

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):

RC2: 'Comment on acp-2021-122', Anonymous Referee #2, 21 Apr 2021

The authors use observations from multiple satellite instruments to create an ozone column dataset above the historically and scientifically significant Halley station in Antarctica, using the Dobson instrument as a calibration anchor. The result is a dataset that can fill in gaps due to a recent ice crack, check for calibration issues affecting the new Dobson data, and fill in gaps caused by other future geophysical or social disruptions.

General comments:

Overall, I enjoyed reading this work. The manuscript is well written and organized. Gaps in scientifically significant long-term datasets – such as the Halley Station’s Dobson – are an important problem. The paper contributes usefully to this topic. I have a few questions and requests for clarification.

Thank you for the positive feedback!

The satellite instruments used in the study have a variety of measurement techniques, advantages and limitations. While it might not be necessary to do a deep dive into this, I think it is at least necessary to comment on the different spatial resolutions (and perhaps vertical sensitivities) and the spatial coincidence criteria used to define co-location with Halley Station.

We agree that this would be useful information to provide, and the revised paper discusses spatial resolution and spatial coincidence criteria in a more standardized format for all satellite instruments (L83-123).

Figure 1 is a key illustration of the datasets involved in this study – i.e., the adjusted average satellite and Halley Station Dobson. The figure is limited to 2013 – 2019, I assume because this allows features on the ~monthly timescale to be seen and because 2013 – 2015 was one of the time windows used to test the technique’s ability to reproduce the Dobson. And 2017 – 2018 was the motivating gap in the Dobson’s timeseries. Nonetheless, I would like to see the full timeseries of the two datasets. Figure 6 does some of this, but I believe it only inserts the satellite dataset into the Dobson gaps (?). That’s useful, but I’d also like to see both fully plotted to get a sense of how closely they agree.

We apologize for the confusion. Figure 1 shows the Halley Station Dobson plotted alongside the **raw** satellite daily averages from all instruments. Its purpose in the introduction is to display the 2017-2018 Dobson gaps that are the focus of our paper as well as illustrate how the satellite instruments provide complimentary information during the gap periods. The caption has been reworded to say: “*Figure 1: **Daily averages for total column ozone measurements by Dobson***”

*instruments at Halley station (in black) overlaid on top of **all** available (**raw**) satellite measurements (in red) from 2014-2019.”*

As requested, we have added additional timeseries plots for the complete time series of the available satellite data versus Halley in the supplement (new Fig. S1-S3).

I am not entirely comfortable with the conclusion that the adjusted satellite average reproduces the Halley Station Dobson to within about 2 DU. I don't think it was sufficiently stated that the satellite reproduction of the Halley Dobson data varies seasonally. In Figure 5, some months show very close agreement; others show much larger differences, e.g., ± 15 DU. 2 DU is the apparent result of averaging large positive differences and negative differences across the year. In addition, while there are similar annual patterns, there are also notable differences between the years shown.

Yes, it was incorrect of us to conclude that the adjusted satellite average reproduces the Halley Station Dobson “*to within an average of 2 Dobson units*” for the reasons you mentioned. We have re-evaluated the numbers carefully, and rewritten the sentence to represent our results more accurately, and it now reads: “***Comparisons to the Dobson suggest that our method reproduces monthly ground-based total ozone values with an average difference of 1.1 ± 6.2 DU for the satellites used to fill in the 2017-2018 gap.***”

In addition, the time periods chosen for the test have a particular combination of satellite instruments that are being used to compare with the Halley Station Dobson. What were the results of the comparison between the adjusted satellite average and the Dobson for wider time periods? How well do these test cases, 1998 – 2002 and 2013 – 2015, generally represent the physical conditions and satellite datasets available to fill other gaps and future gaps?

This is a good point. Given that our study primarily focuses on developing a sound methodology for filling in the 2017-2018 gaps, we chose the 2013-2015 testing period because the instruments in operation are the same as the ones in 2017-2018 and would presumably be used in future efforts (GOME-2A, GOME-2B, SBUV, OMI, OMPS-NM, and OMPS-NP) until they cease operations. Unfortunately, we are unable to test on a wider time period for this particular set of instruments because the majority of them are only available from around 2012-2020. If we widened the training period, we would run the risk of not having enough years to “train” the Δ s on. A similar issue arises with the four instruments available in 1998-2002.

We understand that the decisions made in our tests as well as the methodology may be confusing, so we have rewritten the discussion to clarify it as follows:

*“Consequently, we chose to test the method for the years 2013 to 2015 by pretending data for those years did not exist and characterizing the monthly Δ values averaged over those years **using the rest of the available data for the GOME-2A, GOME-2B, OMI, OMPS-NP, OMPS-NM, and SBUV instruments.** To examine the performance of our method during periods when there were fewer available instruments, we also tested on 1998-2002 using **data from GOME SCIAMACHY, SBUV and EP/TOMS instruments.** **The range of available***

data for each instrument can be found in Figure 2. The training period for each instrument is the available range after excluding the years being tested (and 2019-2020)."

The approach described in the paper, and the resulting dataset, is useful, but I think care needs to be taken in asserting the accuracy in reproducing the Dobson shown here – especially since particular months of the year are often of more significant interest to the study of ozone chemistry than annual averages.

Thank you again for the positive feedback. We agree that particular months of the year are often of more significant interest to the study of ozone chemistry, and our results (i.e., Table 1, now Table 2) suggest that the satellite average does not exhibit anomalously high differences during the months that comprise the ozone hole season. However, we have corrected our conclusions to be more representative of our results—as mentioned above.

Early in the paper (section 2.2) it is stated that both absolute and relative differences are computed. The paper then focuses on the absolute differences. Given the annual pattern to the absolute differences, I'm curious to know what the results of the relative differences were? Do percent differences show the same seasonality as the absolute differences? Would constructing a delta adjustment on the basis of a relative difference or SZA remove some of the annual pattern in the difference?

Yes, relative differences were computed but showed the same seasonality as the absolute differences. Early on in our research process, we examined solar zenith angle dependence as well and determined that the delta is not simply dependent on that. We decided that this topic should be a focus of a later paper, and our satellite and BAS coauthors are aware of the issue. However, we agree that the issue is important to mention in the paper, and a line has been added in section 2.3 (Delta Characterization and Adjustment). It now reads as follows:

*"... the Δ value for each day of year is the average of the absolute differences between each satellite and Dobson for that day of year, across all years in each satellite series. **Relative differences were also computed but displayed the same seasonality as absolute differences.**"*

Figures:

It would be helpful to expand the width of the figures to fill the width of the page. Figure 1, for example, would benefit from this since it can be difficult to see the structure in the data and the comparison between the Dobson and satellite measurements. This is important for understanding the work being described. Figure 3 as well.

We agree with the suggestion. All figures have been widened to better fill the width of the page.

Figure 4: why is there no monthly value for August (month 8) 2019? Caption title could add that 2019 is with the automated Dobson so that this context stands alone in the figure without the text. I'm assuming the error bars are the standard error average? Is the August error bar larger because there are fewer days being used (polar night limitation)?

Thank you for catching this error. The value for August 2019 was cut out of the original figure and the axes limits have now been corrected to include the data point. To make the goal of the figure more clear, we have also changed the caption to read: “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.**”

It is likely that the error bars in August and April are larger due in part to the polar night limitation; we have added a line mentioning this limitation in section 2.2 (data analysis). The line is: “**Months directly bordering the polar night (April and August) contained fewer data points when computing monthly averages.**”

Figure 6: has tick labels that are too small to easily read; they are notably smaller than other figures.

We have fixed the axes font size.

Tables:

Table 1 is awkwardly split across page 5, which has the caption and titles, and page 6, which has all the values. Please note this to the typesetter and check that the proofs correct this.

Thank you for pointing this issue out, the table should no longer be split in the revised paper but we will make a note to the typesetter.

Could be informative to add a column for the delta-corrected satellite average and/or the delta adjustment.

We decided to not introduce the Δ -adjusted satellite average in Table 1 (now Table 2) in order to keep the purpose of Table 1 (now Table 2) as it was: comparing the various (raw) satellite data and justifying our use of the satellite average to fill in the gaps.

Abstract:

In a few cases, it might be helpful to the reader to add specifics. For example,

“by... adjusting overpass data” – adjusting how?

The line originally read as “*bias-correcting overpass data*,” but we were advised by the editor to remove mentions of “bias” throughout the text—hence why we refer to “ Δ s” and “ Δ -adjustment.” The specifics of our adjustment are explained in the Methods section of the text.

“Tests suggest that our method...” – what tests? Or rephrase to say “comparisons to ___ suggest that our method...”

We have changed the sentence to be more descriptive, it now reads: “*Comparisons to the Dobson suggest that our method reproduces monthly ground-based total ozone values with an average difference of 1.1 ± 6.2 DU for the satellites used to fill in the 2017-2018 gap.*”

“... our approach improves on the overall performance...” – what does overall performance refer to? Accuracy of the measurement? Comparison results? Completeness? Also want to be careful because the goal of the paper, as stated at P3L57-58, is not a high-performance dataset / “most accurate” dataset, but “to reproduce what the Dobson instrument would have observed”.

This is a good point. The phrase “*our approach improves on the overall performance*” has been changed to: “*our approach more closely reproduces the Dobson measurements than...*”

“... there was a significant difference between the two.” – I’d suggest being more quantitative. What was the difference? This also brings up a question of what the authors consider to be a threshold for good agreement?

Since the primary focus of our paper is to fill in the Halley Dobson record and reproduce total ozone values as would have been measured by the Dobson, we decided not to delve too deep into the topic of agreement between the different satellite instruments and the Dobson. Throughout the text (i.e., L251-252) we suggest that it could be a topic of future study using our method and datasets).

We agree that the term “*significant difference*” is too vague because of its quantitative implications, so we have replaced it with: “*there were apparent inconsistencies between the two.*” This statement is backed up in the text by Figure 4.

Specific comments:

P2L1: A good paragraph starting sentence. But, suggest rephrasing “now”, since the interruption being discussed was a few years ago.

This is true. The sentence has been rephrased to say: “*In 2017, this remarkable record was interrupted.*”

P2L49: (warning: pedantic point) Most readers will understand what is meant by “With the advanced multi-satellite observing system now...” but the wording here might suggest that the current (and past) suite of satellites are part of a coherent and coordinated “system”. We have a great scientific community and missions are chosen to provide complementary coverage. But I’m not entirely convinced the satellites used in this study, from different agencies, eras, and designs, are “an observing system”.

That is a good point. 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...*”

P4L79: missing parenthesis before the Munroe citation.

Thank you—fixed.

P4L76: “...some also have spectral information at other wavelengths.” – what instruments and what other wavelengths?

The spectral information at other wavelengths was not relevant to our study and the segment has since been deleted.

P4L80: Suggest re-wording “spin off of the” – it isn’t clear what is meant by “spin off”.

The entire section on the GOME and SCIAMACHY satellites have been revised. The word “spin off” is not used anymore.

P4L81: “... of a somewhat improved...” – “somewhat” is hard to interpret. What was improved (or not) that is notable here?

(addressed along with the next comment)

P4L79:82: why put this information about GOME here when the details of this instrument are four paragraphs below? Doesn’t serve the flow.

Thank you for pointing this out. We have reorganized the paragraphs on the satellite instruments to be in a more standardized format. The information about GOME is now consolidated in Lines 83-91.

Because the line in question was not relevant to our study (P4L81, original paper), it was removed when the section was reorganized.

P4L83: SBUV acronym used before defined. Need to carefully define, given it is an instrument and set of instruments.

The SBUV acronym is now defined, the revised sentence reads: “*The Solar Backscatter Ultraviolet (SBUV) record...*”

P5L111: It is important to be clear what has been done to define the spatial coincidence criteria for the satellites and Halley Station.

Spatial coincidence criteria is now discussed in Lines 88-91 (GOME, GOME-2, and SCIAMACHY), Lines 98-99 (SBUV), and Lines 115-118 (TOMS/OMI/OMPS-NM/OMPS-NP).

P5L117: “... due to unusually high differences...” – what were the differences.

Once again, we decided not to delve into the topic of what a threshold for agreement between the Dobson and satellite instruments would be. We have rewritten the sentence as follows: “*lunar Dobson measurements from August 24th, 2015 were excluded due to obviously anomalous differences.*”

P5: Section 2.3: You state that the delta value for each DOY is averaged across all years in each satellite series. Was there any trend or interannual variability in the differences before combining all years?

That is a good point. The uncertainty in the Δ -characterization and adjustment due to interannual variability in the differences is captured by our estimated training error. We now address this in Methods—here is the revised line: *“Uncertainty for the delta-adjustment of the satellite average was calculated by combining, in quadrature, the standard error of the mean for each satellite and accounts for interannual variability.”*

We have also added a phrase explaining the uncertainty metric more clearly in the Results section: *“ $1.8 \pm$ an estimated training error of 6.7 Dobson units (DU)” (L208).*

P6L149: Why is April different?

(addressed along with the next comment)

P8: The observed differences are typically larger than the average delta, which is reduced by large positive and negative values averaging out. If this pattern is due to SZA-dependence, would a bias correction based on (or including) SZA rather than DOY produce a useful result?

For an unknown reason, Δ does not seem to follow a simple solar zenith angle dependence, as evidenced by the much higher values computed in April than in August at similar solar zenith angles. We choose not to speculate on this in our paper, but it is the reason that we chose to do a bias correction based on DOY rather than SZA. We have reworked the discussion of Figure 3 to explain this component of our reasoning and added the following sentences:

“Additionally, Δ does not follow a simple solar zenith angle dependence. Values differ between the onset and end of the polar night for days with the same solar zenith angle, as evidenced by the larger Δ s in April versus August. Therefore, we chose to characterize Δ by day of year rather than zenith angle.”

P8: If there was an unusually high or low level of ozone abundance, will the absolute delta correction sufficiently reproduce the Dobson? Would a relative (%) adjustment scale better?

The agreement at Halley appears to scale with ozone abundance. We are not sure why this is, but the issue does not seem to persist when comparing the satellite instruments to the Syowa station. We have alerted our BAS coauthors to this issue but decided this should be the focus of a different paper, as our primary goal is to reproduce what the Dobson would have seen—not discuss the accuracy of the Dobson.

P9L190: suggest adding a comment about calculating averages for the months that border polar night, where there will be a reduced number of days contributing. I don't know offhand when Halley Station's latitude enters/exits polar night. That might be worth mentioning somewhere since it is relevant to data collection and chemistry.

We agree this is worth mentioning. We have added the following sentence to our discussion of the averaging in section 2.2: ***“Months directly bordering the polar night (April and August) contained fewer data points when computing monthly averages.”***

More information about the measurement season at Halley Station has also been added to section 2.1 (Data). It now includes the lines: “Halley solar data typically end on April 16th as the sun retreats for polar night, and resume on August 27th. There are also some limited lunar measurements.”

P12L223: “Larger difference” – how large? Since this is an important finding, it would be good to state this. Hopefully this prompts investigation into why there might be differences between the measurements.

Figure 4 was developed so the reader could judge this for themselves for their season of interest. We revised the paper to: ***“we found that the preliminary computed data from the automated instrument in 2019 had apparent inconsistencies with the earlier data taken with the manual Dobson when compared to the satellite average (see Fig. 4).”***

P13L227: were any of the satellites used in this study validated using the data collected at Halley by the Dobson?

The WFDOAS (GOME, SCIAMACHY, GOME-2) approach has been validated using Halley Bay data (Weber et al. 2005, Orfanoz-Cheuquelaef et al., 2021) but the rest of the instruments have not. This is now mentioned in the text:

“The WFDOAS approach was validated using Halley station data as reported in Weber et al. (2005) and Orfanoz-Cheuquelaef et al. (2021).”

“None of these instruments [TOMS, OMI, OMPS-NM, OMPS-NP], as well as SBUV, were validated with Halley station data.”

P13: There was quite a bit of discussion about the DOY data but the conclusions focused on the monthly averages. Why not include a plot of the DOY comparison results as well and comment on their comparability to the monthly results?

The results focus on the monthly averages because they are the easiest to visualize and tabulate in the paper. We felt that the DOY plots would be too messy to draw any useful conclusions. With respect to the tests, the limitations of the testing periods (mentioned above) would only allow us to make comparisons for a maximum of three datapoints for each DOY.

We agree that the filled DOY record should be provided, and we have made the data available at: <https://www.ssolomongroup.mit.edu/toolsandproducts>