes care of them anyways?

Although the Δ-adjustment would most likely take care of the Bass-Paur difference as well, we felt that not mentioning and correcting the systematic bias would misrepresent the performance of the GOME-2 instruments in Table 1 (original paper, now Table 2). Additionally, we believe that it makes sense to correct a systematic bias that is already known and explained.

Figure 4: Having Figure 4 so close to Figure 3 confused me (Are they now using monthly Δ-s again? Or daily? Or what?). I guess the only point of Figure 4 is to show that 2019 was very different from the other years. This does not become very clear here. The stars for 2019 are easy to miss in the Figure, and they do not have error bars. It would be helpful to have a clearer Figure, that points out 2019 in a legend in the Figure, not just in the caption.

(Addressed along with the next comment)

Also Figure 4: What are the error bars? Standard deviation of daily data or monthly data? Standard error of the mean? One or two standad deviations?

We apologize for the confusion. The purpose of Figure 4 is to show that 2019 was an anomalous year using monthly Δs. We have made the stars more visible, added a legend, and clarified what the error bars represent. The rewritten caption reads: *"Average Δ over all years (Fig. 2) excluding 2019 for each month with error bars (black). **The monthly Δ values with the automated Dobson in 2019 (red) are much larger than other years. The error bars***

***represent the standard error of each satellite mean, combined in quadrature for each monthly bin.*"**

Lines 169 to 187, and Figure 5: I am confused. Does Fig. 5 show data, where 2003 to 2012 was the training period? Or which training periods were used to generate the data in the two panels of Fig. 5?

For tests done on 2013-2015, we pretended the Dobson data was "missing" for that period and each satellite was Δ-adjusted using Δs calculated from the rest of its available data, excluding 2019-2020. We have reworked the discussion of Figure 5 to clarify our test, particularly in Lines 190-202.

Lines 184, 185: I assume that the numbers are for the trained data? Please state the same numbers for the unadjusted satellite data. Only then you can conclude if the adjusted date are better, or not. Check consistency with numbers in abstract and conclusions!!

Thank you for this suggestion! When we calculated the average differences for the unadjusted satellite data, we found values of 6.5 DU for 1998-2002 and 4.6 DU for 2013-2015, which were much larger than the values for the adjusted data (1.8 ± 6.7 DU and 1.1 ± 6.2 DU, respectively). The uncertainty represents the estimated training error, so there is no comparable value for the raw data. A few lines comparing the two have been added in the discussion of Figure 5 (L209-213).

We have also corrected the abstract and conclusion to say *"with* ***an average difference of 1.1 ± 6.2 DU*"** in order to be more consistent with the results.

Figure 6: The differences between the Dobson monthly means and the Δ-adjusted satellite data look rather large in 2019 and 2020, 5 to 10 DU. Is that consistent with the numbers given in lines 184, 185? Figure 4 shows that the 2019 Dobson data are flawed. Are the 2020 Dobson data flawed as well? Flaws of the Dobson data should be stated, and maybe even marked with different sysmbols in the Figure. How do the Δ-adjusted satellite data look in the other years? It would be good to plot the entire red time series.

Yes, we believe the 2020 Dobson data are flawed as well due to likely inconsistencies between the automated instrument and earlier data (Figure 4). The use of the automated instrument was continued in 2020. We have added the following sentence in the discussion of Figure 4: ***"Because the station continued to use the automated instrument in 2020, we treated the 2020 data as likely inconsistent as well and excluded it from our Δ adjustment."*** We also now note this in the caption of Figure 6 with the following added sentence: ***"Dobson data from 2019 and 2020 were filled in due to apparent inconsistencies between the automated instrument and earlier data."***

Because 2019 and 2020 were excluded from the Δ adjustment, the numbers given in lines 184 and 185 (original paper) are not inconsistent with Figure 6.

We decided not to plot the entire red time series along the Dobson because the primary focus of our study is to fill in gaps in the Dobson as faithfully as possible, not to investigate between the Dobson and satellite instruments. The purpose of Figure 6 is to present the now complete record.

Line 230: Are the 2 Dobson Units the average difference? Is that really relevant? In principle, the average difference should be zero, due to the Δ-adjustment. Of course zero is not realized in every subset / realization of the data. Is not the standard deviation between Dobson and Δ-adjusted satellite data a much more meaningful quantity, to show how well the two data sets agree?

(Addressed along with the next comment)

Also line 230: Check consistency with the numbers in abstract and in lines 184, 185. Please give (also) the standard deviations of Dobson minus Δ-adjusted satellite data on the basis of monthly and daily means.

Yes, the 2 Dobson Units are meant to represent the average difference, but the line has been corrected to read: "average difference of 1.1 ± 6.2 DU for monthly averages," with the uncertainty now included. The average difference was not zero because this was the result of calculating Δs using the rest of the data and using it to fill in "missing" 2013-2015 data (explained above).

We felt that the statement *"we could fill in missing months with a high degree of fidelity"* should be backed up quantitatively, hence why the test results are included. The numbers should now all be consistent across the paper.