

Response to Referee #1:

We thank referee #1 for their helpful comments. Our responses are given below in black with the referee's comments in blue. The new text in the modified manuscript is given in red (italicized).

This is the first review of the paper "The world Brewer reference triad – updated performance assessment and new double triad by Xiaoyi Zhao et al. This paper is of great importance for the WMO Brewer network as it discusses the stability of the world Brewer triad maintained by the ECCO, Canada. Comparisons between the single Brewer triad (BrT) and the double Brewer triad (BrT-D) are reported for the 1999-2019 period. The previous assessment of the BrT performance (Fioletov et al., 2005) is used to verify the stability of the reference instruments over an extended period (1984-2019). Four statistical methods to evaluate the uncertainty of each instrument relative to the BrT and BrT-D baseline, to the independent reference observations (Pandora and eleven satellite records), and to the reanalyses (MERRA-2) are presented and summarized in plots and tables.

The paper is well written, the figure used to demonstrate the analyses are clear. The summary tables support the discussion and allow us to evaluate stability and random uncertainties of the total ozone observation originating from uncertainties in the extra-terrestrial constant (ETC) and the effective absorption cross-section coefficient specific to each instrument in the triad. There are a couple of inconsistencies in the analyses, including grouping of the data in either monthly, 3-months, or 6-months averages. It is not clear why the time periods for averages are changing depending on the analyses. It would make sense to present all data as monthly averages.

We thank referee #1 for the positive feedback on this work. As pointed out by the referee, some of the analyses were done at different frequencies. The analyses made with Models 1 and 2 as well as those analyses with satellite and reanalysis data used a 3-month frequency (e.g., Figs. 1a, 2, 5, 6, and 7). The 3-month frequency is selected due to having a better balance of sufficient co-incident measurements and good temporal resolution. For example, Fig. R1 shows

the number of days that can be used in Model 1 analysis with different analysis frequency (from 1 month to 6 months). A specific day is analyzed with Model 1 only if each of the three instruments has 1) at least ten measurements on that day and 2) at least three measurements in each half-day on that day (see Section 4.1.1). The median values of days used in Model 1 analysis are 11, 32, and 64 for analysis frequencies of 1 month, 3 months, and 6 months, respectively. Using monthly averages will have some undersampling issues, especially in the winter period. In addition, the Models 1 and 2 analyses done in the previous triad assessment (Fioletov et al., 2005) used a 3-month frequency, which was selected to preserve any possible artificial seasonal cycle in ETC errors and also to have as many data points as possible. Thus, to make this new assessment work be directly comparable with the first assessment, we decided to keep using this 3-month frequency, and change other analyses to match this frequency.

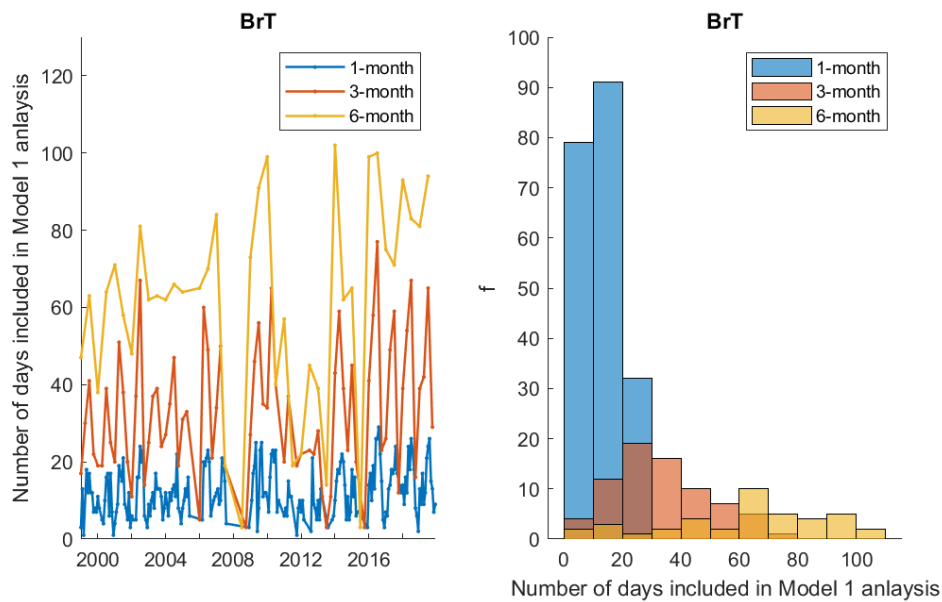


Figure R1. The number of days included in Model 1 analysis for Brewer reference Triad (BrT).

Different frequencies were used when comparing with Pandora data. For example, we selected monthly frequency (Fig. 3, relative difference) to better illustrate the fine-scale variability (e.g., January to February 2017, in the original Fig. 3). However, we agree with the referee that a consistent analysis frequency is a better choice. Thus, Fig. 3 has been modified to a 3-month frequency.

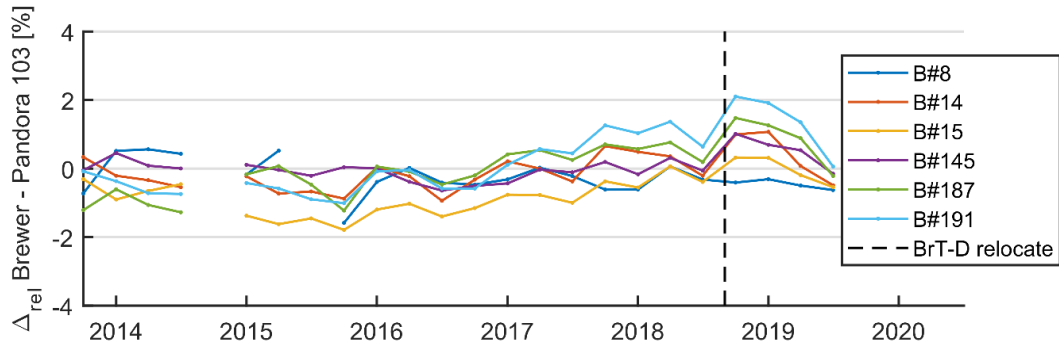


Figure 3. **3-month** relative differences between Brewers and Pandora total column ozone. **3-month** averages are calculated if there are at least ten coincident measurements between Brewer and Pandora for that period. The black dash line represents the time when BrT-D was relocated to Egbert, i.e., Pandora and BrT-D were not co-located.

We also updated Fig. 4 to use a 3-month frequency.

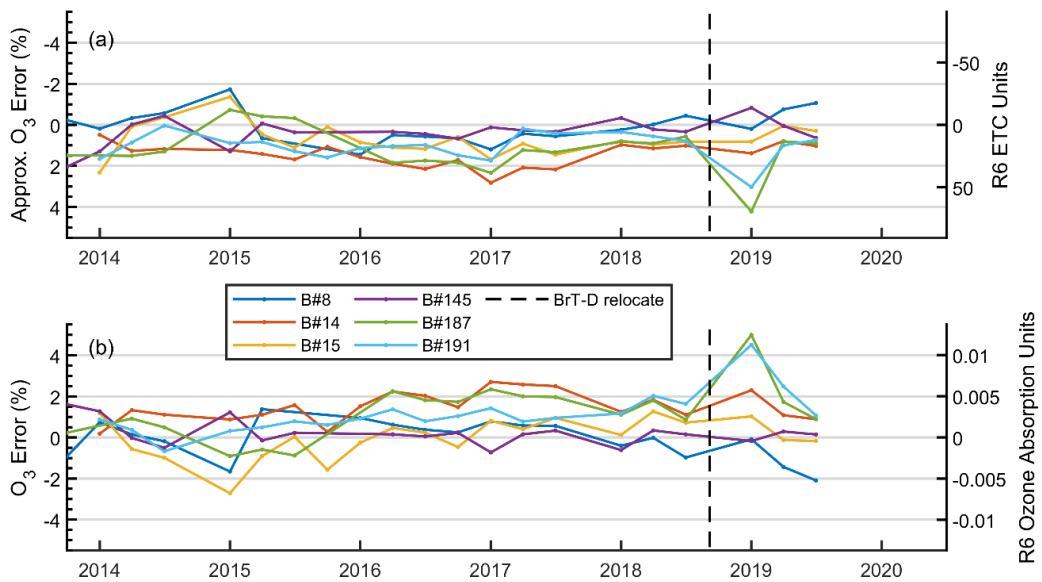


Figure 4. Relative systematic uncertainties in ETCs and effective ozone absorption coefficients estimated using Model 3. Description of y-axes is in Fig. 2. Each point on the graph represents a **3-month** average. The black dash line represents the time when BrT-D was relocated to Egbert.

The 2005 paper analyzed data starting in 1984. Why does this paper exclude the 1984-1988 period? Since the triad is independently calibrated at Mauna-Loa observatory, where the

station Dobson (since 1957) is located, why not to perform comparisons for data collected by triad at MLO? The traveling Brewer reference is used to calibrate station instruments. It would be good to include its record with respect to BrT in this paper.

The suggestions from the referee, i.e., including 1984-1998 data and comparisons with MLO data are very important. We think that the whole four decades of observations should be carefully evaluated and can be useful, e.g., to provide high-precision TCO trends in Toronto. However, the focus of the current work is to provide an updated assessment for triad in the past two decades. Thus, we selected the 1999-2019 period in this assessment work to provide 5 years of overlap with the first assessment (1984-2004).

A comparison between Brewer reference instruments and Dobson instruments at MLO is also possible. However, for each calibration trip, the Brewer reference instrument will only co-locate with Dobson at MLO for about a month. Thus, the dataset will be small, i.e., less than 17 months (see Table 2). Including these analyses will not likely affect the results and conclusion from this work. Moreover, the Dobson operated at MLO is not the Dobson world reference instrument. The world reference, Dobson #83, is calibrated at MLO once every several years. Therefore, it is not possible to compare the triad instruments with the world reference Dobson. Thus, to make the current work more concise, we would prefer to leave this analysis work in a future publication.

In addition, in a joint work with other Brewer groups, we are planning a publication detailing about the absolute calibration procedure, the calibration transfer procedure, and an assessment of travelling standard instruments soon. Together with the triad assessments (Fioletov et al., 2005 and current work), these works will provide the general, but important pictures, of ozone monitoring activities carried out by the global Brewer network.

Here are specific comments:

1) line 68. The text “230 Brewer instruments deployed” is in contradiction with the abstract where 230 instruments are referred to as “produced”. Were all produced instruments deployed?

Some of the Brewers have been retired after years of services and are not currently deployed. The sentence has been modified.

By 2019, there were more than 230 Brewer instruments manufactured, with most of them deployed worldwide within the WMO GAW global ozone monitoring network.

2) Line 70-71. The paper states that 123 instruments are currently in operations and are located at 88 stations. How many countries use Brewer instruments for ozone monitoring? Are there Brewers that are not part of the WMO GAW network and do not submit data to WOUDC for archiving?

This is an interesting question, but outside the scope of this paper. Some instruments are operated by universities and have no connections to the WMO GAW. We can only provide information about sites that were calibrated using the Toronto Brewer triad as a reference. Detailed information on such calibrations and data submissions to the WOUDC is available from International Ozone Services Inc. (IOS) web site at <https://www.io3.ca/Calibrations>. In the last twenty years, the total number of distinct Brewers that have been calibrated by IOS is 148. On average, IOS has transferred world reference instruments' calibration to about 40 Brewers by visiting 15 countries per year. These Brewers are located in 48 countries. To complete all these calibration, IOS took 599 trips. Figure R2 shows the time series of these calibration activities. Some of these information have been included in the revised manuscript.

In practice, each field Brewer instrument receives its ETC constant by comparing ozone values with those of the travelling standard instrument. The travelling standard itself is calibrated against the set of world reference instruments (i.e., world Brewer reference triad). The world reference triad data are used to calibrate the traveling standard, and the traveling is used to calibrate 30-40 Brewers per year, on average, around the world.

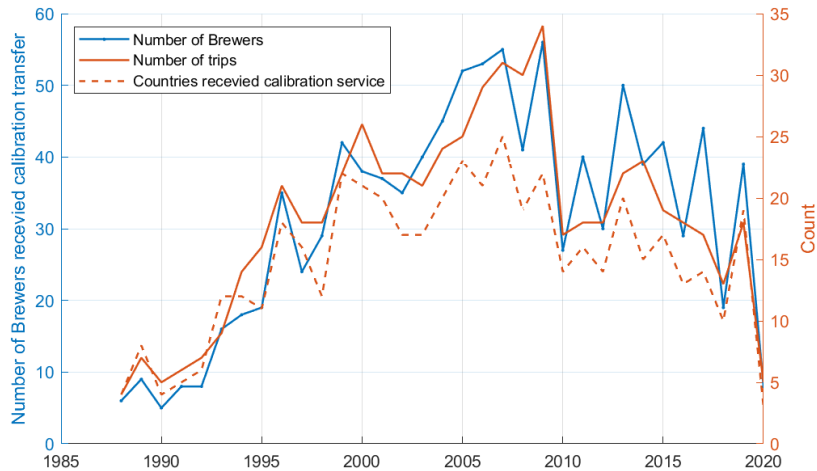


Figure R2. Time series of the calibration transfers done by IOS from 1988 to 2020.

3) Lines 73-74. Do I understand correctly that effective ozone cross-section is determined once after the instruments are produced? Are there in-field instrument adjustments that can change the instrument-specific absorption cross-section, overtime, or abruptly? Is there a method to check the stability of the ozone cross-sections? Is it done when the instrument is calibrated at MLO?

Brewer uses BP (Bass-Paur) ozone cross-section (at 228.3° K, Bass and Paur, 1985), which was measured in the laboratory. The effective ozone cross-section mentioned by the referee should be the effective ozone absorption coefficient ($\Delta\alpha$). This coefficient is generated for each Brewer by performing the dispersion test (DSP) (Savastiouk, 2006) with the use of a group of discharge lamps (e.g. Hg, Cd, In). In general, the slit functions of the Brewer are determined by DSP. Then, $\Delta\alpha$ is calculated as the convolution of slit functions and literature ozone cross-section at the operating wavelengths. It is correct that the in-field adjustments may change $\Delta\alpha$. Thus, after each adjustment work, new $\Delta\alpha$ will be measured via DSP. This work can be done in the field. The stability of $\Delta\alpha$ is directly related to the stability of the wavelengths setting in the Brewers. This is regularly checked using the stable solar spectrum as the reference using the so-called Sun Scan test. Some of these information has been included in the revised manuscript.

For example, the effective ozone absorption coefficients ($\Delta\alpha$) are determined for each individual instrument in laboratories via dispersion test, and are regularly checked using the stable solar spectrum as the reference using the so-called Sun Scan test (Savastiouk, 2006).

4) Line 80. “reference instrument ... is independently calibrated every 2-6 years”. What are the WMO GAW requirements for the frequency of calibrations of the triad? Is it consistent with the requirement for the infield instrument calibrations? Table 1 shows that some instruments were not calibrated for 6 years. Would this affect the triad stability? What is the requirement for the traveling standard calibration?

When the primary calibration has been done for one of the reference instruments at MLO, this instrument can be used to validate the status of other reference instruments in Toronto. Therefore, to satisfy the 2-3 year interval between calibrations requirement, it is sufficient if at least one triad Brewer is calibrated at MLO every 2-3 years.

The traveling standards need to be calibrated against a World Brewer Reference – traceable instrument before and after every calibration trip. This ensures the quality of the transferred calibration and a complete understanding of the traveling standards’ performance.

5) Line 109, Table 1, right column, row 6 – “Significantly less instrumental stray light than in single instrument” – please quantify what it means, include information about the level of rejection of the stray light, i.e. 10^{-4} , 10^{-5} in the wings? Is stray light here attributed to the out-of-of band light? How much does it contribute to the total column ozone error at representative air mass over Toronto?

The strength of stray light effect depends on the slant ozone amount and not on air mass. The median air mass factor over Toronto is 2 ($\mu = 2$), for which the stray light effect is weak. As illustrated in Fig. R3, BrT and BrT-D start to have more than 1% relative difference when $\mu > 3.5$ (equivalent to slant ozone 1200 DU). Thus, data only with $\mu \leq 3.5$ are used in this assessment work (except Fig. A2). In conditions with representative air mass values (e.g., μ values about 2), Brewers have a median standard deviation of about 1.2 DU (see Fig. A1). Details of the stray

light issue are provided in Appendix A. Following suggestions from the referee, the description in Table 1 has been updated.

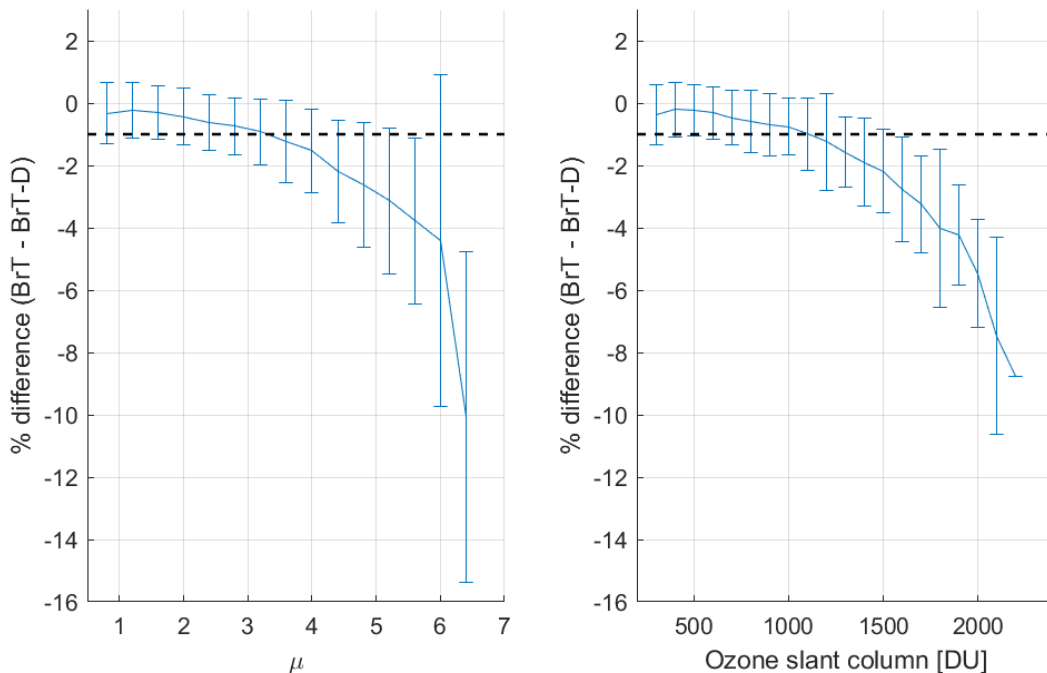


Figure R3. The relative difference between BrT and BrT-D, in terms of air mass factor (μ) and slant column ozone. The error bars represent 1σ of the relative difference values. The black dash lines show the -1 % relative difference.

Significantly less instrumental stray light (out-of-band, stray light fraction 10^{-7}) than in the single monochromators (10^{-5}) (Fioletov et al., 2000).

6) Line 145. The period of evaluation includes 2019 which is after the BrT was moved to a different location in 2018. Why not exclude 2018-2019?

It was the BrT-D that been moved to the Egbert site temporarily since September 2018. We decided to include this period to demonstrate some fine-scale (spatial) variability of stratospheric ozone field (i.e., the two monitoring sites are only 55 km apart). A detailed example is provided in Section 5 (see Fig. 8), which shows the fine-scale variation may have a significant impact on the validations of high-resolution satellite TCO product (e.g., TROPOMI).

Also, please note that for Models 1 and 2 analyses, since the baseline ozone was formed by each triad, this re-location will not affect the assessment results. Although we see the difference between BrT and BrT-D in 2019 January to February as in Figs. 3 and 4, these differences will not be reflected in Figs. 1 and 2. Thus, including this period will not affect our major conclusions about BrT and BrT-D's long-term stability (via Model 1 and 2).

7) Line 170 – “seasonal mean” – is it the same 3-month averages that are discussed later (line 432)?

No. The values were estimated with monthly averages in Zhao et al., 2016. The sentence has been modified to clarify this.

In general, after correction, the multiplicative bias in Pandora ozone data can be decreased from 2.92 to -0.04 %, with the seasonal difference decreased from ± 1.02 to ± 0.25 % (see Fig. 11 in Zhao et al., 2016; i.e., comparing to Brewer, corrected Pandora data has -0.04 + 0.25% offset in summer and -0.04 – 0.25% offset in winter).

8) Line 181, another mentioning of the “good stray-light control”. Please be more specific In Zhao et al. (2016) “good” is define as low AMF dependence up to 81.6 degrees SZA, or within 1% up to AMF=5.5

Done.

The Pandora and BrT-D instruments have good stray-light control, and under typical ozone conditions (i.e., slant column ozone less than 1500 DU), their air mass dependence is comparably low up to 81.6° SZA (within 1% up to AMF = 5.5; Zhao et al., 2016).

9) Lines 197-198, “bi-weekly” means two weeks? Are you referring to the fact that the SBUV total ozone data are selected within the box centered on the station location, +/-2 degrees in latitude and +/-20 degrees in longitude, and then distance weighted to create the station overpass? What is the uncertainty of SBUV total ozone overpass over Toronto? When comparing to satellite overpass data, do you use the satellite data uncertainty in the estimate of the agreement with Brewers?

Yes, due to a small field-of-view, the SBUV instruments provide global coverage about every two weeks. In other words, for some sites, the true sampling frequency of SBUV instruments can be as low as every two weeks. Thus, the overpass algorithm is used to increase this sampling frequency to daily (Labow et al., 2013), even if the SBUV measurements were not directly overhead of the ground site. It is correct that these daily values are obtained by weighted-interpolating data measured within the box centred on the station locations. Labow et al. (2013) reported that the smoothing errors (the largest error) for total ozone retrievals are mostly less than 0.5%. The uncertainty of individual SBUV total ozone overpasses over Toronto are not available. When comparing to satellite overpass data, we did not include satellite data uncertainty in the estimate of the agreement with Brewers. Most of the published satellite data products used here (except TROPOMI) do not have reported uncertainties associated with each measurement.

Unlike TOMS, OMI or TROPOMI, which provides daily global coverage, the non-scanning, nadir viewing SBUV instruments provide full global coverage approximately bi-weekly. The SBUV ozone column data used in this work is produced by the overpass algorithm to create daily overpass values (Labow et al., 2013; by weighted-interpolating data measured within the box centred on the station location ($\pm 2^\circ$ in latitude and $\pm 20^\circ$ degrees in longitude)).

10) Line 200, the reference to “+/- 1%” is one or 2 standard deviation? This number is based on the monthly averaged comparisons. How does it compare to the results in Table 5 where one standard deviation is provided based on 3-month averaged data?

The $\pm 1\%$ agreement reported by Labow et al. (2013) is the yearly relative difference (time series comparison) between ground-based instruments and SBUV (see their Fig. 1). There are no 3-month or 1-month standard deviations of relative differences that can be used to compare with the current study (Table 5). However, the results from Labow et al. (2013) can be compared with Fig. 5 in this study, but should be interpreted with some level of cautions. For example, the relative difference defined in this work is

$$\Delta_{rel} = \frac{Brewer - SBUV}{\frac{1}{2}(Brewer + SBUV)}.$$

Whereas in Labow et al. (2013), the relative difference was defined as:

$$\Delta_{rel} = \frac{SBUV - Brewer}{SBUV}.$$

In addition, the results in Labow et al. (2013) used an average of 33 northern hemisphere sites. Fig. 1b in Labow et al. (2013) (note 1b is the TOMS V8 total ozone data products that used in this study, 1a is the profile integrated total ozone) shows that the relative differences are in a range of -2 to 4 % (monthly mean) in 1999 to 2010 period (with yearly averages in a range of 0 to 2.5 %). The results from Fig. 5 of this work shows the 3-month relative differences are in a range of -3 to 6 % (also see Fig. R4, which only shows the results from SBUV). We also calculated yearly relative difference which shows Brewers and satellites TCO agrees well within -2 to 3 % (except for SBUV 19 in 2019, which has very sparse coincident observations), as shown in Fig. R5. Thus, we think that the comparison results in this work is in good agreement with previous studies. The description for SBUV series and Fig. 5 (see Section 4.2) has been updated.

Labow et al. (2013) reported that the total column ozone data from Brewers and SBUVs show an agreement within ± 1 % over 40 years (1970-2010; yearly relative difference).

Figure 5 shows the relative differences between satellite and Brewer measurements for seasonal (3 months) values are within ± 4 % and yearly values are within ± 3 % (not shown here) in these two decades (1999-2019).

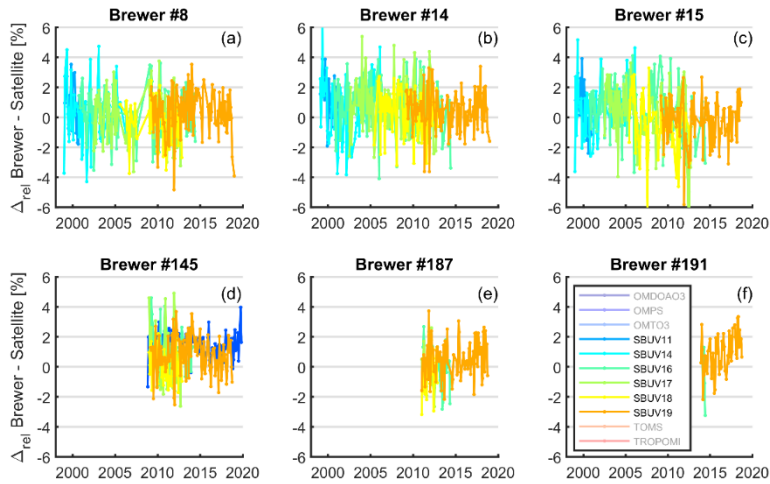


Figure R4. The relative difference between satellites and the world Brewer reference triads (BrT and BrT-D). Same as Fig. 5, but only shows SBUV satellites.

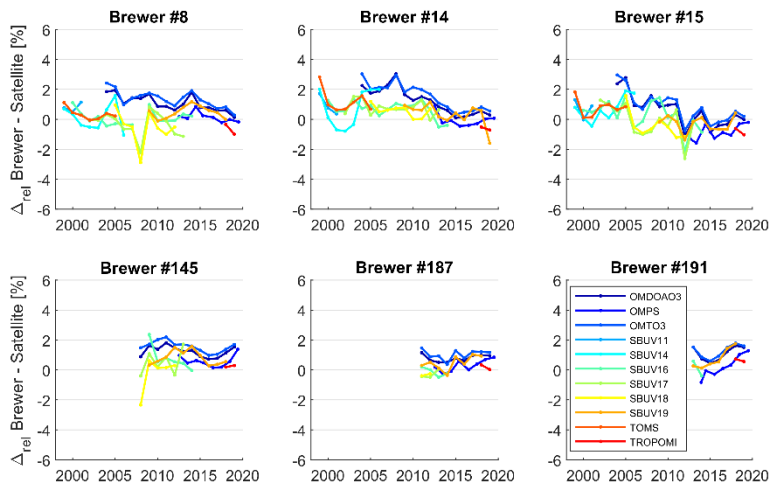


Figure R5. The yearly relative difference between satellites and the world Brewer reference triads (BrT and BrT-D).

11) Lines 251-232. Please explain why the instrument with more points would not dominate the forming of the baseline. Is it in reference to the previous method where three Brewers are used to establish a baseline? In the 3d party method, the baseline is derived for each instrument separately, therefore the 3d party instrument represents the “baseline”?

The referee's comments are correct. For the Model 1 analysis, the baseline ozone is formed by fitting a 2nd order polynomial function with observations from three Brewers for each day. Thus, if one of the three instruments produced more observations than the other two, the calculated baseline ozone will be more representative of that particular instruments (no matter if the real data quality from that instrument is good or not). This issue has been addressed by introducing the index matrix in Eqn. 2, which calculates three "baseline ozone". These three "baseline ozone" share the common curvature (i.e., 2nd and 3rd order terms) but have unique offsets (i.e., A_1 , A_2 , and A_3). However, it is still likely the instrument with more observations may contribute more to the curvature terms.

On the other hand, when using the third-party scheme, i.e., use Pandora TCO as the baseline ozone, we can avoid the issues mentioned above. In other words, the Brewer instrument (no matter if it has more or fewer observations in that particular day) can be "fairly" compared with baseline ozone that is independent of its own observations. The sentence has been modified as to clarify this.

Moreover, when using coincident Pandora ozone data, the baseline will not have the sampling or weighting issues; i.e., the Brewer instrument that reported more data points will not dominate the forming of the baseline (i.e., as the baseline formation in Model 1, see Eqn. 2).

12) Line 268. In this method, B and C are shared between the instruments. In case one of the instruments have a stray light contribution that is larger than in the other two instruments, would it create the offset in the B and C coefficients? Is there a weighting method used to determine these coefficients?

The referee is correct that if one of the instruments has a strong stray light issue, then it may artificially contribute to the curvature of the fitted baseline ozone. For this reason, we do not recommend to use Model 1 to analyze data measured in large AMF conditions ($\mu > 3.5$) for single Brewers. As discussed in the previous question and Appendix A, for moderate AMF ($\mu < 3.5$), both single and double Brewers have reasonable good stray light control, thus currently, we did not use any weighting method in the determination of these coefficients.

13) Line 288. Would the effective absorption cross-section value change with the solar zenith angle due to the presence of the stray light? Do you restrict data comparisons to SZAs that have limited impact of the stray light?

The effective ozone absorption coefficient ($\Delta\alpha$) is quantified by instrument slit functions (determined in DSP test) and the published ozone cross-section. The measured slit functions were acquired with discharge lamps, which might not fully represent the true slit functions of the instruments, especially when stray light became an issue. Thus, the referee is correct that $\Delta\alpha$ is different at different SZA due to stray light. To avoid this, only observations with AMF < 3.5 were included in this work to minimize chance of high slant ozone. Similar to the answer to the previous comment from the referee, we do not recommend to use measurements with large AMF in Model 2 analysis. The sentence has been modified to include these recommendations.

Thus, the difference of total column ozone between the individual instrument and Model 1 is allocated to the “error” of ETC and effective ozone absorption values. As the stray light issue in high μ conditions may affect the formation of the baseline ozone (see Eqns. 2 and 3), all Brewer DS ozone data used in this study have $\mu \leq 3.5$.

14) Line 319 – “only good quality satellite data are used in the analyses”. What are these criteria? Please discuss the QA criteria (flags in section 2.4.

Done.

The EP/TOMS total ozone data from 1996 to 2005 with a quality flag of zero were used in this work (McPeters et al., 1998).

The SBUV ozone column data used in this work is produced and quality assured by the overpass algorithm to create daily overpass values (Labow et al., 2013; by weighted-interpolating data measured within the box centred on the station location ($\pm 2^\circ$ in latitude and $\pm 20^\circ$ degrees in longitude)).

In this work, OMPS-NPP L2 Nadir Mapper (NM) Ozone Total Column swath orbital v2.1 data (only good sample, with a QualityFlags of zero) from the OMPS-NM module is used.

In this work, the OMDAO3 and OMT03 OVP data are used, with L2 quality flag equal to 0 or 1 and bit 6 is not set are included (see <https://avdc.gsfc.nasa.gov/pub/data/satellite/Aura/OMI/V03/L2OVP/OMDAO3/>).

The offline (OFFL v010107) total ozone column data (Garane et al., 2019) are used in this work (only L2 data with $qa \geq 0.75$ are included).

15) Line 341. What was the reason to select 3-months averages for the presentation of results?

This question has been addressed at the beginning of this document. A sentence has been included here.

Using the analytical method from the first assessment work (Fioletov et al., 2005), the deviations and residuals are reported with frequencies of 3 months and 1 year, respectively, in Fig. 1. These frequencies were used because they provide a good balance between sampling frequency and sufficient co-incident measurements as well as preserve a potential seasonal component in the differences.

16) Line 357-358. What is the 3-month mean TO and mean air mass in Toronto in each season? Is it comparable to 330 DU and $\mu=2$?

These values (TCO = 330 DU and $\mu = 2$) were selected based on the statistic of TCO and μ values in Toronto. The time series of 3-month TCO and mean air mass factor (μ) in Toronto are shown in Fig. R6. Also, the histograms of TCO and air mass factors are shown in the right column of Fig. 6. The histograms show that 330 DU is the median TCO values in Toronto, and $\mu = 2$ represents the mid-point of TCO seasonal (3-month) air mass variations from 1.5 to 2.5.

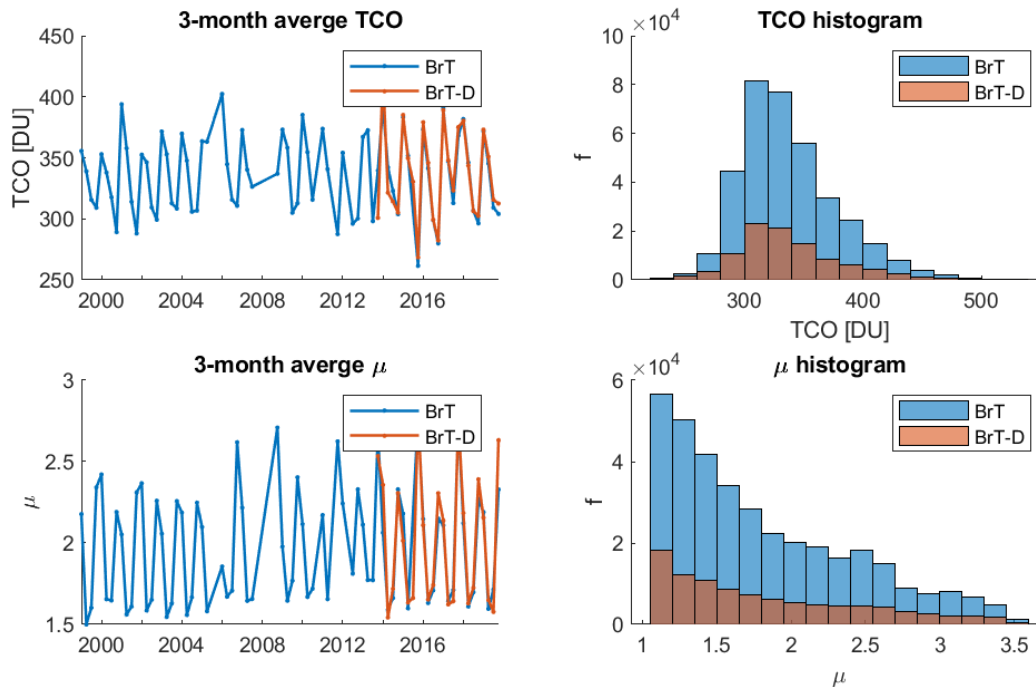


Figure R6. Time series and histogram of TCO and air mass factor (μ) in Toronto.

17) Lines 364-369. Figure 2 suggests a drift in Brewer #14 between 1999-2004 and in Brewer #8 between 2007-2013. Were the drifts corrected in the data archived in WOUDC? According to Table 2, Brewer # 14 was calibrated in 2000, 2005, 2008, and 2013. If the drift is detected between independent calibrations, is there a method to post correct the data prior to the latest calibration reference? Brewer #145 shows a large spike in both errors with the opposite sign. What caused it?

These drifts described by the referee (observed in Fig. 2) have not been modified since they are still within the acceptable error budgets of Brewers. As pointed out in Section 4.1.2, the errors in ETCs and ozone absorption coefficients can largely compensate for each other, thus the derived TCOs may still have “reasonable” values (i.e., 3-month deviations within $\pm 1\%$, as shown in Fig. 1). However, if the errors in ETCs and ozone absorption coefficients are too large, the TCOs measured in a day will have artificial curvature in high μ conditions. Thus, a reprocessing of data from reference instruments was made only when these errors account for more than 3 or 4 % of TCO. An example of this practice was provided in Appendix B, using Model 3. In that

case, Brewer #145's ETC error in early 2014 was as large as 4 % and the ozone absorption error was about 3 % in the operational processing version. Detailed data scrutiny was made and the cause was found (see Appendix B for more details). The re-processing was made to address the issue, with the ETC and ozone absorption errors both being decreased to the acceptable level (± 2 %).

If a drift is detected between the independent calibrations, a detailed investigation will be made together with Brewer technicians and researchers. As an example provided in Appendix B, a post-correction can be made if solid evidence was not only found by models, but also confirmed by Brewer technicians.

The opposite signs in the estimated ETC and ozone absorption errors are due to the nature of the models (2 and 3). The models distribute the residuals (mismatch between observed ozone and baseline ozone) into two parts, i.e., X and Y terms in Eqns. 3 and 4. Thus, retrieved X and Y are negatively correlated. The section has been updated to reflect this information.

The large errors in ETCs and ozone absorption coefficients may largely compensate for each other and not be evident in the Model 1 analysis. This is because Model 2 distributes the residuals (mismatch between observed ozone and baseline ozone) into two parts, i.e., X and Y terms in Eqn. 3, which made the retrieved errors negatively correlated.

18) Line 383, "although empirical correction methods have been applied, the residual effect still exists". Figure 3 shows sudden changes in biases in 2016 and 2017, winter season. Does it have anything to do with this Brewer calibration in 2015? Can you explain instrument issues in this section while discussing Figure 3?

The sudden changes in late 2016 to early 2017 were due to the "imperfect" empirical correction made for Pandora TCO (not due to Brewers). Since the effective temperature used in the Pandora TCO correction model (see Eqn. 10 in Zhao et al. 2016) is calculated with modelled ozone and temperature profiles (ERA-Interim), we might have a large bias in some extreme conditions, especially in winter (e.g., see Fig. 10 in Zhao et al. 2016). On the other hand, after correction, although Pandora TCO has reduced its seasonal deviations from Brewers, one still

can see small residuals (i.e., about 0.4 % seasonable variations, see Fig. 13c in Zhao et al. 2016). This “winter shift” is correlated to the temperatures and still can be seen in Fig. 3 (its depends on the weather; for some years, it is relatively mild).

The original Fig. 3 was made using a monthly frequency to illustrate this issue. In the new Fig. 3 (3-month, see the figure provided in the general comments part), this effect is no longer prominent. Since this feature is not prominent in the modified Fig. 3, we remove the sentence to avoid any misunderstanding of this point.

~~*For example, the absolute differences from the six Brewer instruments all shifted towards positive in the January to February 2017 period.*~~

19) Line 412, Was eq(3) use to derive errors in the ETC and ozone absorption cross-sections? Was total ozone from Pandora used for this assessment?

It was done with Eqn. 4, in which baseline ozone values are not derived from Model 1 (see Eqns. 2 and 3), but observations from Pandora. Total ozone from Pandora is used as the third-party baseline ozone in this analysis.

20) Line 426, When issues with ozone absorption cross-section for Brewer 145 are discussed, what period is referred to? It is not clear from Figure 5(d)

This was discussed with details and provided in Appendix B. Results in Fig. 5d are using the re-processed data from Brewer #145. Thus, it is correct that no clear deviations can be seen. The sentence has been modified.

For example, when analyzing Brewer #145 data, it was revealed by the Model 3 analysis that its absorption coefficients were not ideal (in 2014, see Appendix B for more details).

21) Line 453, Table 5 results need to be discussed in greater detail. For example, if all comparison periods are included in the assessment of the BrT’s errors relative to satellite

overpass data, the mean bias increases to 0.625 %, which is larger than the BrT-D bias. Another interesting fact is that OMTO3 shows the largest bias from both BrT and BrT-D, whereas OMDOAS bias is much lower. TROPOMI bias is negative wrt BrT-D, and it is almost of the same magnitude as of the OMDOAO3, but of the opposite sign. Are the OMDOAO3 and TROPOMI biases related to TROPOMI higher spatial resolution or their respective ozone absorption cross-sections? There seems to be a difference in relative biases for BrT and BrT-D, where BrT-D is often higher (although the difference is not statistically significant) Is there any reason for this? It would be of interest to know of each Brewer calibration results and how much the calibration was changed.

It is correct that if we include results from SBUV11, SBUV14, and TOMS into the calculation, the mean bias will increase. The difference between OMDOAO3 and OMTO3 was also reported by other research works. For example, Antón et al., 2009 reported that TCO from OMTO3 is on average 2.0 % lower than Brewer data, whereas for OMDOAO3 data the bias is only 1.4 %. Thus, we think the findings here are in good agreement with previous works, i.e., OMDOAO3 (0.84 %) has a lesser bias to Brewer data than OMTO3 (1.14 %).

The comparison for TROPOMI is a bit more complicated since the viewing geometry (line of sight) plays a more important factor for such high-resolution satellite data. In this part of the work, to make it a fair comparison, we only used TROPOMI true overpass data (i.e., same as other satellites), without taking into account the difference of viewing geometries between Brewers and satellite. However, this could cause some issues when the stratospheric ozone field has a large gradient, as discussed in Section 5 (see Fig. 8). In short, the opposite signs for BrT and BrT-D's relative biases to TROPOMI should be interpreted with extra caution. In addition, to truly validate the high-resolution satellite ozone data, one will need an improved coincident data selection algorithm. We think that this result will not affect the assessment that we made for the Brewer reference instruments, which is the main goal of this project. We also decided to leave this part within the general satellite comparison and discussion sections to bring the attention of the research community to this high-resolution satellite validation topic.

Regarding the relative differences between BrT and BrT-D, we think it is better to see this effect from Figs. 3 and 7. On average, the relative differences between these two reference groups are within 1 %. Since limited by the random uncertainties of the current Brewer 5-wavelength algorithm (about 0.5 %, as discussed in Section 4.1.1), decreasing these differences will be very challenging. Some of this information has been included in the revised manuscript.

The standard deviation of the Brewer-OMTO3 (OMDOAO3) difference (for 3-month averages) calculated for six instruments is 0.99 % (1.06 %), about 0.5 % higher than Brewers' standard random uncertainties calculated in Section 4.1.1. It is also found that Brewers have lower relative differences compared with OMDOAO3 than OMTO3, which is in agreement with previous researches (e.g., Antón et al., 2009). For high-resolution satellites, such as TROPOMI, the interpretation of the results should be made with extra cautions as the line-of-sight of ground-based and satellite instruments should be accounted for (see more details in Section 6).

22) Line 463, “ the same long-term stability of the Brewer reference instruments when compared with Pandora or satellite instruments”. Figure 7 indicates that the mean bias (by eye) of Brewers in 1999-2004 is near 0% relative to MERRA-2, then it changes to ~2% in 2005 (MERRA-2 change?). The bias in 2005-2015 shows a slow ~1% drift. There is a step-change in 2015 and then it rises to 1% in 2017. Brewer #15 is the lowest in 2006. Brewer #08 is the lowest in 2017-2018. Are all these differences related to the MERRA-2 changes of assimilated data?

The 2% jump in 2005 was due to MERRA-2 changing its assimilation sources from SBUV to MLS/OMI. This information was included in the manuscript (Line 472). As shown in Table 5, we expect that the bias between SBUV and OMI will be propagated to the reanalysis data. The caption of Fig. 7 has been modified to make this more clear.

Figure 7. The relative difference between the reference Brewers and MERRA-2 reanalysis. Each point represents a 3-month average. The green dash line represents the time when MERRA-2 changed its assimilation sources from SBUV-2 to MLS/OMI (causing about 2% relative difference). The black dash line represents the time when BrT-D was relocated to Egbert.

There might be a slow 1% drift as described by the referee in the 2005 to 2015 period. However, given the fact that Brewers also have 0.5 % random uncertainties, we are not sure if this drift is statistically significant. Without uncertainties from the model (in addition to the propagated uncertainties from its assimilation sources), it is difficult to give any solid conclusion on such a small level of variations (i.e., if this 1 % drift is statistically significant or not). It is not always possible to determine the origin of small $\pm 1\%$, differences between different datasets (for example, as in 2015-2017).

The other difference, such as Brewer #015 was the lowest in 2006, is possibly due to a real bias in the instrument. Some of these features can also be found when comparing with Pandora (see Fig. 3). However, the variations that we see in these 3-month relative difference plots are a combination of random uncertainties and biases from both Brewers and the other TCO dataset (e.g., MERRA-2 or Pandora). Similar to our answer to the previous question about satellite comparison, limited by the uncertainties in the dataset (not just from Brewers, but also from other instruments or models), detecting any variation or trend within 1% is very challenging. For example, if we simply assume that both Brewer and the other instrument have 0.7 % total uncertainty, the propagated uncertainties in relative difference will be on an order of 1%. Thus, further fine-tuning or interpretation of the current relative biases found between Brewers and other instruments (or reanalysis) may not be possible.

*The relative difference in time series is shown in Fig. 7, which demonstrated the **similar** long-term stability (i.e., the relative difference within $\pm 2\%$) of the Brewer reference instruments when compared with Pandora or satellite instruments.*

23) Line 476, after October 2004 instead of 2014?

Thanks. The typo has been corrected.

For example, the mean Brewer #014 – MERRA-2 relative bias was 0.11 % (Δ_{rel}') for the SBUV-based data assimilation, but it increased to 1.07 % after October 2004, probably due to some bias in OMI data as mentioned previously in Section 4.2.

24) Line 505-508. The issue with strong temperature dependence in Brewer #15 is discussed. The optical frame was fixed in 2017. Was the data prior to 2017 corrected?

The strong temperature dependence (TD) in Brewer #015 is not ideal, but in this case only, it had limited effects on the data. This is because the wavelength calibration tests (HG) are done regularly, which can largely reduce the impact. However, we should point out that if the time interval between the HG tests is large enough, some measurements can be affected. We included this to illustrate how Brewer hardware problems can affect the overall instrument performance. The relevant text has been modified to clarify this issue.

For example, it was found that Brewer #015 has a particularly strong temperature dependence where the optical frame was expanding significantly faster than any other Brewer instrument. As a result, the wavelength calibration tests (HG) had to be scheduled more frequently to reduce the impact. However, we should point out that if the time interval between the HG tests is large enough, some measurements can be affected. This issue was fixed in 2017 by replacing the optical frame (details of instrument repair and upgrade history is provided in the supplementary information).

25) Lines 508-510. Wavelength drift in Brewer #145 is discussed. It would make sense to mention instrumental issues while discussing results in Figures 3 and 6.

We agree with the referee that it makes sense to provide more details of instrumental issues while discussing the figures provided. So, we included some discussions for Brewer #015 and #145.

A second example is the original configuration of Brewer #145 micrometer was found to have developed wear and became unreliable, causing some wavelength drifts, and as a result,

relatively high uncertainties for Brewer #145 as shown in Table 7 (also see larger variations of 3-month deviations from Brewer #145 compared to Brewers #187 and #191 in Fig. 1a).

26) Line 513 – Hardware replacement issues, ie. ND filter and mercury bulb. What is the recommendation to the BrT and BrT-D data reprocessing? Please make sure to refer here to Appendix B.

We think that it is more appropriate to include some recommendations for BrT and BrT-D data reprocessing in the next paragraph.

However, this approach raises the question of reproducibility of the obtained results and must be carefully documented. For BrT and BrT-D's data reprocessing, we recommend using the statistical models developed in relevant studies to help the identifications of potential hardware or software issues. To keep the integrity of the world reference instruments, data reprocessing could be done only if solid evidence of imperfection of hardware or software have been found and confirmed by Brewer technicians and researchers.

Data availability section: There is no link to TROPOMI data Brewer data for triad is not accessible through WOUDC.

The Brewer triad data has been uploaded to WOUDC.

TROPOMI data are available from <http://www.tropomi.eu/data-products/total-ozone-column>, last accessed: October 2020.

Response to Referee #2:

We thank referee #2 for their helpful comments. Our responses are given below in black with the referee's comments in blue. The new/revised text in the modified manuscript is given in red (italicized).

This is a good update of the work of Fioletov et al. [2005] addressing the precision on the WCC triad, with interesting model comparison introducing external instruments in the assessment. However, three important topics are not addressed in this work, 1. How absolute calibration is done. 2. How the calibration are maintained between absolute calibrations. 3. How the calibration is transferred to the traveling instrument and then to the Brewer network.

As recommended by the referee, absolute calibration procedure, maintenance, calibration transfer, and assessment of travelling standard should be detailed described and published. Together with the assessment of triads, these works will provide some general, but important pictures of the Brewer ozone monitoring network. Thus, the suggested work has been included in our project plan. We will coordinate with other relevant institutes and prepare the second publication in the near future since it was discussed and recommended at the recent meeting of the WMO GAW Scientific Advisory Group on Ozone and UV. However, the purpose of this study is to demonstrate the long-term stability of the existing Brewer reference standard (the Brewer triad). Some of these information has been included in the revised manuscript.

Thus, it is critical to review and assess the world reference instruments' performance on a regular basis. This study's focus is on the demonstration of the long-term stability of the existing reference instrument. Absolute calibration procedure, maintenance, calibration transfer, and assessment of travelling standard will be a subject of a separate study.

Simultaneous observations are required for the calibration transfer of the Brewer, so it seems feasible to have enough simultaneous measurements over a month to derive the calibration constants of the Brewer triads, using every Brewer as a reference to calibrate the others. This will produce a monthly series of the calibrations constants (F_0 , α) to compare with model results.

This was essentially done by the statistical Model 2. Instead of comparing constants instrument by instrument (that makes it difficult to interpret the results), the Model 2 estimates deviations of the

constants for each instrument from the “best” value based on all measurements for each 3-month period for the entire 20-year long triad record. If a calibration constant is different from the value “prescribed” by the two other instruments, that the estimated errors for the instrument would appear as an outlier. This gives information about long-term changes in the constants and the overall triad stability.

There is no mention of the number of observations in the study. In contrast with other studies there is no plot of the simultaneous measurements (see for example Figure 3 of [Stübi et al., 2017]). Observing at the hourly data set used for the comparison with the reanalysis, we can almost get a view of the differences without using any average. In general, the figure are difficult to see, especially if they are printed, because the several curves in the figure are not easily to distinguished. I suggest extending both axis for a clearer view, and using consistent symbols for BrT and BrD representation. In addition, I also suggest indicating the dates of the calibrations on the graphs.

The analysis was done based on individual measurements and only the results are presented in the form of long-term plots. Following suggestions from the referee, we also plotted the measurements from six reference instruments with the absolute calibration dates indicated. This figure has been included in the supplementary information. In the manuscript, all six reference instruments were plotted with consistent unique colours (e.g., consistently using blue colour for Brewer #008 and consistently using red colour for Brewer #014 when results from all six instruments were presented together).

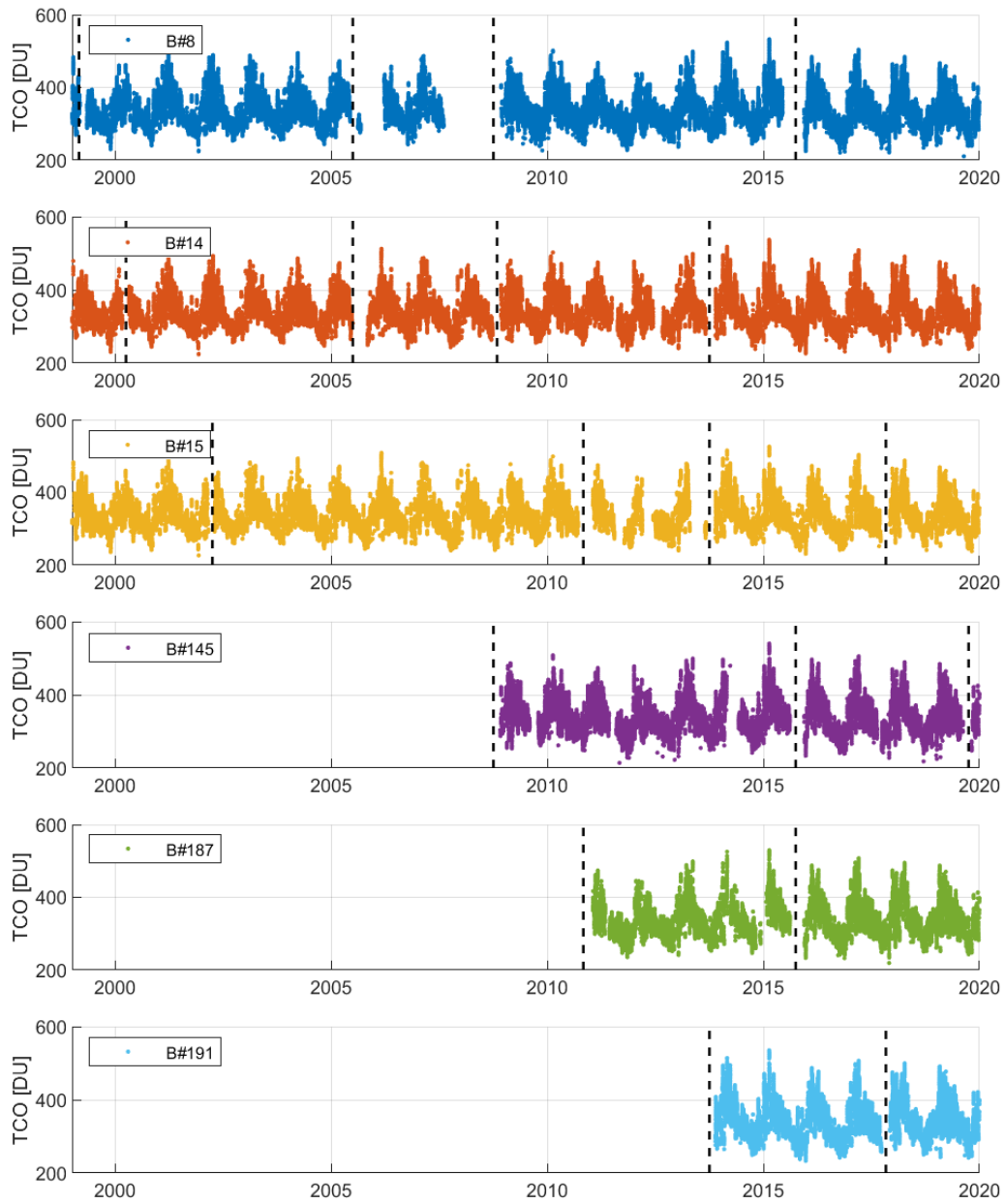


Figure S1. Time series of Brewer TCO observations in Toronto. Vertical black dash lines indicate the time of primary calibrations as shown in Table 2.

General Comments:

1. The independent calibration of the instruments is not described. As the authors say, (line 80) The absolute calibration is "critical to review and assess the ... instrument performance", but there is no description of the methodology used, the results of the calibration and the level of agreement with the results of this work.

The purpose of this work is to evaluate the triad performance based on the existing calibration results. The sentence has been modified to provide a reference to the procedures of independent calibration. The results of the independent calibration are ETC values, which have been used to produce the TCO values reported. The TCO values have been examined by Model 1 and compared with Pandora, satellites, and reanalysis data. The ETC values themselves were been evaluated via Models 2 and 3 analyses, which provided the estimated errors of the ETCs. In general, we think this work has already provided an assessment of the results of the independent calibration via the analyses mentioned above. Please note that Models 2 and 3 were designed to estimate the errors of ETC and effective ozone absorption coefficient, but not ETC or effective ozone absorption coefficient themselves.

The extraterrestrial calibration constant (ETC) has to be determined in the field by one of the two means: 1) the independent calibration method, i.e., the Langley plot calibration method or the so-called zero airmass extrapolation technique, or 2) the calibration transfer method (e.g., transfer ETC from well-calibrated reference instruments to field instruments) (see more details about calibration procedures in Kerr, 2010).

2. The number of calibrations of the instruments is low, in the period of 20 years analyzed BrT instruments were calibrated four times, on average every 5 years. While brewer instruments of the Network for detection of the Atmospheric Climate Change are requested to be calibrated every year and WMO recommends a two-year cycle calibration. It is crucial to know how the calibrations are maintained between absolute calibrations.

In a perfect world, Brewer would be calibrated just once and then most of the changes in the instrument characteristics would be tracked and corrected by mercury and halogen lamp tests. We have seen examples of such long-term stability at the South Pole where the instrument was operated without any additional calibrations for seven years.

The requirement about frequent calibrations is mostly based on the need of regular instrument maintenance that many operators cannot carry out by themselves (and to characterize the changes of the slits). In the case of the triad, such maintenance is done regularly. Between the calibrations, the constants were tracked by the lamp tests and the ETC was adjusted accordingly.

When the primary calibration has been done for one of the reference instruments at MLO, this instrument can be used to validate the status of other reference instruments in Toronto. So, to satisfy the 2-3 years interval between the calibration requirement, it is sufficient if at least one triad Brewer is calibrated at MLO every 2-3 years.

3. The transfer method from the triad to the travelling reference need to be clarified. Which of the instruments are used for transfer? What ozone data do you use for the transfer? That from the BrT or the straylight-free data? The observations from the BrT, BrT-D or an average of all six instruments? Which period of time is used for the calibration of the traveling reference.

The calibration process for the traveling references is the same as for any transferred calibration (Savastiouk, 2006): the instrument to be calibrated is assessed to make sure that its hardware is working properly, all the necessary characterization tests are done, simultaneous direct-sun data are collected with the triad instruments, and an average of BrT is used using $1.2 \leq \mu \leq 3.2$ for $\text{TCO} \leq 350$ DU to establish the ETC. However, the ozone calibration transfer is beyond the scope of this paper.

4. Different updated versions of the model Fioletov et al. [2005] have been used to establish the performance of the Brewer instrument, but this method is not used for the satellite and reanalysis comparison. For validation of this model a comparison of the triads using hourly observations (as reanalysis) may be of interest.

In this work, we used three statistic models. Models 1 and 2 used here are strictly following the model designs described in Fioletov et al. 2005. Model 3 is a new one, or more precisely a modified Model 2. Here, Model 1 was used to directly assess the performance of reference instruments by examining their measured TCO values. Models 2 and 3 were used to examine the errors of ETC and the effective ozone absorption coefficient. The major difference between Models 2 and 3 are that they are using different “baseline” ozone. The former one uses the baseline ozone values derived from Model 1; the latter one uses the values from Pandora measurements.

We thank the referee for the suggestion to compare triads with satellite and reanalysis data using Model 3. The referee is correct that for Model 3, we can use any baseline ozone, as long as it is from a third party. However, we should note that this baseline ozone should have equivalent or better sampling frequency than that from Brewers. Given that satellites have a much lower frequency (about daily), we did not use them in Model 3. It might be possible to use the hourly reanalysis data, but the model also has uncertainties propagated from its assimilation sources (e.g., see Fig. 7, the “shift” in 2004). Thus, using reanalysis data would make the assessment for triad more complicated (i.e., we cannot easily separate the errors from the reanalysis model, satellite instrument, and Brewers). On the other hand, Pandora was selected because of its good precision (about 0.5 DU, see Zhao et al. 2016) and high sampling frequency (less than 5 minutes).

5. The Methods 2 and 3 also evaluate the error in the Extraterrestrial constant and absorption coefficient. These parameters are also obtained during the calibration, but no comparison is made between the model-derived parameters and those obtained when the instrument is calibrated.

As provided in previous responses, Models 2 and 3 do not generate an estimation of ETC or effective ozone absorption coefficient themselves, but their estimated errors. Thus, the model-derived parameters were used to evaluate the performance of the Brewers but not directly compared with those calibration constants (ETC or effective ozone absorption coefficient). As discussed in Sections 4.1.2 and 4.1.3, by the nature of errors in ETC and or effective ozone absorption coefficient, they may compensate each other and produce “reasonable” final TCO data products, despite the errors of themselves might be relatively large. Thus, we recommend not only examining the deviations of TCO values (e.g., Model 1), but also performing suggested Models 2 and 3 analyses for Brewer triads. An example of this practice was provided in Appendix B.

6. The Stray light effect on the ozone is the power law of the ozone slant column Karppinen et al. [2015] Moeini et al. [2019], although the observations are limited by air mass (3.5) and not by ozone slant column. A Brt to BrtD comparison against the ozone slant column may give us the correct limitation of the ozone slant column for the analysis.

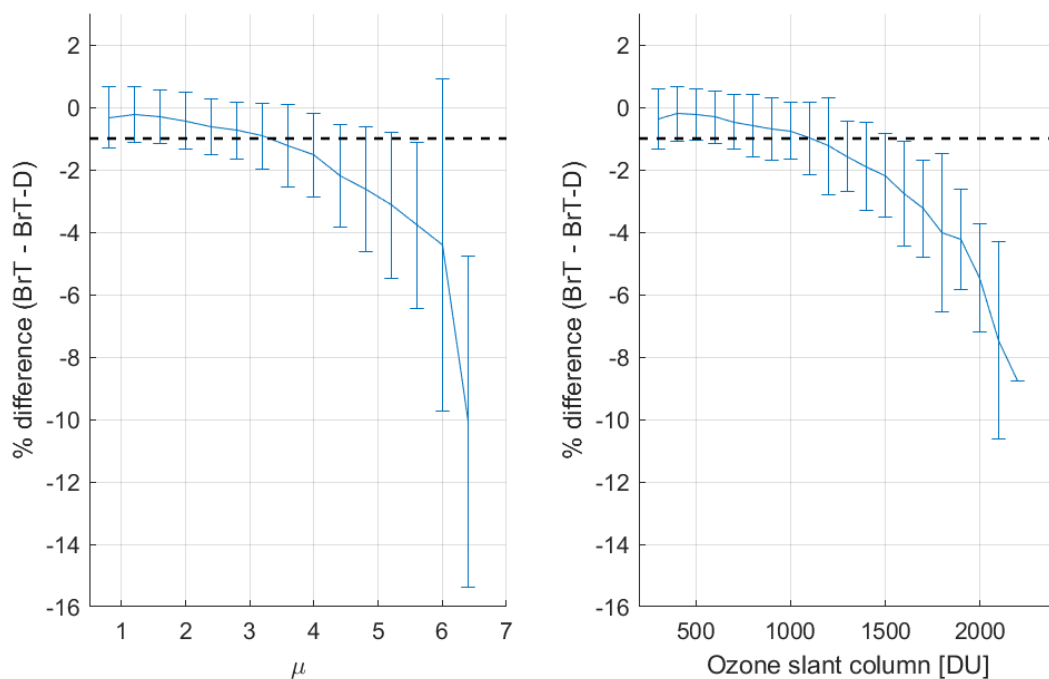


Figure R3. The relative difference between BrT and BrT-D, in terms of air mass factor (μ) and slant column ozone. The error bars represent 1σ of the relative difference values. The black dash lines show the -1 % relative difference.

We fully agree with the referee that the slant column is a governing factor. The 3.5 limits of air mass factor were examined and validated by previous works conducted in Toronto with Brewers, which should address the stray light effect in single Brewers sufficiently. As suggested by the referee, analysis of the percentage difference between BrT and BrT-D is provided in Fig. R3, in term of a function of both air mass factor (μ) and slant column ozone. In general, they provided the same picture, i.e., BrT and BrT-D start to have more than 1% relative difference when $\mu > 3.5$ (equivalent to slant ozone 1200 DU).

Unless stray light correction is implemented, a filter based on slant ozone amount is impractical as it can easily allow poor data to go through if TCO is high, e.g., TCO = 300 DU ($\mu = 4$ and slant column = 1200 DU) will be calculated as TCO = 250 DU ($\mu = 4$ and slant column = 1000 DU) if stray light is present and pass the filter of 1000 DU. Moreover, any filtration based on slant column may introduce a bias in the data since low values would pass through the filter, while high values would not.

The manuscript has been revised to include some of these information and the Figure R3 has also been included in the supplement file (as Fig. S2).

It is found that the air mass dependencies of BrT and BrT-D are consistent within these two periods. Further information on relative difference between BrT and BrT-D, in terms of air mass factor and slant column ozone are provided in Fig. S2.

7. The use of different timescales, monthly, three monthly or six monthly make the comparison of the different models difficult Please unify the results.

Following this suggestion, analyses are now using a consistent 3-month frequency.

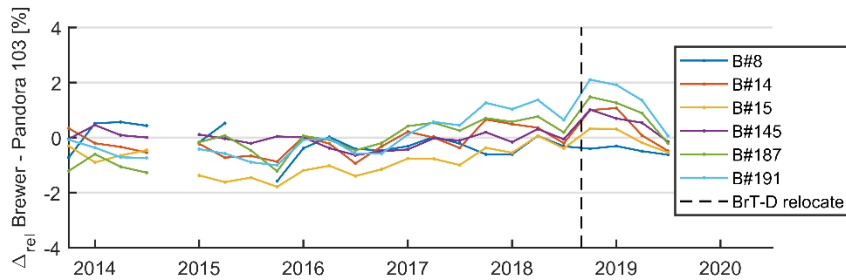


Figure 3. **3-month** relative differences between Brewers and Pandora total column ozone. **3-month** averages are calculated if there are at least ten coincident measurements between Brewer and Pandora for that period. The black dash line represents the time when BrT-D was relocated to Egbert, i.e., Pandora and BrT-D were not co-located.

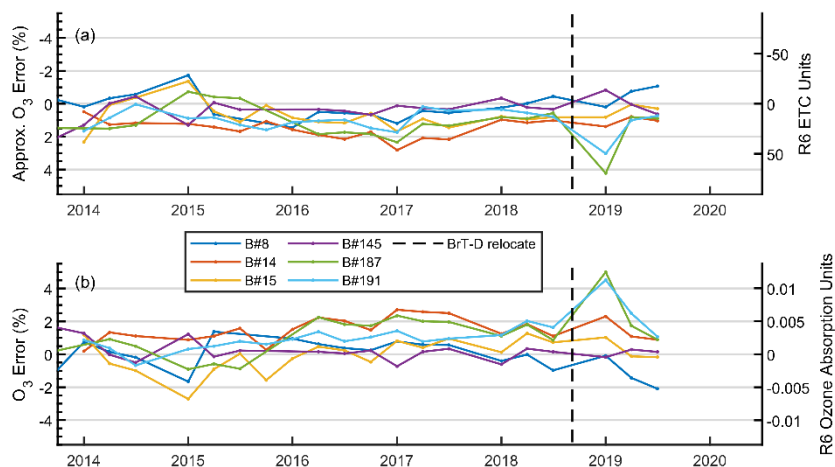


Figure 4. Relative systematic uncertainties in ETCs and effective ozone absorption coefficients estimated using Model 3. Description of y-axes is in Fig. 2. Each point on the graph represents a 3-month average. The black dash line represents the time when BrT-D was relocated to Egbert.

8. Results of the regular standard lamp tests of the Brewers, normally a good indicator of the stability of the instrumental calibration. A comparison of these measured SL-test records with the presented statistical parameters should be included and hopefully show the same good stability.

The measured SL-test records are included in the data processing. Thus, the ETC values have been corrected based on the SL tests. The SL test is not a measure of data quality, but a measure of instrument’s spectral sensitivity changes that are applied to the data processing. Having relatively stable SL results are of little importance if not properly used in data processing and, conversely, even large variability in SL test results can be successfully used to correct the data (Lam et al., 2007). As the SL corrections have been made within BPS and the Model 2 used BPS outputs as input, directly comparing the SL records and Model 2 outputs may not be very meaningful. We have included more discussions about the data quality assurance in the revised manuscript.

The way that the data are processed also affects the results. Siani et al, (2018) concluded that the ozone data processed by different software agree at the 1 % level; however, some differences can be found depending on the software in use. They also recommended “a rigorous manual data inspection” of the processed data and to be careful with how Standard Lamp (SL) test results are used. Visual data screening was also used by Stübi et al., (2017b) to eliminate outliers. However, this approach raises the

question of reproducibility of the obtained results and must be carefully documented. *For BrT and BrT-D's data reprocessing, we recommend using the statistical models developed in relevant studies to help the identifications of potential hardware or software issues. To keep the integrity of the world reference instruments, data reprocessing could be done only if solid evidence of imperfection of hardware or software been found and confirmed by Brewer technicians and researchers.*

Specific Comments

3.1. Page 1 Line 27 Reference to the WMO requirements document is missing.

The reference to WMO has been included.

The random uncertainties of individual reference instruments are within the WMO/GAW requirement of 1 % (WMO, 2001; 0.49 % and 0.42 % for BrT and BrT-D, respectively as estimated in this study).

3.2. Page 1 Line 27 Reference to the uncertainty analysis is missing.

The uncertainty analysis, i.e., 0.49% and 0.42% reported here, was made by this work.

The random uncertainties of individual reference instruments are within the WMO/GAW requirement of 1 % (WMO, 2001; 0.49 % and 0.42 % for BrT and BrT-D, respectively as estimated in this study).

3.3. Page 2 Line 49 random uncertainty? Please use standard meteorologic terminology

Data analysis from this study shows that the precision of individual observations are within ± 1 % in about 90 % of all measurements.

3.4. Page 2 Line 53 Please update Stray Light correction references, [Karppinen et al., 2015], [Rimmer et al., 2018]

New references have been included.

Internal instrumental stray light affects measurements made with the single-monochromator instruments; therefore, corrections are applied to the data when necessary (Bais et al., 1996; Fioletov et al., 2000; Karppinen et al., 2015; Rimmer et al., 2018).

3.5. Page 3 Line 63 The Arosa triad is now in Davos at PMOD World Radiation Center([Stübi et al., 2017])

This information has been updated.

The Arosa triad (Stahelin et al., 1998; Stübi et al., 2017b), formed in 1998, was the second Brewer triad worldwide (composed of two Mark II and one Mark III instruments; now in Davos at PMOD World Radiation Center (Stübi et al., 2017a)).

3.6. Page 3 Line 65 Reference comparisons are described in [Redondas et al., 2018]

This information has been updated.

The regional reference instruments are regularly compared to the world reference instruments via a travelling standard (Redondas et al., 2018).

3.7. Page 3, Line 80 The instrument calibration every 2-6 years ?, the range looks 3-8 years.

This information has been updated.

Each individual reference instrument is independently calibrated at the Mauna Loa Observatory (MLO), Hawaii (19.5° N, 155.6° W, 3400 m asl), every 3-8 years (see Table 1) via the Langley plot calibration method.

3.8. Page 5, Line 116 Please detail the configuration of the BPS. Was this software also used for the previous Fioletov et al. [2005] analysis? Which are the main differences?

The software used in this study is the same as that used in Fioletov et al. (2005). The text has been revised to reflect this information.

Brewer data was processed by Brewer Processing Software (BPS) developed by ECCC (Fioletov and Ogyu, 2008). The same processing software was used in Fioletov et al. (2005).

3.9. Page 5, Line 113 Please associate the references with the corresponding product

The sentence has been modified to associate the references with corresponding products.

The Brewer spectrophotometer provides data products that include column ozone (e.g., Kerr, 2002; Kerr et al., 1981), column sulphur dioxide (SO₂; e.g., Fioletov et al., 1998; Zerefos et al., 2017), column nitrogen dioxide (NO₂, by Mark IV only; e.g., Cede et al., 2006; Kerr et al., 1988), spectral UV radiation

(e.g., Bais et al., 1996; Fioletov et al., 2002), aerosol optical depth (e.g., Kazadzis et al., 2005; Marenco et al., 2002), and effective ozone layer temperature via group-scan technique (Kerr, 2002).

3.10. Page 6, Line 140 What is an independent calibration technique? Please clarify.

Additional information has been included.

As previously described, to maintain the high precision of all Brewer instruments (i.e., transfer the F_0 value), the world reference instruments (BrT and BrT-D) receive their F_0 values via the independent calibration technique. In short, these high-precision F_0 values were determined by fitting the measured F values as a linear function of air mass factor (see Eqn. 1). For example, in clear sky conditions with stable ozone values, if measurements are made under a range of air mass factors throughout a day, then the intercept of the linear fitting of $(F + \Delta\delta m)$ versus μ will be F_0 . More technical details, such as calibration periods, averaging, and why MLO is the ideal site for this practice are provided in details in Kerr 2010.

3.11. Page 7, Line 180 Please indicate the Pandora calibration.

Unlike Brewers, Pandora instruments do not need to perform the independent calibration at MLO. Some of these details, e.g., construction of extraterrestrial spectrum, were provided in the first paragraph of this section (Section 2.2). A new sentence has been included to indicate the Pandora calibration as suggested.

The Pandora and BrT-D instruments have good stray-light control, and their air mass dependence is comparably low up to 81.6° SZA (within 1% up to $AMF = 5.5$; Zhao et al., 2016). Benefitting from the TOAS technique, unlike Brewers, Pandora instruments do not need the independent calibration at MLO (Tzortziou et al., 2012).

3.12. Page 8, Line 165 Are Serdyuchenko cross sections used in this work? Please clarify.

All Brewer data used in this study are based on Bass-Pour 1985 effective ozone absorption coefficient. Although Serdyuchenko cross sections are recommended, they have not been widely implemented on the global Brewer network yet. Pandora (entire PNG) data was using Serdyuchenko cross sections.

Another major difference between the Brewer and Pandora retrieval algorithms is their selection of ozone cross-section, i.e., the Brewer uses BP (Bass-Paur) ozone cross-section (at 228.3° K, Bass and Paur, 1985) and the Pandora uses *Serdyuchenko* ozone cross-section (at 225° K, Serdyuchenko et al., 2014).

3.13. Page 8 , Line 170 Can you please summarize the differences between the official Pandora observations at Downsview that can be obtained from the Pandora Global Network, and the ones used in this work? Are the observations used here also publicly available?

The effective temperature-corrected Pandora TCO data is not available on PGN. The major difference between the official Pandora TCO and the corrected TCO is their temperature sensitivity using empirical formula as described by Zhao et al., (2016). We have modified the sentence to summarize the differences between the official and corrected Pandora TCO. We have upload the corrected TCO data to ECCC's public data server and can be downloaded from:

https://collaboration.cmc.ec.gc.ca/cmc/arqi/Zhao_et_al_amt-2020-324/

The effective temperature was calculated from temperature and ozone profiles provided by ERA-Interim (Dee et al., 2011). In general, after correction, the multiplicative bias in Pandora ozone data can be decreased from 2.92 to -0.04 %, with the seasonal difference (estimated with monthly data) decreased from ±1.02 to ±0.25 % (see Fig. 11 in Zhao et al., 2016; i.e., comparing to Brewer, corrected Pandora data has -0.04 + 0.25% offset in summer and -0.04 – 0.25% offset in winter).

3.14. Page 8, Line 170 StrayLight (ozone slant column dependence), see general note 9.

The general note 9 is missing in the referee's report. Line 170 is not directly related to stray light. The information of Pandora stray light was discussed in comparison with Brewers in Appendix A.

3.15. Page 10, Line 220 Can you quantify the good quality of MERRA total ozone, for example, the BIAS and standard deviation with ground base?

The relative differences, biases, and standard deviations between Brewers and MERRA-2 were provided in Section 4.2 (see Fig. 7 and Table 6).

3.16. Page 10, Line 245 "the baseline is only needed to adjust for the time difference in ozone measurements by individual Brewers" How large is the time difference between the measurements of

the Brewers of the Triads? Couldn't they run in sync? If all the Brewers were in sync, would the baseline calculation (Ai coefficients still be needed?)

It is correct that if all reference instruments' observations are synchronized, then we will not need to follow the design of Model 1 (i.e., use Ai coefficients to evaluate the deviations). Normally, Brewers can have one DS ozone observation about every 4 to 5 minutes, so theoretically it is possible. However, depending on the measurement schedules, Brewers may be operated in several different modes. For example, we plotted the DS TCO measurement intervals (i.e., the time gap between two successive DS TCO observations) in Fig. R7. It shows that the true DS TCO observation intervals can vary from about 5 to 30 minutes. Only less than 50 % of the observations were made with a "perfect" time interval, i.e., about 5 min. Also, a complete synchronization of schedules is not possible since the instruments perform different tasks, e.g., Brewer #015 is used for Umkher measurements and Brewer #014 is the main instrument for spectral UV measurements at Toronto.

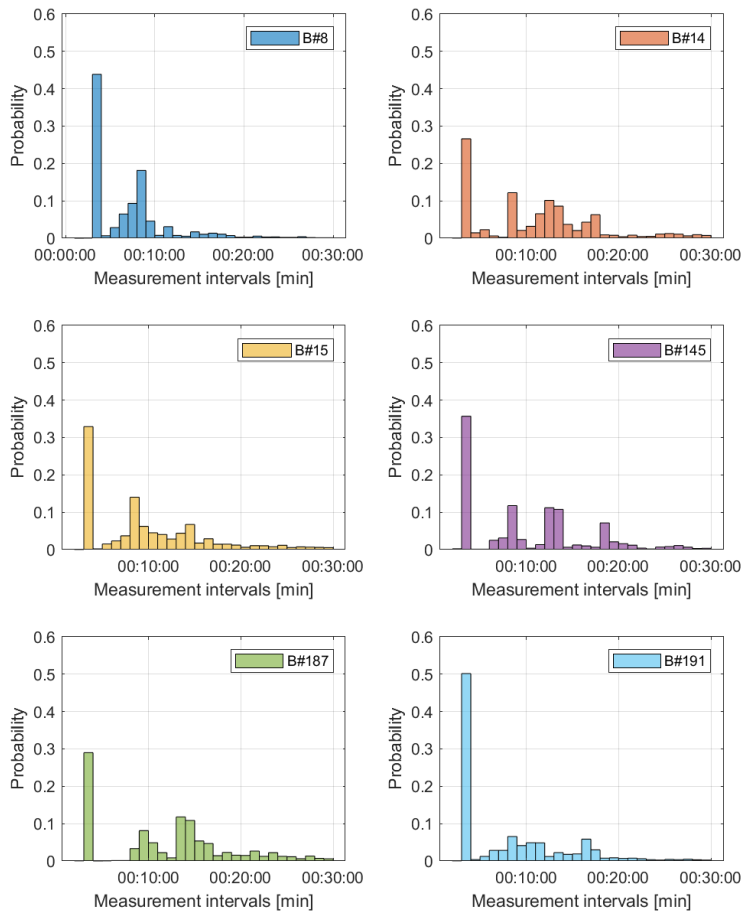


Figure R7. Probability of Brewer reference instruments' direct-sun TCO observation interval.

3.17. Page 12, Line 307 As the triads receive its ETC independently, can be used as ozone for the model 3.

The referee is correct that a Brewer that received ETC via independent calibration can be used in Model 3 to provide the “baseline”. However, we are reluctant to do this due to the sampling issue discussed in the previous question. Also, as the goal of this work is to assess the performance of all six reference instruments, we think that using one of them as a “baseline” is not ideal.

3.18. Page 13, Line 325 The total ozone above 400 DU are usual in Toronto and with 3.5 airmass limit means 1400 ozone slant column, so this observations are seriously affected by stray light. Why the double brewer are also limited in airmass?

A 3-month time series and histogram of TCO and air mass factor in Toronto were made when answering another question from referee #1. Here we presented it again below (see Fig. R3). For Toronto, the median TCO values are about 330 DU. Also, the stray light effect has been discussed and proved to be low with current selection of filters (see previous answers and Appendix A). We should point out again that unless stray light correction is implemented, a filter based on slant ozone amount is impractical as it can easily allow poor data to go through if TCO is high.

The referee is correct that double Brewers have much better stray light control and can provide data up to air mass factor 5 (see Figs. A2 and R3). However, in this work, since we want to provide the same assessment for both BrT and BrT-D, the same filtrations were made, i.e., air mass factor ≤ 3.5 . However, the stray light performance of BrT and BrT-D was also examined and discussed in Appendix A.

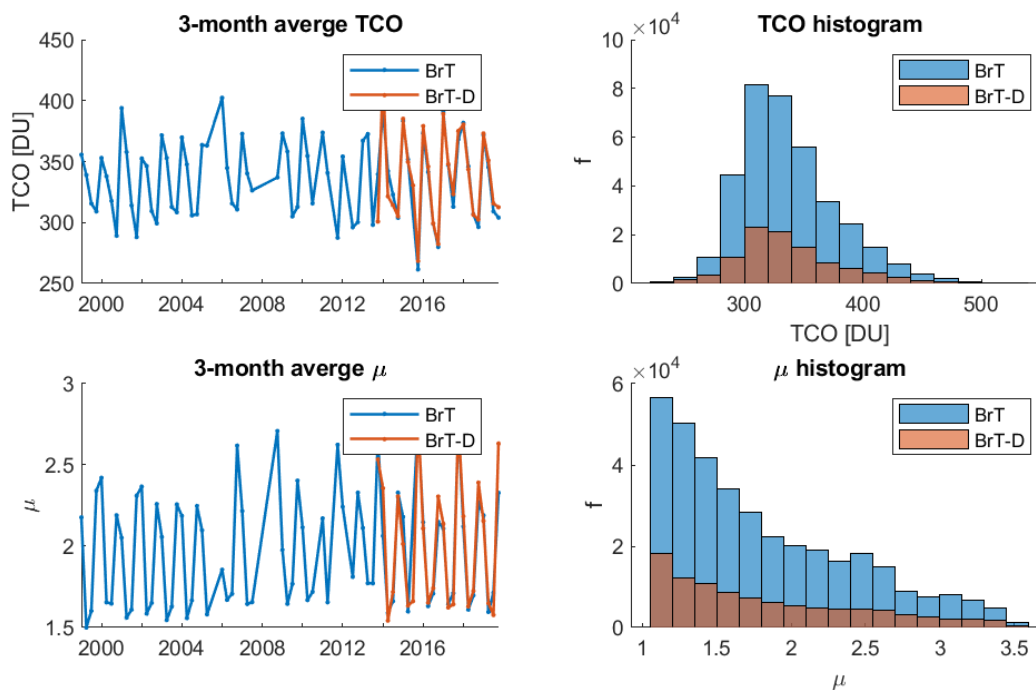


Figure R6. Time series and histogram of TCO and air mass factor (μ) in Toronto.

3.19. Page 14, Line 341 σ' is not defined is it the mean? In that case, it would be better to use $\bar{\sigma}$

The symbols have been updated.

The standard deviations (σ) of the 3-month averages plotted in Fig. 1a are 0.43 %, 0.36 %, and 0.42 % ($\bar{\sigma} = 0.40\%$) for Brewers #008, #014, and #015, which are comparable to the reported values from 1984 to 2004 (0.40 %, 0.46%, and 0.39 %). The double triad also shows good long-term stability with the Model 1 analysis, where all measurements are within $\pm 1\%$ compared to its baseline. The standard deviations are 0.44 %, 0.26 %, and 0.33 % ($\bar{\sigma} = 0.34\%$) for Brewers. #145, #187, and #191. From this, assuming that the instrument uncertainties are independent, the standard uncertainty of Brewers (δ) can be estimated as $\sqrt{1.5}\bar{\sigma}$, i.e., 0.49 % and 0.42 % for BrT and BrT-D, respectively.

3.20. Page 15, Line 358 It looks like there is a factor 10 missing on the formula.

R6 is simply a linear combination of measured intensities in a modified scale. Depending on the values' scale, it may or may not need to be divided by a factor of 10 to get column ozone value. In the Brewer software, R6 is a linear combination of $10000 \times \log(I)$ so to get ozone in DU it need to divide by 10. To make this more clear and consistent with Eqn. 1, the formula has been modified. More description of the models has also been included.

Here, the errors in the ETCs and effective ozone absorption coefficients are estimated in R6 ratio units (the units used in the actual Brewer processing algorithm; R6 values corresponding to measured slant column, i.e., $\Omega = \frac{(R6 - ETCO_3)}{10\Delta\alpha\mu}$ in DU; $ETCO_3 = -10^4 \times F_0$). The errors are converted from R6 ratio units to percentages of total column ozone by using typical conditions for Brewer measurements in Toronto (i.e., $\Omega = 330$ DU, $\Delta\alpha = 0.34$, and $\mu = 2$), to provide more straightforward values to assess the impact of errors in the ETCs and effective ozone absorption coefficients. For example, if we have a model estimated error of $ETCO_3$ as 50 R6 ratio unit, it will correspond to $\frac{x}{10\Delta\alpha\mu\Omega} = 2.2\%$ of total column ozone using the typical conditions described above.

3.21. Page 15, Line 360 For the uncertainties of ETC, the goal is to have it within ± 5 R6 ratio units. Please can you clarify which are the typical conditions, and how are these threshold parameters are obtained?

The goal in a calibration is to have an uncertainty in ETC of less than 1% effect on the TCO. Having an average TCO = 300 DU, $\Delta\alpha = 0.33$ and $\mu = 1$ (worst case for error), we can calculate that a 5 unit uncertainty in ETCO₃ gives about 1.5 DU, or 0.5% uncertainty in ozone.

3.22. Page 15, 370 The goal of ETC and ozone absorption coefficient should be plotted also as reference.

The real output of Brewer is TCO, and the goal of its random error is ± 1 %. This is achieved by high precision ETC and $\Delta\alpha$ together. The estimates of ETC and $\Delta\alpha$ errors just provide values that best distribute the fitting residuals between “baseline ozone” (from Model 1) and measured ozone from one instrument. This means that the estimated errors here are the upper limits of the real errors within the ETC and $\Delta\alpha$. As shown by Fig. 2, these two estimated errors will compensate each other and make the “combined” error of TCO (or more precisely, the real error of TCO) within ± 1 %. Thus, plotting the goal of ETC and $\Delta\alpha$ on Fig. 2 will be misleading, e.g., reader might think the calibration results failed to meet the goal.

3.23. Page 16, Figure 2 Could you add the calibration dates to this figure For Brewer #008, it looks like the error is increasing over the last three years of the period between the 2008 and 2015 calibrations?

Figure R8 has been made below with primary calibration dates for Brewer #008 included. Though the Triad instruments are exceptionally important and we attempt to ensure they run without a flaw, we cannot prevent component failure. In this case, the analog-to-digital (A/D) board failed just before the dates indicated by the referee (see more details in the supplementary information). All indications were that the replacement of the A/D board returned the instrument to normal operation and the replacement of this particular board does not affect Brewer characteristics; however, the instrument was disturbed. It was opened for repair, which does open the possibility for some issues, such as humidity changes (which may initiate a change in the NiSO₄ band pass filter). These conditions would not be easily noted, especially when the differences between reference instruments were within ± 1 %. The 2015 instrument review is a normal review of the instrument in preparation for absolute calibration. Doing the review before-hand minimizes instrument refurbishment time and maximized absolute calibration measurements in MLO.



Figure R8. Modified Fig. 2, with primary calibration dates for Brewer #008 indicated on panel (a) by vertical dash lines.

3.24. Page 17, Line 395 Figure is difficult to see.

We have adjusted the figure to have larger fonts. Following the suggestion from referees, the figure also was updated with the new analysis frequency, i.e., 3-month.

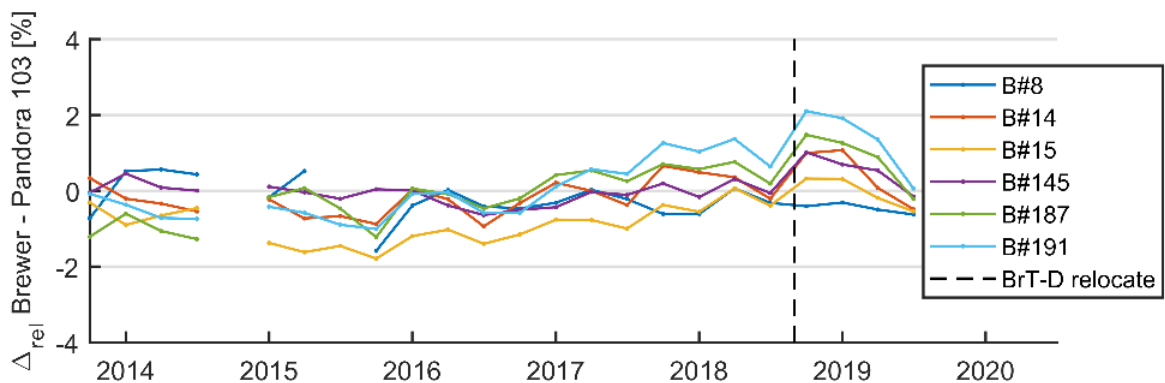


Figure 3. **3-month** relative differences between Brewers and Pandora total column ozone. **3-month** averages are calculated if there are at least ten coincident measurements between Brewer and Pandora for that period. The black dash line represents the time when BrT-D was relocated to Egbert, i.e., Pandora and BrT-D were not co-located.

3.25. Page 18, Line 420 Table 4: for comparison, we suggest to include the results of Model 2

Following the suggestion, the results of Model 2 are also included in Table 4. Please note here Model 3 results have also been updated since Fig. 4 has been changed to have a 3-month analysis frequency.

Table 4a. Mean errors of $\Delta\alpha$ and ETC for Brewer reference instruments (2013-2019) estimated with Model 3.

Brewer serial no.	Mean error of $\Delta\alpha$ [R6 absorption unit]	Mean error of ETC [R6 ETC unit]	Mean error of $\Delta\alpha^*$ [%]	Mean ETC-related error [#] [%]
#008	-0.0002	-1.77	-0.07	-0.08
#014	0.0051	-32.87	1.50	-1.45
#015	-0.0001	-15.64	-0.03	-0.69
#145	0.0007	-8.01	0.21	-0.35
#187	0.0043	-26.84	1.27	-1.19
#191	0.0039	-23.27	1.15	-1.03

* Mean % error in total column ozone, related to error in ozone absorptions; # Mean % error in total column ozone, related to error in ETC, corresponding to X when $\mu = 2$, $\Delta\alpha = 0.34$, and $\Omega = 330$ DU (see Eqn. 3).

Table 4b. Mean errors of $\Delta\alpha$ and ETC for Brewer reference instruments estimated with Model 2.

Brewer serial no. [period]	Mean error of $\Delta\alpha$ [R6 absorption unit]	Mean error of ETC [R6 ETC unit]	Mean error of $\Delta\alpha^*$ [%]	Mean ETC-related error [#] [%]
#008 [1999-2019]	-0.0011	6.79	-0.33	0.30
#014 [1999-2019]	-0.0005	3.26	-0.15	0.14
#015 [1999-2019]	0.0006	-3.79	0.17	-0.17
#145 [2013-2019]	-0.0011	5.68	-0.33	0.25
#187 [2013-2019]	0.0026	-0.61	0.08	-0.03
#191 [2013-2019]	0.0026	-1.05	0.08	-0.05

3.26. Page 19, Figure 5 It is difficult to see anything, probably it would be better to have one plot for every satellite.

Following the suggestion, we made the plot that grouped time series by satellites (see Fig. R9). However, given the number of satellites and ground-based instruments included in this work, it is still very difficult to distinguish the difference between each pair. Also, the focus of this study is assessing the performance of Brewer reference instruments, thus we still prefer the original figure which is ground-based instrument oriented. We think Fig. R9 is also very useful when assessing the performance of each satellite data products. However, we think this is beyond the scope of current study.

The more detailed analysis results were provided in Fig. 6 and Table 5. In addition, the purpose of Fig. 5 is to provide the time series for each instruments and to give a general indication to reader about the Brewers' performance.

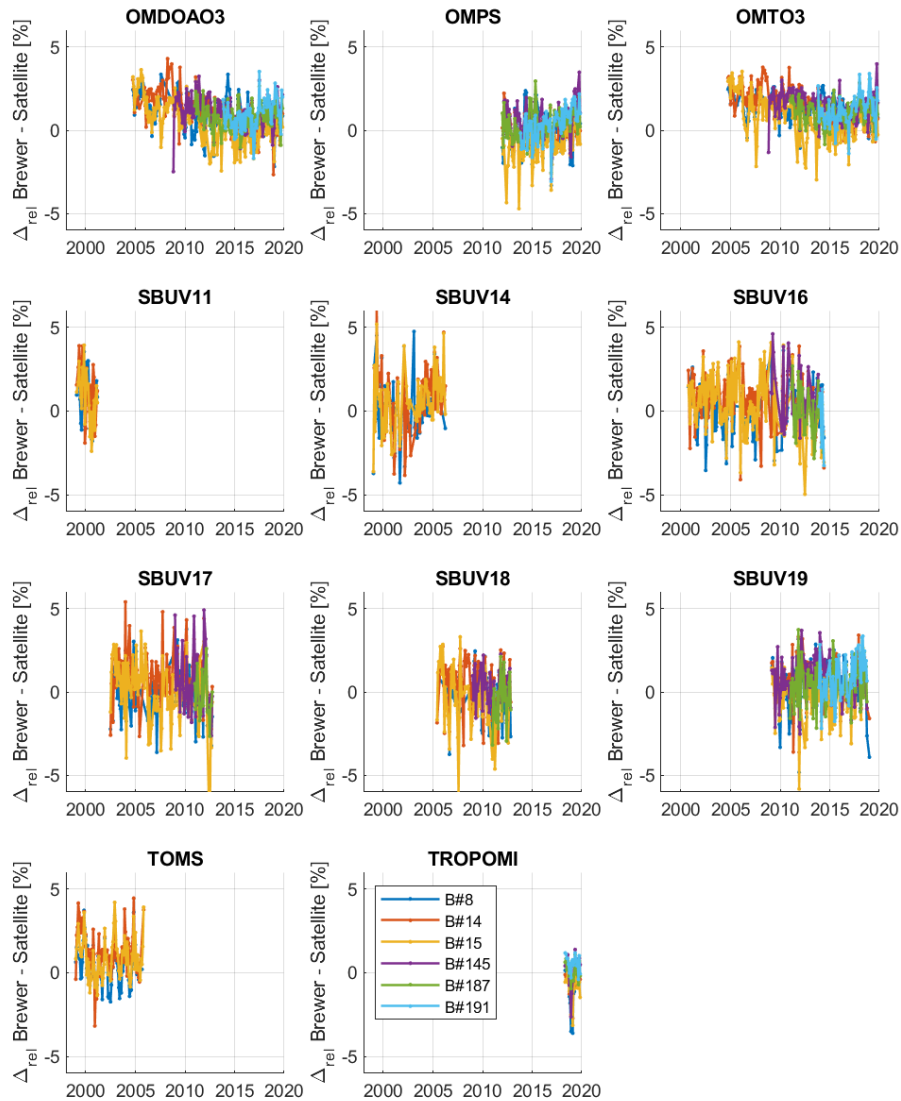


Figure R9. The relative difference between satellites and the world Brewer reference triads (BrT and BrT-D). Each point represents a 3-month average. Brewers and satellite data are paired with the criteria shown in Table 3.

3.27. Page 20, Figure 6 Could you please add a plot with the standard deviation?

The standard deviations were reported in Table 5.

3.28. Page 23, Line 515 There is a correction method to account for the filter non linearity Rimmer et al. [2018] why is not applied?

Currently, the BPS does not have ETC corrections for different filters. The ETC values were generated via a modified Langley method, which fit measurements from various ND filters together. The fitting for the Langley can be done per filter, if data is available. However, this method has other issues such as sampling differences (i.e., less measurements with high ND filters in low SZA conditions). On the other hand, using results from the so-called FI test to correct ND filter non-linearity has been done since 1993 and was described in Savastiouk (2006), which is same as the correction method in Rimmer et al. (2018). As described in Savastiouk (2006), considerable tests were made (i.e., involving twenty instruments) and found the errors caused by this filter non-linearity is within ± 20 DU. Although this number might sounds high enough, the ETC value (which has contributions from many ND filters) will have an overall partial compensation for the effect. In other words, if we apply the FI correction for well-maintained Brewers, we would only expect to see changes in the final data for about ± 1.0 DU, or within ± 0.4 %, for most extreme cases. Besides being responsible for the world reference instruments, the ECCC group maintains the largest number of Brewers within this community (i.e. more than 40 Brewers). Meeting the WMO/GAW requirement and performing centralized data processing with minimal intervention are critical to the ECCC Brewer programme. We fully agree with the referee that further data improvements, such as this ND filter non linearity correction would further improve the quality of the data.

In summary, the goal of this study is to evaluate the overall performance of the current long historical triad record, and any further data improvement can be performed when a higher-precision reprocessing is needed and called upon by WMO/GAW.

3.29. Page 24, Line 523 Please clarify, the instrument is not described before and it is not clear how affect to the ozone measurement and when this issue affects to the results.

The sentence has been modified as suggested.

If the combined focus of the monochromator mirrors of the instrument (see Savastiouk, 2006 for more details of instrument's optical elements) is not optimized and the illuminated filament of the mercury bulb is located in a significantly different location than the illuminated filament from the original bulb, as

much as a 5 micrometer step (one micrometer step is 0.7 pm) change may be seen. For reference, the effective ozone absorption changes by approximately 1% every 3 steps, so a 5 steps shift, which is extreme, can give an error of almost 2% in TCO.

3.30. Page 24, Line 530 See General comment 8

We have answered referee's general comment 8.

3.31. Page 24, Line 545 Please explain how Figure 8b is obtained.

The sentence has been modified to explain this.

As shown by the green and purple lines, the Downsview Brewers were sampling stratospheric ozone over Hamilton, while the Egbert Brewers were sampling stratospheric ozone over west of Brampton (the Brewers' sampling areas were estimated with viewing geometry of Brewers and MERRA-2 ozone profiles, ground projections of the intersections between the Brewer's line-of-sight and the modelled stratospheric ozone layer; Brampton is about 30 km west of Downsview, Hamilton is about 70 km south-west of Downsview).

3.32. Page 24, Line 575 The determination with the model 2 of the ETC and ozone absorption coefficient cannot be define as error budget- The results of the Model 2 are quite far from the goal (the axis limits of Figure 2 are +/- 100 R6 ETC units but the goal is +/-5 R6 units).

The goal of Model 2 is to identify a potential source of errors in total ozone (i.e., errors in ETC or ozone absorption coefficient). As for the goal in ETC values, 5 R6 units of ETC error corresponds to 0.25 % shift in ozone for $\mu = 2$ and ozone = 300 DU. Most of the data in Figure 2 are within ± 20 units of ± 1 % under the same conditions.

Further detailed error ~~budget~~ analysis shows the impacts of ETC and ozone absorption coefficients errors for both reference triads are within ± 2 % when the statistical Model 2 is used.

3.33. Page 26, Line 585 The uncertainty of the Brewer triad is not established on this work, only its long term precision. The highly precise "group scan" is not discussed on this work and shouldn't be in the conclusions.

The sentences have been modified as suggested by the referee.

The *precision* of the Brewer triad instruments are under 0.5 %, while the differences with the best satellite instruments and reanalysis data are close to or slightly lower than 1 %. Further improvement of Brewer total ozone observation precision may be limited by the present Brewer 5-wavelength algorithm and Brewer hardware itself. ~~If highly precise Brewer total ozone measurements are required, then the “group-scan” algorithm (Kerr, 2002) that can deliver measurement uncertainties of individual measurements as low as 0.5 DU or 0.15 to 0.2 %, should be considered instead of the present 5-wavelength method.~~

Response to SC1:

We thank Dr. León-Luis et al. for their very helpful comments and fast response to the issue that we identified and reported in our paper. The results provided in the document are very important and we have adapted some of this information into the revised paper. Our responses are given below in black with the comments from León-Luis et al. in blue. The new/revised text in the modified manuscript is given in red (italicized).

1 Comparison with the RBCC-E triad in León-Luis *et al.* (2018)

In the paper the authors claim that the comparison with the RBCC-E Triad presented in León-Luis *et al.* (2018) should not be carried out because the calculation is not consistent with the results of Model 1 in the present paper by Zhao et al. Note Model 1 was proposed in Fioletov *et al.* (2015).

5 In León-Luis *et al.* (2018), we calculate a quadratic polynomial fit for every Brewer as

$$O_3 = A - i - B \cdot (t - t_0) - i - C \cdot (t - t_0)^2 \quad (1)$$

obtaining for each instrument the corresponding values of A , B and C . Model 1 in Fioletov *et al.* (2015) however calculates common B and C values for all instruments.

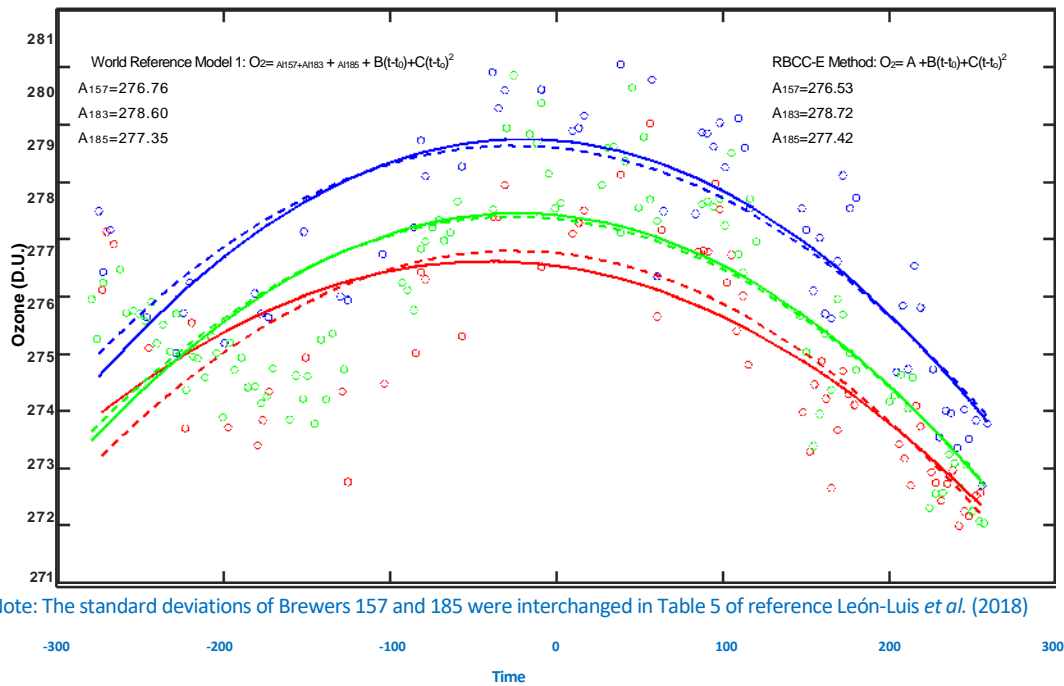
We take the opportunity of the open discussion of this paper to update the calculations of the RBCC-E Triad to be consistent
10 with Fioletov *et al.* (2015), and also to compare the results of both Eq. 1 and Model 1 from Fioletov *et al.* (2015).

Table 1 contains the 3-month standard deviation of the A_i coefficients obtained when the RBCC-E data are re-evaluated using Model 1, together with our previous published results. As can be observed, the values for each Brewer change slightly, depending on the method applied. However, the mean value of the Triad is similar, 0.23% versus 0.27%. This result confirms that there is very little difference between both methods when are applied to the RBCC-E Triad data.

15 This point can be better understood with an example. Fig. 1 demonstrates the total ozone column recorded on November 16th, 2016 (Fig. 4 in León-Luis *et al.* (2018)), where the data have been fitted used the two methods previously described. Table 2 contains the A , B and C coefficients calculated by both methods. As can be seen, regardless of the method used,

Table 1. RBCC-E and World Reference Triads: 3-month standard deviation. We include the values of the World Reference Triad from Zhao *et al.* (2020) for comparison.

RBCC-E			World Reference			
Brewer	$\sigma_{3month},^{Eq.1}$	$\sigma_{3month}, Model 1$	Brewer	$\sigma_{3month}, Model 1$	Brewer	$\sigma_{3month}, Model 1$
#157	0.20 ₁	0.19	#008	0.43 (0.40)	#145	0.44
#183	0.31	0.26	#014	0.36 (0.46)	#187	0.26
#185	0.29	0.23	#015	0.42 (0.39)	#191	0.33
Mean	0.27	0.23		0.40 (0.42)		0.34



Note: The standard deviations of Brewers 157 and 185 were interchanged in Table 5 of reference León-Luis *et al.* (2018)

Figure 1. Ozone values measured on November 16th, 2016, marked with circles. Solid and dotted lines correspond to the 2nd order polynomial fitted using Eq. 1 (RBCC-E method) and Model 1 from Fioletov *et al.* (2015) (World Reference Model method), the Time units are the minutes from solar noon. The A coefficients calculated with both methods are also shown.

the derived A coefficients are very similar. Therefore, the mean daily value of the RBCC-E Triad, the relative errors for each instrument, and the standard deviation, calculated from these coefficients, should not differ significantly. Furthermore, the B and C coefficients calculated by both methods are similar, which suggests that the adjusted functions will exhibit the same behavior as shown in Fig. 1. In conclusion, although both calculation methods are not the same, the results in the case of the RBCC-E Triad are very close. A similar result is achieved when no mathematical adjustment is used and the mean from the simultaneous measurements is calculated directly.

Table 2. Coefficients calculated with the two methods for the RBCC-E Triad data of November 16th, 2016 A, B, and C coefficients

RBCC-E	$A_{157} = 276.53$, $B_{157} = -0.0040$, $C_{157} = -4.855e - 5$ $A_{183} = 278.72$, $B_{183} = -0.0025$, $C_{183} = -6.337e - 5$ $A_{185} = 277.42$, $B_{185} = -0.0028$, $C_{185} = -6.033e - 5$
World Reference	$A_{157} = 276.76$, $A_{183} = 278.60$, $A_{185} = 277.35$ $B = -0.0030$, $C = -5.8122e - 5$

Table 3. Percentage difference of the mean of the three instruments, mean and its standard deviation and the percentage of observations 1% 0.5% and 0.25% of the five minutes simultaneous measurements and daily mean

	Brewer	Mean	σ	<1%	<0.5%	<0.25%
5 min	#157	-0.041	0.342	0.994	0.909	0.687
	#183	0.023	0.372	0.991	0.900	0.701
	#185	0.018	0.342	0.99	0.921	0.758
daily	#157	-0.002	0.245	0.999	0.979	0.816
	#183	-0.005	0.309	0.999	0.931	0.757
	#185	0.007	0.267	0.992	0.954	0.866

Table 3 shows the mean and standard deviation of the ratios for the 5 minutes simultaneous measurements and daily mean 25 values and note that the standard deviations values in this table are fairly similar to those in Table 1, even though the periods used for the calculations are not the same (5 minutes, daily and 3 months).

Up to this point, we have shown that both methods produce a similar result. The difference between the standard deviation reported by both Brewer Triads could then be associated to others factors which have not been considered in these works, such as e.g. the intra-day ozone variability or the number of ozone (Direct Sun) measurements made per day at each station.

30 These factors can affect the robustness of the mathematical fits and, hence, introduce small differences between the calculated A coefficients that are difficult to evaluate for two stations so far apart.

We thanks the Izana Atmospheric Research Centre group from AEMET for providing such detailed recalculations for the RBCC-E data, with the accurate method proposed in Fioletov et al., 2005. We fully agree with the group that using the index matrix method or the simple second-order fitting would not alter the results of the estimated random errors for Brewer triads. As the most accurate and precise total ozone observation instruments, well-maintained Brewers can have less than 0.5% random uncertainties. The difference caused by the selected fitting

algorithm should be well within this uncertainty levels for most of the cases. We pointed this issue out due to two reasons, 1) a part of the results (in León-Luis et al., 2018) was referred to as the use of Model 1 designed in Fioletov et al., 2005, which is not accurate. We thank the AEMET group for providing these re-calculated comparable results. 2) Moreover, the design of Model 1 is just a part of the whole evaluation scheme that been proposed. Model 2 designed in Fioletov et al. 2005 needs a “baseline” ozone, which is calculated from Model 1. For cases illustrated by this comments/report (i.e., RBCC-E data in 2016 Nov. 16), the baseline for this day can be defined as $(A_{157} + A_{183} + A_{185})/3 + B(t-t_0) + C(t-t_0)^2$, if B and C term of the fitting is “common” factors shared by all three instruments. However, when one selects to use a simple second-order fit for each of the instrument, then one will have three B terms and three C terms. One may argue that we also can average B and C terms to receive a “baseline” ozone; however, for some case, the B and C terms can be very different from instruments to instruments (if one only apply the simple fitting). We want to emphasize that the design of Model 1 is only a starting part of the evaluation scheme. The designed models work together to provide a guide in evaluating the performance of Brewers, if no other high-quality reference data can be used as a “referee”.

2 Additional comments

In this section we include some other comments to Zhao et al. (2020), but first we want to acknowledge the effort of the World Reference Triad to maintain all these instruments during decades with such a high precision. Once the precision of the Triads has been established the challenge is to quantify the uncertainty, especially that produced by the described instrumental issues and include them in the analysis.

We thank the AEMET group again for providing us with such important comments and suggestions. The collaborations within the Brewer network is important not just for Brewer researchers, but also for the global ozone monitoring activities and related ozone research studies. The ECCC group designed and maintained the Brewer instruments since the 1970s. Almost a half-century of dedications to ozone monitoring work is a big accomplishment made by all Brewer scientists, technicians, and managers, and more importantly from our collaborators. As noted in the table of world reference instruments’ primary calibration trips, some of the reference instruments were calibrated at Izana with great help from the AEMET group.

1. We do not agree with the comment that the 0.5% level cannot be achieved due to limitations of the Brewer hardware. Some of the issues described, such as for example the filter non linearity, can be addressed, and indeed are accounted for

in the processing performed at Eubrewnet. Eubrewnet's processing also takes into account the issue described for Brewer #15 – the observations not compensated with mercury tests are automatically filtered out.

The purpose of this work is to evaluate the triad performance based on the existing calibration results. Numerous works have been done by researchers worldwide in past decades to improve Brewers' accuracy and precision. However, not all of them have been implemented to the current reference instruments' results. We also provided a more detailed reply on issues such as filter non-linearity correction to referee #2. ECCC group maintains the largest number of Brewers within this community (i.e. more than 40 Brewers). Meet the WMO/GAW requirement and performing centralized data processing with minimal intervention are critical to the ECCC Brewer programme. In summary, the goal of this study is to evaluate the overall performance of the current long historical triad record, and any further data improvement can be performed when a higher-precision reprocessing is needed and called upon by WMO/GAW. ECCC group welcomes and is looking forward to continue and further our collaborations with AEMET group on these activities in future.

2. The cited Pandora manual has more than 150 pages, so it is difficult to find the ozone processing details. It could be better to refer to the ozone processing in the user guidelines available at https://www.pandonia-global-network.org/wp-content/uploads/2020/01/LuftBlick_FRM4AQ_PGNUserGuidelines_RP_2019009_v1.pdf. Furthermore, if we understand it correctly, the data used in the present paper by Zhao et al. is not the operational one that is available to the public for download.

We have included this in the reference. The effective temperature corrected Pandora ozone is not available on the PGN website. We have upload the corrected TCO data to ECCC's public data server and can be downloaded from: https://collaboration.cmc.ec.gc.ca/cmc/argi/Zhao_et_al_amt-2020-324/

Additional information on Pandora calibrations, operation, retrieval algorithms and correction method can be found in Cede (2019; Cede et al., 2019), Tzortziou et al., (2012), and Zhao et al., (2016).

Cede, A., Tiefengraber, M., Gebetsberger, M. and Kreuter, M.: TN on PGN products "correct use" guidelines, Pandonia Global Network. [online] Available from: https://www.pandonia-global-network.org/wp-content/uploads/2020/01/LuftBlick_FRM4AQ_PGNUserGuidelines_RP_2019009_v1.pdf (Accessed 13 November 2020), 2019.

3. It looks that there is a trend on the Merra comparison from 2005 to 2015, with Brewer #015 going from +2% to -2%

We are not sure Brewer #015 is the only instrument that shows a “trend”. For example, Brewers #008 and #014 look like they also have a decreasing trend in this period, just not as “strong” as the one observed by Brewer #015. It is difficult for us to determine if this trend is real. To continue this investigation, one will need to collaborate with the reanalysis group and investigate not just Brewers, but also any upgrade or modifications in the model and its source of assimilation. However, we think that this interesting topic could be a standalone research topic and is beyond the scope of the current study.

4. Appendix A. The standard deviation of the ozone measurement is strongly affected by clouds and is also used as cloud mask to filter the AOD measurements affected by rapid moving clouds (López-Solano et al., 2017). Some of the brewer are equipped with full sky cameras, are the observations reported in Zhao et al. (2020) also filtered by clouds?

The standard data screening procedure used the standard deviations of 5 individual Brewer measurements to remove DS values obtained under cloudy conditions or under moving clouds. In very rare occasions when the standard deviations of such measurements are within the established limits, but the Sun is obscured by the clouds, an additional screening is done using the absolute intensity at the longest wavelength (320nm). Sky cameras are not very useful for such screening since the sky cloud coverage does not really affects DS measurements unless the clouds block the Sun. Sky camera data was not used in this study.

5. Appendix A. Figure A2 shows the dependence with the ozone air mass factor (AMF), as the stray-light is a function of AMF (Karppinen et al. (2015)) , but in the text the discussion is focused on the solar zenith angle and air mass

Further discussions of the stray light issue were provided in our answers to referees #1 and #2. Please refer to our answers in those replies.

6. Appendix A. An statistical approach to estimate the single triad stray light Diemoz et al. (2015) or the determination of the empirical correction by comparison with the double one (Redondas et al., 2018) could be performed to the dataset.

We thank the AEMET group for this very useful suggestion. When re-processing of the Brewer reference instruments records are needed, we will try to implement these proposed methods.

The world Brewer reference triad – updated performance assessment and new double triad

Xiaoyi Zhao¹, Vitali Fioletov¹, Michael Brohart¹, Volodya Savastiouk², Ihab Abboud¹, Akira Ogyu¹,
Jonathan Davies¹, Reno Sit¹, Sum Chi Lee¹, Alexander Cede^{3,4}, Martin Tiefengraber^{4,5}, Moritz Müller^{4,5},
5 Debora Griffin¹, Chris McLinden¹

¹Air Quality Research Division, Environment and Climate Change Canada, Toronto, M3H 5T4, Canada.

²International Ozone Services Inc., Toronto, Canada.

³NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA.

⁴LuftBlick, Innsbruck, Austria.

10 ⁵Department of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria.

Correspondence to: Xiaoyi Zhao (xiaoyi.zhao@canada.ca)

Abstract. The Brewer ozone spectrophotometer (the Brewer) was designed at Environment and Climate Change Canada (ECCC) in the 1970s to make accurate automated total ozone column measurements. Since the 1980s, the Brewer has become
15 a World Meteorological Organization (WMO) Global Atmosphere Watch (GAW) standard ozone monitoring instrument. Now, more than 230 Brewers have been produced. To assure the quality of the Brewer measurements, a calibration chain is maintained, i.e., first, the reference instruments are independently absolutely calibrated, and then the calibration is transferred from the reference instrument to the travelling standard, and subsequently from the travelling standard to field instruments. ECCC has maintained the world Brewer reference instruments since the 1980s to provide transferable calibration to field
20 instruments at monitoring sites. Three single-monochromator (Mark II) type instruments (serial numbers #008, #014, and #015) formed this world Brewer reference triad (BrT), and started their service in Toronto, Canada in 1984. In the 1990s, the Mark III type Brewer (known as the double Brewer) was developed, which has two monochromators to reduce the internal instrumental stray light. The double Brewer world reference triad (BrT-D) was formed in 2013 (serial numbers #145, #187 and #191), co-located with the BrT. The first assessment of the BrT's performance was made in 2005, covering the period
25 between 1984 and 2004 (Fioletov et al., 2005). The current work provides an updated assessment of the BrT's performance (from 1999 to 2019) and the first comprehensive assessment of the BrT-D. The random uncertainties of individual reference instruments are within the WMO/GAW requirement of 1 % (WMO, 2001; 0.49 % and 0.42 % for BrT and BrT-D, respectively as estimated in this study). The long-term stability of the reference instruments is also evaluated in terms of uncertainties of the key instrument characteristics: the extraterrestrial calibration constant (ETC) and effective ozone absorption coefficients
30 (both having an effect of less than 2 % on total column ozone). Measurements from a ground-based instrument (Pandora

spectrometer) and satellites (eleven datasets, including the most recent high-resolution satellite, TROPOspheric Monitoring Instrument), and reanalysis model (the second Modern-Era Retrospective analysis for Research and Applications, MERRA-2) are used to further assess the performance of world Brewer reference instruments and to provide a context for the requirements of stratospheric ozone observations during the last two decades.

35 1 Introduction

Ozone (O_3) is one of the most well-known and critical atmospheric trace gases (WMO, 2018), with remote sensing monitoring of atmospheric ozone being traced back to 1926 (Dobson, 1968). In the late 1970s to early 1990s, stratospheric ozone has become an important scientific topic and a matter of intense interest after discovery and subsequent studies of the Antarctic ozone hole (Farman et al., 1985; Solomon et al., 1986; Stolarski et al., 1986) and ozone depletion on the global scale (Ramaswamy et al., 1992; Stolarski et al., 1991). To perform long-term, automated, ground-based total column ozone monitoring, the Brewer instrument was proposed by Alan Brewer (Brewer, 1973) and developed with James Kerr, Tom McElroy and David Wardle in the early 1980s at Environment and Climate Change Canada (ECCC) (Kerr, 2010; Kerr et al., 1981). In 1988, the Brewer was designated as the World Meteorological Organization (WMO) Global Atmosphere Watch (GAW) standard instrument for total column ozone measurements. ECCC has maintained the world Brewer reference instruments since the 1980s to provide transferable calibration to field instruments at monitoring sites. In practice, three Mark II type instruments (serial numbers #008, #014, and #015) formed this world Brewer reference triad (BrT), and started their service in Toronto (43.781° N, 79.468° W, 187 m a.s.l.), Canada in 1984 (Fioletov et al., 2005). The long-term performance of these three instruments was previously evaluated using direct sun total column measurements for a 20-year period between 1984 and 2004 (Fioletov et al., 2005). Data analysis from this study shows that the precision of individual observations are within ± 1 % in about 90 % of all measurements.

Internal instrumental stray light affects measurements made with the single-monochromator instruments; therefore, corrections are applied to the data when necessary (Bais et al., 1996; Fioletov et al., 2000; Karppinen et al., 2015; Rimmer et al., 2018). To significantly reduce this effect, in 1992, ECCC scientists introduced the Brewer Mark III spectrophotometer that uses the same concept of the Mark II model version, but has a second monochromator (Wardle et al., 1996). In 2013, a second world reference standard, known as the double Brewer reference triad (BrT-D), consisting of three Brewer double spectrophotometers (serial numbers #145, #187 and #191) was co-located with the original triad in Toronto (Zhao et al., 2016). The two triads run in parallel. These two triads serve as a calibration reference for travelling standard instruments that are used for calibration of Brewer spectrophotometers deployed across the world in the GAW Programme run under the auspices of the WMO. There are other Brewer triads formed and operated by the Swiss Federal Office of Meteorology and Climatology (Meteo Swiss; the triad is known as the Arosa triad) and the State Meteorological Agency of Spain (AEMET; the triad is known as the Regional

65 Brewer Calibration Center Europe (RBCC-E) triad). The Arosa triad (Staehelin et al., 1998; Stübi et al., 2017b), formed in 1998, was the second Brewer triad worldwide (composed of two Mark II and one Mark III instruments; now in Davos at PMOD World Radiation Center (Stübi et al., 2017a)). To better coordinate the Brewer network at the regional scale (León-Luis et al., 2018; Redondas et al., 2018), the RBCC-E triad was formed in 2003 (composed of three Mark III instruments). The regional reference instruments are regularly compared to the world reference instruments via a travelling standard (Redondas et al., 2018).

70 By 2019, there were more than 230 Brewer instruments manufactured, with most of them deployed worldwide within the WMO GAW global ozone monitoring network. From 1999 to 2019 (the period within which the world Brewer reference instruments' data are evaluated in this work), the World Ozone and Ultraviolet Radiation Data Centre (WOUDC, woudc.org) received Brewer ozone observations from 123 instruments at 88 stations. As a large global monitoring network, the measurement stability is maintained via strict laboratory calibrations (e.g., ozone absorption coefficients from dispersion test) and field calibration (i.e., deriving the extraterrestrial calibration constant). For example, the effective ozone absorption
75 coefficients ($\Delta\alpha$) are determined for each individual instrument in laboratories via dispersion test, and are regularly checked using the stable solar spectrum as the reference using the so-called Sun Scan test (Savastiouk, 2006). The extraterrestrial calibration constant (ETC) has to be determined in the field by one of the two means: 1) the independent calibration method, i.e., the Langley plot calibration method or the so-called zero airmass extrapolation technique, or 2) the calibration transfer method (e.g., transfer ETC from well-calibrated reference instruments to field instruments) (see more details about calibration
80 procedures in Kerr, 2010). In practice, each field Brewer instrument receives its ETC constant by comparing ozone values with those of the travelling standard instrument. The travelling standard itself is calibrated against the set of world reference instruments (i.e., world Brewer reference triad). The world reference triad data are used to calibrate the traveling standard, and the traveling is used to calibrate 30-40 Brewers per year, on average, around the world. Each individual reference instrument is independently calibrated at the Mauna Loa Observatory (MLO), Hawaii (19.5° N, 155.6° W, 3400 m asl), every 3-8 years
85 (see Table 1) via the Langley plot calibration method. Thus, it is critical to review and assess the world reference instruments' performance on a regular basis. This study's focus is on the demonstration of the long-term stability of the existing reference instrument. Absolute calibration procedure, maintenance, calibration transfer, and assessment of travelling standard will be a subject of a separate study.

90 Previously, the assessment for the BrT, carried out by Fioletov et al. (2005), examined its twenty-year long record of direct sun (DS) total ozone measurements (1984-2004). It was found that the BrT's precision over these two decades was better than ± 1 % (Fioletov et al., 2005). There is no further published assessment for the world reference instruments after that period, and no formal assessment made for the BrT-D yet. In addition, with the increasing number of satellite observations (e.g., OMI,

95 TROPOMI) and ground-based observations from emerging technologies (e.g., Pandora spectrometer) of total ozone columns, it is important to compare the triad datasets with these measurements.

100 This paper provides a more recent assessment for the BrT (1999-2019) and reports the first assessment of the BrT-D (2013-2019). It is organized as follows: Section 2 describes the ground-based ozone measurements, satellite ozone measurements, and the model reanalysis ozone data. In Section 3, the standard and the new evaluation schemes are introduced, with a detailed description of a new third-party evaluation model. In Section 4, the world Brewer reference instruments (BrT and BrT-D) data products are evaluated by the standard and new schemes. Lastly, Section 5, discusses the challenges for Brewer instruments to measure ozone at a level better than 1%, in the context of the comparison between the world reference triads, regional triads and high-resolution satellite data. Conclusions are given in Section 6.

2 Datasets

105 2.1 Brewer

110 There are several model versions of the Brewer instrument. The Mark I prototype instruments were tested and operated since the 1970s (Kerr et al., 1981). The first production version (Mark II) was introduced in the early 1980s. In the 1990s, the double monochromator (Mark III) was developed to reduce the internal instrumental stray light, which allows high-quality total column ozone measurements in large slant column ozone (e.g., low sun elevation) conditions. There were other versions of Brewers developed in the late 1990s (i.e., Mark IV and V) to extend the measuring wavelengths and to measure other trace gases. Today, only the Mark III version of the Brewer is manufactured. Table 1 summarizes some of the specific similarities and differences between the single and the double Brewer reference triads. More details about Mark II and III's measurement standard deviations and stray light characteristics are provided in Appendix A.

Table 1. Specific features of single and double Brewer reference triads

	Single Brewer	Double Brewer
Model Version	Mark II	Mark III
Serial No.(s)	#008, #014 and #015	#145, #187 and #191
Start of triad observations	September 1984	October 2013
Optical and spectral characteristics	Single monochromator: a dispersing monochromator with an 1,800 line/mm holographic diffraction grating.	Double monochromator: a top dispersing monochromator with a 3,600 line/mm holographic grating, and a bottom recombining monochromator that is a mirror image of the dispersing monochromator
	Spectra measured by a single monochromator that is affected by the internal instrumental stray light in the UV region (Bais et al., 1996; Fioletov et al., 2000).	Significantly less instrumental stray light (out-of-band, stray light fraction 10^{-7}) than in the single monochromators (10^{-5}) (Fioletov et al., 2000). Thus, increased accuracy of ozone and UV measurements under certain conditions (Bais et al., 1996; Wardle et al., 1996).
Output	Solar radiation at six UV wavelengths is measured with the spectrometer. The wavelengths are 303.2 nm (almost exclusively for wavelength calibration, i.e., spectral reference test) and five operating wavelengths (306.3 nm, 310.1 nm, 313.5 nm, 316.8 nm and 320.1 nm) used to measure total column ozone and sulphur dioxide using the sun, sky or near full moon as a light source.	
	Provides high-quality ozone measurements with a slant ozone column amount up to 1000 DU, which for the global average total ozone column of 300 DU corresponds to an ozone air mass factor of 3.33 and a solar zenith angle (SZA) of about 73° (Zanjani et al., 2019).	Provides high-quality ozone measurements with a slant ozone column amount up to 2000 DU, which for the global average total ozone column of 300 DU corresponds to an ozone air mass factor of 6.67 and a SZA of about 81° (Savastiouk, 2006).

115

The Brewer spectrophotometer provides data products that include column ozone (e.g., Kerr, 2002; Kerr et al., 1981), column sulphur dioxide (SO₂; e.g., Fioletov et al., 1998; Zerefos et al., 2017), column nitrogen dioxide (NO₂, by Mark IV only; e.g., Cede et al., 2006; Kerr et al., 1988), spectral UV radiation (e.g., Bais et al., 1996; Fioletov et al., 2002), aerosol optical depth (e.g., Kazadzis et al., 2005; Marengo et al., 2002), and effective ozone layer temperature (Kerr, 2002). However, the main data product provided by the Brewer is the total column ozone via direct-sun observations. In this work, we focus on the Brewer direct-sun total column ozone data product only, although total column ozone also can be retrieved using solar zenith-sky radiance, solar global spectral UV irradiance, and lunar direct irradiance (Fioletov et al., 2011; Kerr, 2010). Brewer data was processed by Brewer Processing Software (BPS) developed by ECCO (Fioletov and Ogyu, 2008). **The same processing software was used in Fioletov et al. (2005)**. The software demonstrated good performance in a recent comparison of available processing software tools for Brewer total ozone retrievals (Siani et al., 2018).

120

125

The Brewer spectrophotometer is a modified Ebert grating spectrometer that was designed to measure almost simultaneously the intensity of radiation at six selected channels in the UV (nominally 303.2, 306.3, 310.1, 313.5, 316.8, and 320.1 nm). The first channel is almost exclusively used for wavelength calibration. The four longer wavelengths are used for the total column ozone (Ω) retrieval via the following equation:

$$F + \Delta\beta \cdot m = F_0 - \Delta\alpha \cdot \Omega \cdot \mu \quad (1)$$

where, m and μ are the enhancement factors for the slant pathlength of the direct radiation relative to vertical path for air and the ozone layer respectively (also known as the air mass factors). F , $\Delta\alpha$, and $\Delta\beta$ are the linear combinations of the logarithms of the measured intensity (base 10), the effective ozone absorption and the Rayleigh scattering coefficients, respectively. For example, $F = \log(I_3) - 0.5\log(I_4) - 2.2\log(I_5) + 1.7\log(I_6)$, where I_3 to I_6 are the photon count rates at channel number three to number six. F_0 is the instrument response (F) if there were no atmosphere between the instrument and the sun, it is also known as the ETC. Details about the standard Brewer ozone retrieval algorithm can be found in Kerr (2010) and the references cited there. In the standard Brewer algorithm, $\Delta\beta$, m , and μ are determined and pre-calculated, and are not instrument-dependent. F_0 and $\Delta\alpha$ (calibration constants) are unique for each instrument and depend on the exact wavelengths and band passes of the slits of each instrument (Kerr et al., 1985). After laboratory and field calibration (to determine $\Delta\alpha$ and F_0 , respectively), Ω is then readily calculated for each field observation (i.e., F).

As previously described, to maintain high quality data of all Brewer instruments (i.e., transfer the F_0 value), the world reference instruments (BrT and BrT-D) receive their F_0 values via the independent calibration technique. In short, these high-precision F_0 values were determined by fitting the measured F values as a linear function of air mass factor (see Eqn. 1). For example, in clear sky conditions with stable ozone values, if measurements are made under a range of air mass factors throughout a day, then the intercept of the linear fitting of $(F + \Delta\beta m)$ versus μ will be F_0 . More technical details, such as calibration periods, averaging, and why MLO is the ideal site for this practice are provided in details in Kerr 2010. These values are transferred to the travelling standard and then to the field Brewers via co-located field calibration routines (i.e., calibration transfer method). The primary calibration history of the world Brewer reference instruments is summarized in Table 2. Due to building roof work at the Toronto site, the BrT-D was temporarily moved to Egbert, Canada (44.230° N, -79.780° W) at the beginning of September 2018 and deployed on the roof of the ECCC Centre for Atmospheric Research Experiments building (CARE, 251 m a.s.l.). The CARE building is located in a rural area, which is surrounded by farmlands. For this period between September 2018 and December 2019, the BrT-D was located about 55 km north west from the BrT. This period of data is still used in the analysis to study and illustrate some fine scale variations in the ozone field. More details about reference instruments' repair and upgrade review are provided in the supplementary information.

Table 2. Independent calibration history of world Brewer reference instruments.

Serial no. (Model version)	Operation since	Independent calibration
No. 008 (Mark II)	1984	March 1999 MLO*
		July 2005 MLO
		Instrument failure in July 2007
		Oct 2008 Izaña
		Oct 2015 MLO
No. 014 (Mark II)	1984	Apr 2000 MLO
		July 2005 MLO
		Nov 2008 MLO
		Oct 2013 MLO
No. 015 (Mark II)	1984	Apr 2002 MLO
		Nov 2010 MLO
		Oct 2013 MLO
		Nov 2017 MLO
No. 145 (Mark III)	1998	Oct 2008 Izaña
		Oct 2015 MLO
		Oct 2019 MLO
No. 187 (Mark III)	2007	Nov 2010 MLO
		Oct 2015 MLO
No. 191 (Mark III)	2009	Oct 2013 MLO
		Nov 2017 MLO

*MLO: Mauna Loa Observatory.

2.2 Pandora

The Pandora instrument records spectra between 280 and 530 nm with a resolution of 0.6 nm (Herman et al., 2009, 2015; Tzortziou et al., 2012). It uses a temperature-stabilized Czerny-Turner spectrometer with a 2048 × 64 pixels CCD detector. The spectra are analyzed using total optical absorption spectroscopy (TOAS) technique (Cede, 2019), in which absorption cross-sections for multiple atmospheric absorbers such as ozone, NO₂, and SO₂, are fitted to the spectra. Different from the Brewer, which only uses intensities measured at four wavelengths, the Pandora instruments use the entire spectrum from 310 to 330 nm (at 0.6 nm resolution, with more than 160 pixels) in its ozone retrieval. The current Pandora standard ozone column retrieval algorithm uses a literature reference spectrum (composite of Kurucz (2005), Thuillier et al. (2004), van Hoosier (1996) and Gueymard (2004), details in Cede (2019)), and does not retrieve the effective ozone temperature. Thus, Pandora standard ozone data products have a temperature dependence (Herman et al., 2015), i.e., 0.25 % K⁻¹ when compared to Brewer measurements (Zhao et al., 2016). This temperature dependence introduces a 1 to 3 % seasonal bias between the Pandora and the Brewer standard data products. Another major difference between the Brewer and Pandora retrieval algorithms is their

175 selection of ozone cross-section, i.e., the Brewer uses BP (Bass-Paur) ozone cross-section (at 228.3° K, Bass and Paur, 1985)
and the Pandora uses Serdyuchenko ozone cross-section (at 225° K, Serdyuchenko et al., 2014). As a result of temperature
dependency and different selection of ozone cross-sections, a two percentage multiplicative bias between the Pandora and
Brewer standard ozone column products were found in Zhao et al. (2016). Thus, in this work, the Pandora ozone data are
corrected by an empirical method with the ozone-weighted effective temperature (Zhao et al., 2016). The effective temperature
180 was calculated from temperature and ozone profiles provided by ERA-Interim (Dee et al., 2011). **In general, after correction,
the multiplicative bias in Pandora ozone data can be decreased from 2.92 to -0.04 %, with the seasonal difference (estimated
with monthly data) decreased from ± 1.02 to ± 0.25 % (see Fig. 11 in Zhao et al., 2016; i.e., comparing to Brewer, corrected
Pandora data has $-0.04 + 0.25\%$ offset in summer and $-0.04 - 0.25\%$ offset in winter).** An effective ozone temperature retrieval
algorithm is under development for the Pandora to minimize its temperature dependence effect (Cede, 2019). Additional
185 information on Pandora calibrations, operation, retrieval algorithms and correction method can be found in Cede (2019; **Cede
et al., 2019**), Tzortziou et al., (2012), and Zhao et al., (2016).

Pandora instrument no. 103 has been making direct-sun measurements in Toronto (co-located with BrT and BrT-D) since 2013
(Zhao et al., 2016). The instrument has made almost daily measurements since its deployment, except during a filter upgrade
190 in 2017. The seven years of data (2013-2019) have been re-processed and harmonized by the Pandonia Global Network (PGN)
to ensure the high quality of its ozone data product. In this work, only high-quality Pandora ozone data products are used
(Pandora level 2 (L2) data product quality flag = 0; Cede, 2019). Originally, Pandora no. 103 was operated in DS mode only
and Pandora DS ozone data had a one-minute resolution. Starting in 2018, it was operated in the combination mode (i.e., direct-
sun, zenith-sky, and multi-axis) and Pandora DS ozone data had a five-minute resolution. The Pandora and BrT-D instruments
195 have good stray-light control, **and under typical ozone conditions (i.e., slant column ozone less than 1500 DU),** and their air
mass dependence is comparably low up to 81.6° SZA (**within 1% up to AMF = 5.5;** Zhao et al., 2016). **Benefitting from the
TOAS technique, unlike Brewers, Pandora instruments do not need the independent calibration at MLO (Tzortziou et al.,
2012).**

2.3 Satellites

200 The BrT's performance was evaluated against the Total Ozone Mapping Spectrometer (TOMS) and reported in Fioletov et al.,
(2005). With more satellite instruments reporting total ozone columns, here we present a data comparison between the Brewer
reference instruments (BrT and BrT-D) and multiple satellites, including TOMS, NOAA Solar Backscatter Ultraviolet
Radiometer-2 (SBUV) series (nos. 11, 14, 16, 17, 18, 19), Ozone Mapping and Profiler Suite (OMPS), Ozone Monitoring
Instrument (OMI), and TROPospheric Monitoring Instrument (TROPOMI).

205 2.2.1 TOMS

There were four TOMS in orbit: on Nimbus-7 satellite launched in 1978, on Meteor-3 in 1991, and on ADEOS and Earth Probe (EP) in 1996. Total column ozone was derived from incident solar radiation and backscattered ultraviolet sunlight measurements. TOMS total column ozone has been widely used for verification of ground-based measurements (e.g., Fioletov et al., 1999; Kyrö, 1993). Fioletov et al. (1999) reported that about 80 % of the Dobson and Brewer data have standard deviations of monthly mean difference with TOMS that are less than 2.5 %. The EP/TOMS total ozone data from 1996 to 2005 with a quality flag of zero were used in this work (McPeters et al., 1998).

2.2.2 SBUV series

Total column ozone from NOAA SBUV series (nos. 11, 14, 16, 17, 18, 19) is used in this work. Unlike TOMS, OMI or TROPOMI, which provides daily global coverage, the non-scanning, nadir viewing SBUV instruments provide full global coverage approximately bi-weekly. The SBUV ozone column data used in this work is produced and quality assured by the overpass algorithm to create daily overpass values (Labow et al., 2013; by weighted-interpolating data measured within the box centred on the station location ($\pm 2^\circ$ in latitude and $\pm 20^\circ$ degrees in longitude)). Labow et al. (2013) reported that the total column ozone data from Brewers and SBUVs show an agreement within ± 1 % over 40 years (1970-2010; yearly relative difference).

220 2.2.3 OMPS Nadir Mapper

The OMPS on the Suomi National Polar-orbiting Partnership (Suomi NPP) satellite was launched in 2011 (Flynn et al., 2014; Kramarova et al., 2014). OMPS includes nadir and limb modules to measure both profile and total column ozone concentrations. In this work, OMPS-NPP L2 Nadir Mapper (NM) Ozone Total Column swath orbital v2.1 data (only good sample, with a QualityFlags of zero) from the OMPS-NM module is used. Flynn et al. (2014) reported that the OMPS column ozone (from an earlier v1) has a bias with other records (e.g., OMT03) on the order of -3 %.

2.2.4 OMI

The OMI instrument on the Earth Observing System Aura satellite was launched in 2004. OMI has two standard data products, OMDOAO3 (J. P. Veefkind et al., 2006) and OMT03 (Bhartia and Wellemeyer, 2002), which are produced using DOAS and TOMS-like techniques, respectively. The mean difference between the two data products varies from 0 to 9 DU (0-3 %) with latitude and season (Kroon et al., 2008). In this work, the OMDOAO3 and OMT03 OVP data are used, with L2 quality flag equal to 0 or 1 and bit 6 is not set are included (see <https://avdc.gsfc.nasa.gov/pub/data/satellite/Aura/OMI/V03/L2OVP/OMDOAO3/>).

2.2.5 TROPOMI

235 TROPOMI, onboard the Copernicus Sentinel-5 Precursor satellite, was launched in 2017. The offline (OFFL v010107) total
ozone column data (Garane et al., 2019) are used in this work (only L2 data with $qa \geq 0.75$ are included). Garane et al., (2019)
reported that the mean bias and the mean standard deviation of the percentage difference between TROPOMI and Brewer
ground-based total ozone column data are within 1 % and 2.5 %, respectively.

2.3 MERRA-2 reanalysis data

240 The second Modern-Era Retrospective analysis for Research and Applications (MERRA-2) is an atmospheric reanalysis from
NASA's Global Modeling and Assimilation Office (GMAO). MERRA-2 assimilates partial total column ozone retrievals from
the SBUV series from 1980 to 2004. From October 2004, MERRA-2 assimilates ozone profiles and total column data from
the Microwave Limb Sounder (MLS) and the OMI, respectively (Wargan et al., 2017). MERRA-2 column ozone data has been
found to be of good quality when compared with satellite and ground-based observations (e.g., Rienecker et al., 2011; Wargan
et al., 2017; Zhao et al., 2017, 2019). In this work, the MERRA-2 total column ozone ($0.5^\circ \times 0.625^\circ$, version 5.12.4) with 1-
245 hour temporal resolution is used as an input in the third-party comparison model (see Section 4 for more details).

3 Comparison methods

Multiple Brewer instruments at the same site may not measure ozone at exactly the same time. To compare the ozone column
data provided by each Brewer reference instrument, a “baseline” ozone column value at the time of each measurement should
be established. Ideally, if the true ozone column values are known, then the performance of each instrument can be evaluated
250 as simple as calculating the discrepancies between true ozone and measured ozone. However, this approach is not possible in
reality. Several other means to form these (daily or time-resolved) baseline ozone values were used in the past: 1) the average
of all satisfactory measurements for each instrument (Kerr et al., 1998), 2) a second-order time-resolved statistical model
(Fioletov et al., 2005), 3) a third-order simple polynomial fit (Stübi et al., 2017b), and 4) a fourth-order time-resolved statistical
model (León-Luis et al., 2018). In general, these approaches aim to define the best baseline total column ozone values for each
255 day, which are as close to true ozone values as possible. Apparently, the first method (i.e., simple daily mean) is not ideal since
it includes the effects of ozone changes during the day combined with differences in the timing and number of measurements
by each instrument (Fioletov et al., 2005) and instrument uncertainties are overestimated. The second method takes the daily
baseline ozone values as a second-order function, which are fitted using all satisfactory measurements for all three instruments
together, but also give the individual instrument a degree of freedom in offsets. The third method takes the ozone changes into
260 account, but it is still affected by the number of measurements from each instrument (i.e., the instrument reporting more data
points will dominate the baseline). The advantage of the time-resolved model (second or fourth method) is that it takes both

effects of ozone changes into account and minimizing the impact of sampling (i.e., all three instruments share the same first and second-order terms, while the offset terms are unique for each instrument; see more details in the following section). It should also be noted that third- or higher-degree polynomial fit does not really change the results much because the baseline is only needed to adjust for the time difference in ozone measurements by individual Brewers. Thus, to make the current work directly comparable to previously reported results for the world reference instruments, we only use the second approach in the analysis (i.e., second-order time-resolved statistical model; following Fioletov et al., (2015) referred to as Model 1).

In addition to constructing the baseline with the individual Brewers' data, we can use third-party (e.g., co-located, independent total column ozone measurements from Pandora) measurements as the baseline ozone in the evaluation. The Pandora instrument typically has a better temporal-resolution than Brewers, and therefore, can capture most of the daily ozone variations better. Moreover, when using coincident Pandora ozone data, the baseline will not have the sampling or weighting issues; i.e., the Brewer instrument that reported more data points will not dominate the forming of the baseline (i.e., as the baseline formation in Model 1, see Eqn. 2). However, when using this third-party baseline, we should be cautious about the difference between Pandora and Brewer ozone data products, i.e., their seasonal and multiplicative bias. Details about how to interpret the third-party assessment results are provided in Section 4.

3.1 Comparison with ground-based instruments

3.1.1 The original method

Two statistical models have been developed to evaluate Brewer reference instruments' performance by Fioletov et al., (2005). The first model is a time-resolved second-order model (referred to as Model 1) to provide the baseline ozone and applied to the reference triad data from each day:

$$\Omega = A_1 \cdot I_1 + A_2 \cdot I_2 + A_3 \cdot I_3 + B \cdot (t - t_0) + C \cdot (t - t_0)^2 \quad (2)$$

where, Ω is an ozone measurement from one of the three Brewers (e.g., BrT, or here with arbitrary serial nos. 1, 2, and 3), t is the corresponding time of the measurement and t_0 is the local solar noon time. The I_1 , I_2 , and I_3 are the indicator functions for each of the three Brewers. For example, if the ozone value Ω is measured by Brewer no. 1, I_1 is set to 1 (and set to 0 for the two other Brewers). The coefficients A_1 , A_2 , A_3 , B , and C can then be estimated by the least-squares method. Please note here the ozone values for this day are then represented by three second-order curves, which share the common curvatures (B and C terms), but have a different offset (i.e., A_1 , A_2 , and A_3). In other words, each instrument formed its own daily time-resolved ozone variations, but these variations are not totally independent from each other (since they share the B and C terms). Then, the average of the three coefficients $A = (A_1 + A_2 + A_3)/3$ is used as the benchmark to evaluate the performance of individual

instruments. For example, $(A_I - A)$ represents the deviation of Brewer no. 1 from the baseline ozone (i.e., corresponding to $\Omega = A + B(t-t_0) + C(t-t_0)^2$).

295 In general, with contributions from all three instruments, this model removes the diurnal ozone variations relative to the noon ozone value. Meanwhile, the model preserves the instrumental differences as much as possible by assigning different offsets for each baseline (i.e., corresponding to an assumption that there is only an additive bias between Brewers).

For a well-calibrated and well-maintained Brewer instrument, its major uncertainties in derived ozone column data came from two instrument constants assigned to it (i.e., F_0 and $\Delta\alpha$). Next, to further break down the uncertainty budgets, Model 2 is designed by combing Eqns. 1 and 2 as:

$$F + \Delta\beta \cdot m = (F'_0 + X) - (\Delta\alpha' + Y) \cdot (A + B \cdot (t - t_0) + C \cdot (t - t_0)^2) \cdot \mu \quad (3)$$

305 where, F'_0 and $\Delta\alpha'$ are the assigned ETC and effective ozone absorption coefficient values. X and Y are the assigned uncertainties to these two instrument constants. Here, the total column ozone amount (Ω) is replaced by the Model 1 defined baseline ozone. Next, X and Y can be estimated for each of the three instruments using the least-squares method for each 3-month season. In general, Model 2 assumes that the baseline ozone provided by Model 1 is the ground-truth (i.e., true ozone values). Thus, the difference of total column ozone between the individual instrument and Model 1 is allocated to the “error” of ETC and effective ozone absorption values. **As the stray light issue in high μ conditions may affect the formation of the baseline ozone (see Eqns. 2 and 3), all Brewer DS ozone data used in this study have $\mu \leq 3.5$.**

3.1.2 Third-party scheme

The design of Model 2 is based on our assumption of the high quality of Brewer ozone data, i.e., the Brewer-derived baseline ozone (Model 1 ozone) is close to the true ozone. In general, for well-calibrated and well-maintained Brewer instruments, this assumption is valid. For example, if Brewer nos. 1 and 2 are in good condition, but Brewer no. 3 is not, Model 1 will show the discrepancy. Then, we can easily identify the issue and re-calibrate Brewer no. 3. However, if Brewer nos. 1 and 2 are the instruments with larger discrepancies from true ozone and Brewer no. 3 is in good condition, then things will become more complex. In addition, whenever we select three instruments to form a “triad” and use Models 1 and 2 to performing the analysis, we also selected baseline ozone defined by those three instruments. In other words, the Models 1 and 2 analyses applied to BrT and BrT-D cannot reflect their relative difference, i.e., BrT uses BrT’s baseline, whereas BrT-D uses BrT-D’s baseline. Thus, to better evaluate and compare BrT and BrT-D’s performances, we need to use a third-party ozone column data as the baseline. Here, Model 3 is designed as:

$$F + \Delta\beta \cdot m = (F'_0 + X) - (\Delta\alpha' + Y) \cdot \Omega_{3rd-party} \cdot \mu \quad (4)$$

325

where, the only difference compared to Model 2 is that we replaced the Model 1 defined baseline ozone with a new third-party baseline ozone ($\Omega_{3rd-party}$). The new baseline can be supplied by either other co-located and independent ozone column observations (e.g., Pandora ozone data) or reanalysis data (e.g., MERRA-2). Please note here the $\Omega_{3rd-party}$ has to be independent of Brewer reference instruments. For example, they cannot be measurements from another Brewer (e.g., another co-located field Brewer instrument), unless it received its ETC constant via the independent calibration method.

330

When a third-party baseline ozone exists, it is easy to evaluate the deviation of each Brewer from the baseline ozone. Thus, in this work, when using the third-party baseline ozone, we simply report their absolute and relative differences defined as:

335

$$\Delta_{abs} = \Omega_{Brewer} - \Omega_{3rd-party} \quad (5)$$

$$\Delta_{rel} = \frac{\Omega_{Brewer} - \Omega_{3rd-party}}{\frac{(\Omega_{Brewer} + \Omega_{3rd-party})}{2}} \times 100\% \quad (6)$$

3.2 Comparison with satellites

Regression analyses between Brewer and satellite observations were made by using the following coincident criteria: (1) nearest (in time) measurement that was within $\pm x$ hr of satellite overpass time, (2) closest satellite ground pixel (having a distance (d , in km) from the ground pixel centre to the location of the Brewers less than y km). These coincident criteria are summarized in Table 3. Only good quality satellite data are used in the analysis. For example, OMTO3 with only error flag = 0 (good sample) are used.

340

Table 3. Satellite comparison criteria.

Satellite (product)	Time criteria	Spatial criteria
	$ \Delta t \leq x$ hr	$d \leq y$ km
OMI (OMDOAO3)	1	30
OMI (OMTO3)	1	30
SBUV-11	2	200
SBUV-14	2	200
SBUV-16	2	200
SBUV-17	2	200
SBUV-18	2	200
SBUV-19	2	200
OMPS	2	50
TOMS	2	50
TROPOMI	0.5	10

345

4 Assessment results

The assessment of the Brewer reference instruments was performed using Models 1, 2, and 3 defined in Section 3. To ensure the assessment is based on good quality data, the data were strictly filtered (i.e., data from single and double spectrometer instruments with reported standard deviation > 3 DU or $\mu > 3.5$ are removed). Using 3 DU (Fioletov et al., 2005) instead of the standard 2.5 DU (Fioletov and Ogyu, 2008) yields more data points and, therefore, more days suitable for comparison, but does not improve the comparison since the additional measurements are the noisiest.

350

4.1 Comparison of ground-based instruments

4.1.1 Model 1

To perform Model 1 analysis, additional criteria are applied. A specific day is analyzed with Model 1 only if each of the three instruments has 1) at least ten measurements on that day and 2) at least three measurements in each half-day (defined by local solar noon time) on that day. The Model 1 analysis was done for BrT and BrT-D separately. The deviations of each individual instrument from their baseline are shown in Fig. 1a, which are comparable to the results in Fig. 1 from Fioletov et al., (2005). The residuals from Model 1 include some remaining instrument uncertainties, but also some short-term fluctuations in ozone, which are not reflected by the second-degree polynomial model. The uncertainties include the effects of instrument temperature fluctuations and the differences in the characteristics of the neutral density (ND) filters. The 5th and 95th percentiles of the Model 1 residuals are shown in Fig. 1b, which are comparable to the results in Fig. 2 from Fioletov et al., (2005). The standard deviation of the residuals is about 2.4 DU or 0.72 %. In general, these updated results show that the performance of the BrT in

360

the last two decades (1999-2019) is comparable to its reported values from 1984 to 2004. The long-term instrument drifts are still typically within $\pm 1\%$. Using the analytical method from the first assessment work (Fioletov et al., 2005), the deviations and residuals are reported with frequencies of 3 months and 1 year, respectively, in Fig. 1. These frequencies were used because they provide a good balance between sampling frequency and sufficient co-incident measurements as well as preserve a potential seasonal component in the differences. The standard deviations (σ) of the 3-month averages plotted in Fig. 1a are 0.43 %, 0.36 %, and 0.42 % ($\bar{\sigma} = 0.40\%$) for Brewers #008, #014, and #015, which are comparable to the reported values from 1984 to 2004 (0.40 %, 0.46%, and 0.39 %). The double triad also shows good long-term stability with the Model 1 analysis, where all measurements are within $\pm 1\%$ compared to its baseline. The standard deviations are 0.44 %, 0.26 %, and 0.33 % ($\bar{\sigma} = 0.34\%$) for Brewers. #145, #187, and #191. From this, assuming that the instrument uncertainties are independent, the standard uncertainty of Brewers (δ) can be estimated as $\sqrt{1.5\bar{\sigma}}$, i.e., 0.49 % and 0.42 % for BrT and BrT-D, respectively.

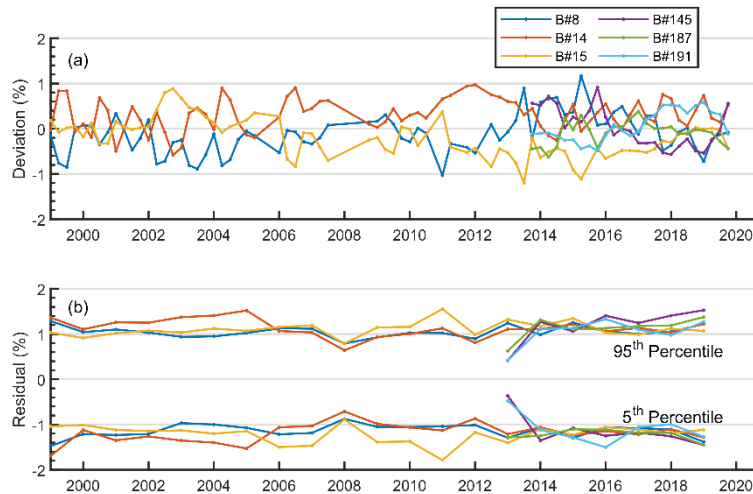


Figure 1. Model 1 estimated deviations and residual of ozone values. (a) Deviations of ozone values of individual triad Brewers from the mean of the three instruments. Each point on (a) represents a 3-month average. Panel (b) shows the 5th and 95th percentiles of the residuals of the Model 1 analysis. Each point on (b) is based on one year of data.

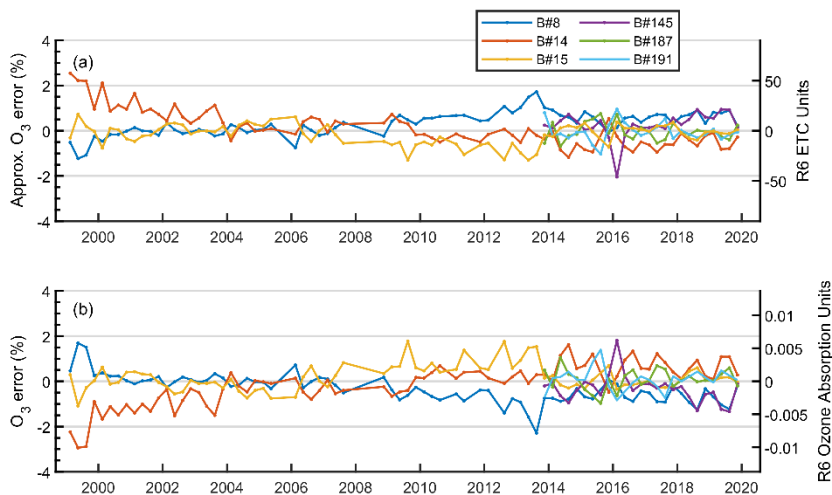
4.1.2 Model 2

The Model 2 analysis was performed for BrT and BrT-D. Figure 2 corresponds to Fig. 4 in Fioletov et al., (2005). In general, Fig. 2 shows the errors in the ETCs and effective ozone absorption coefficients account for up to $\pm 2\%$ of total column ozone, as indicated in Fioletov et al., (2005). Here, the errors in the ETCs and effective ozone absorption coefficients are estimated in R6 ratio units (the units used in the actual Brewer processing algorithm; R6 values corresponding to measured slant column, i.e., $\Omega = \frac{(R6 - ETCO_3)}{10\Delta\alpha\mu}$ in DU; $ETCO_3 = -10^4 \times F_0$). The errors are converted from R6 ratio units to percentages of total column

385 ozone by using typical conditions for Brewer measurements in Toronto (i.e., $\Omega = 330$ DU, $\Delta\alpha = 0.34$, and $\mu = 2$), to provide
 more straightforward values to assess the impact of errors in the ETCs and effective ozone absorption coefficients. For
 example, if we have a model estimated error of ETCO_3 as 50 R6 ratio unit, it will correspond to $\frac{x}{104\alpha\mu\Omega} = 2.2\%$ of total column
 ozone using the typical conditions described above. In typical conditions, the uncertainties of ozone absorption coefficient are
 within ± 1 micrometer step based on the dispersion test, which corresponds to approximately $\pm 0.3\%$ of total column ozone.
 For the uncertainties of ETC, the goal is to have it within ± 5 R6 ratio units.

390

The large errors in ETCs and ozone absorption coefficients may largely compensate for each other and not be evident in the
 Model 1 analysis. This is because Model 2 distributes the residuals (mismatch between observed ozone and baseline ozone)
 into two parts, i.e., X and Y terms in Eqn. 3, which made the retrieved errors negatively correlated. For example, during 2013,
 there were significant errors in the assigned ETCs and absorption coefficients to #008 that was truly caused
 395 by wavelength range limitations of this early model Brewer. A measurement type was added to the schedule of this instrument,
 that when ran, reached the extent of physical travel of the micrometer causing a 2 nm shift in the measurement from the forward
 to the backward scan of the micrometer. The Model 2 results show that the BrT-D has a similar performance compared to the
 BrT since 2013. The errors in ETCs and ozone absorption coefficients from BrT-D (within $\pm 1\%$) are even smaller than those
 from BrT in the most recent period (2017-2019).



400

Figure 2. Relative systematic uncertainties in ETCs and effective ozone absorption coefficients estimated using Model 2. The right y-axes represent the values in the units used in the actual Brewer algorithm (i.e., R6 ratio units); the left y-axes demonstrate the % values of these errors in total ozone values. Each point on the graph represents a 3-month average.

405 **4.1.3 Model 3**

For a third-party-based ozone analysis (Model 3), Brewer and Pandora data are both averaged into 10-minute bins and then paired. Note that the Pandora instrument sampling frequencies were reduced from one measurement every 1.5 minutes in 2013-2017 to one measurement every 5 minutes in 2018-2019 due to change in the observation schedules.

Differences between the Pandora observations and the measurements by individual Brewers are shown in Fig. 3. The gaps in the Pandora record are caused by an instrumental failure in winter 2014. The absolute differences between Brewer and Pandora data are within ± 8 DU. They are slightly larger in wintertime due to the temperature dependency in Pandora ozone data (although empirical correction methods have been applied, the residual effect still exists, e.g., Fig. 13 in Zhao et al., (2016)). Thus, when using Pandora data as a third-party baseline, it is more important to examine the variation of relative differences (i.e., Δ_{rel} of one Brewer minus Δ_{rel} of another Brewer). In the period of the example, the relative differences between Brewer #015 and Brewer #145 are within 5 DU. Thus, the Brewers' performance was, in fact, stable in that period. Figure 3 shows the relative differences, indicating that compared to Pandora, all Brewer reference instruments have long-term stability within ± 2 %. This result is not as good as the prediction from Model 1 (which shows ± 1 % deviations) because even corrected, Pandora data still have some residual seasonal bias. For shorter periods (e.g., summer 2016), all six Brewers have a relative difference within the range from 0 to -2 %, which is comparable to a ± 1 % when Brewer themselves are used as baselines.

420 We also can assess the performance of individual instruments from a third-party ozone baseline. For example, when compared to any other reference instruments, Brewer #015 gave the lowest ozone in the period from 2015 to 2017. Another example is the period after the BrT-D was relocated to Egbert, in which the discrepancy between BrT and BrT-D data became obvious (up to 4 % relative difference between Brewers #008 and #191).

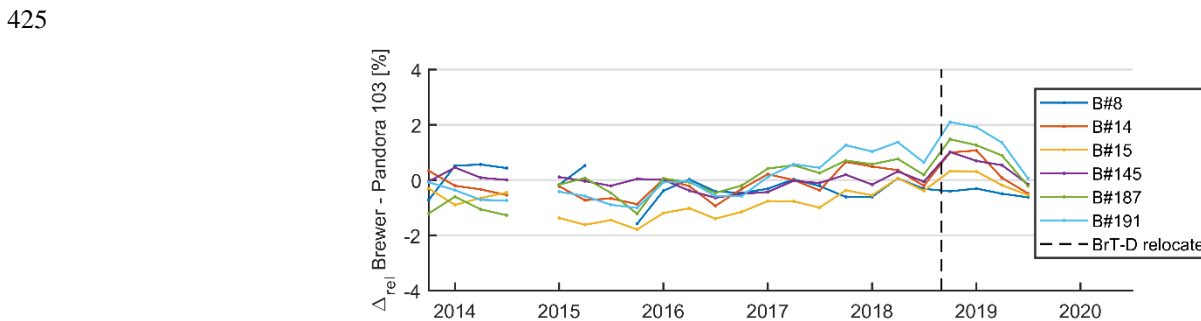


Figure 3. 3-month relative differences between Brewers and Pandora total column ozone. 3-month averages are calculated if there are at least ten coincident measurements between Brewer and Pandora for that period. The black dash line represents the time when BrT-D was relocated to Egbert, i.e., Pandora and BrT-D were not co-located.

430 The Model 3 analysis results are shown in Fig. 4, where the errors in ETCs and ozone absorption coefficients from each Brewer are reported independently. They show that the quality of these instrument “constants” can drift in time due to the nature of

the calibration and maintenance work performed on the instruments. In general, Fig. 4 shows that in most cases, the estimated ETC and effective ozone absorption errors for all reference instruments are within $\pm 2\%$, i.e., similar to the Model 2 results (see Fig. 2).

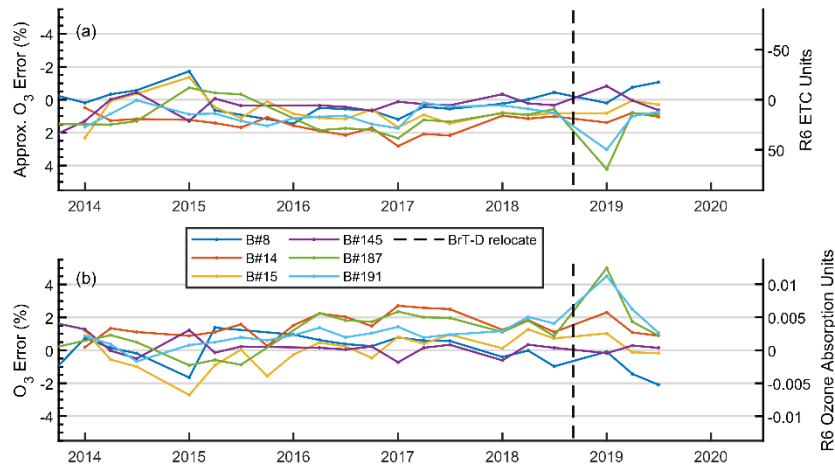


Figure 4. Relative systematic uncertainties in ETCs and effective ozone absorption coefficients estimated using Model 3. Description of y-axes is in Fig. 2. Each point on the graph represents a 3-month average. The black dash line represents the time when BrT-D was relocated to Egbert.

When compared to Model 2, Model 3 provides independent estimates of ETC and effective ozone absorption errors, i.e., errors for BrT and BrT-D can be compared directly. For example, in Fig. 2a, we cannot directly compare the ETC errors from Brewer #014 with those from Brewer #145 because they were evaluated by different baselines. However, with Fig. 4, we can conclude that Brewer #145 has about 1 % lower ETC errors than those for Brewer #014. The detailed results of ETC and ozone absorption coefficients errors are summarized in Table 4. In general, for this assessment period (2013-2019), Brewers #008, #015, and #145 have lower ETC and effective ozone absorption coefficients errors (within $\pm 0.5\%$) when compared to the other Brewer reference instruments.

Table 4a. Mean errors of $\Delta\alpha$ and ETC for Brewer reference instruments (2013-2019) estimated with Model 3.

Brewer serial no.	Mean error of $\Delta\alpha$ [R6 absorption unit]	Mean error of ETC [R6 ETC unit]	Mean error of $\Delta\alpha^*$ [%]	Mean ETC-related error [#] [%]
#008	-0.0002	-1.77	-0.07	-0.08
#014	0.0051	-32.87	1.50	-1.45
#015	-0.0001	-15.64	-0.03	-0.69
#145	0.0007	-8.01	0.21	-0.35
#187	0.0043	-26.84	1.27	-1.19
#191	0.0039	-23.27	1.15	-1.03

* Mean % error in total column ozone, related to error in ozone absorptions; [#] Mean % error in total column ozone, related to error in ETC, corresponding to X when $\mu = 2$, $\Delta\alpha = 0.34$, and $\Omega = 330$ DU (see Eqn. 3).

Table 4b. Mean errors of $\Delta\alpha$ and ETC for Brewer reference instruments estimated with Model 2.

Brewer serial no. [period]	Mean error of $\Delta\alpha$ [R6 absorption unit]	Mean error of ETC [R6 ETC unit]	Mean error of $\Delta\alpha^*$ [%]	Mean ETC-related error [#] [%]
#008 [1999-2019]	-0.0011	6.79	-0.33	0.30
#014 [1999-2019]	-0.0005	3.26	-0.15	0.14
#015 [1999-2019]	0.0006	-3.79	0.17	-0.17
#145 [2013-2019]	-0.0011	5.68	-0.33	0.25
#187 [2013-2019]	0.0026	-0.61	0.08	-0.03
#191 [2013-2019]	0.0026	-1.05	0.08	-0.05

As discussed in Section 4.1.2, sometimes, Model 2 may also overlook issues if two out of three instruments have the compensation effect (i.e., errors in ETCs and ozone absorption coefficients compensate each other). For example, when analyzing Brewer #145 data, it was revealed by the Model 3 analysis that its absorption coefficients were not ideal (in 2014, see Appendix B for more details). The issue was not observed with Model 2 due to Brewer #191 also has a similar issue in the same period. Thus, besides providing independent uncertainties, the Model 3 analysis can provide an important additional quality control process. Details about this additional quality control process are provided in Appendix B.

4.2 Comparison with satellite and reanalysis data

Eleven satellite overpass column ozone datasets are used for data verification of the Brewer reference instruments. Figure 5 shows the relative differences between satellite and Brewer measurements for seasonal (3 months) values are within ± 4 % and yearly values are within ± 3 % (not shown here) in these two decades (1999-2019). The standard deviation ($\sigma_{3\text{month}}$) of the 3-

month Brewer-satellite relative differences is 1.38 %. Detailed regression analysis was also performed and some results are summarized in Fig. 6.

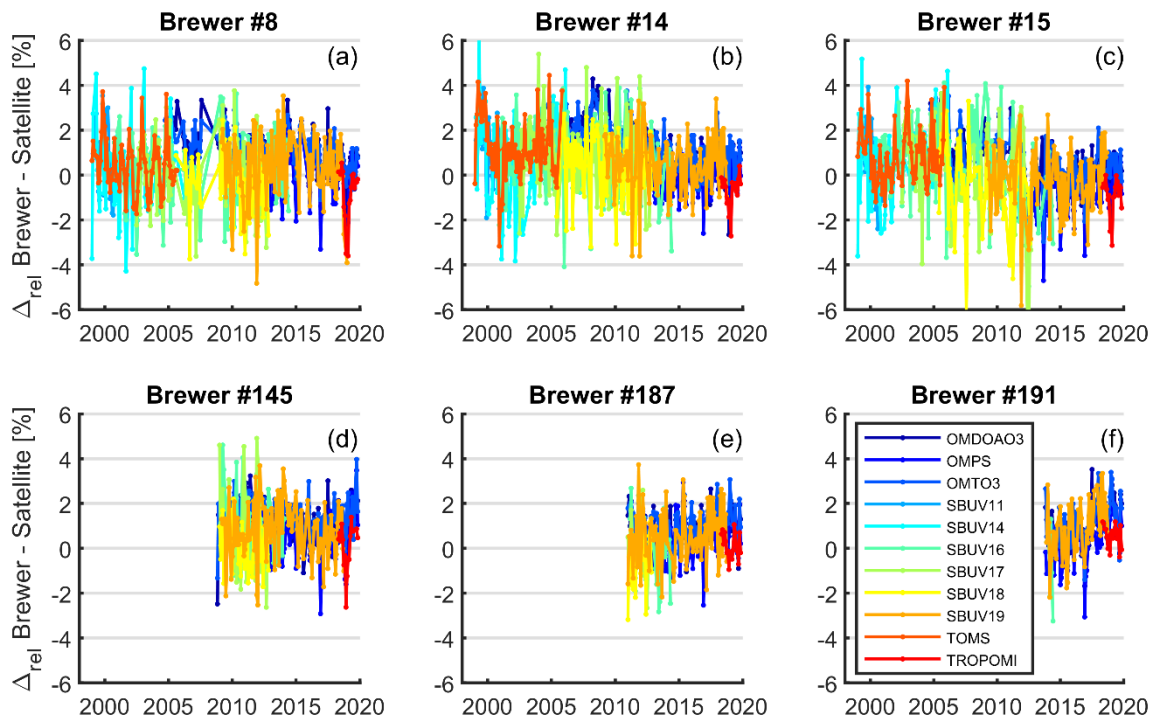
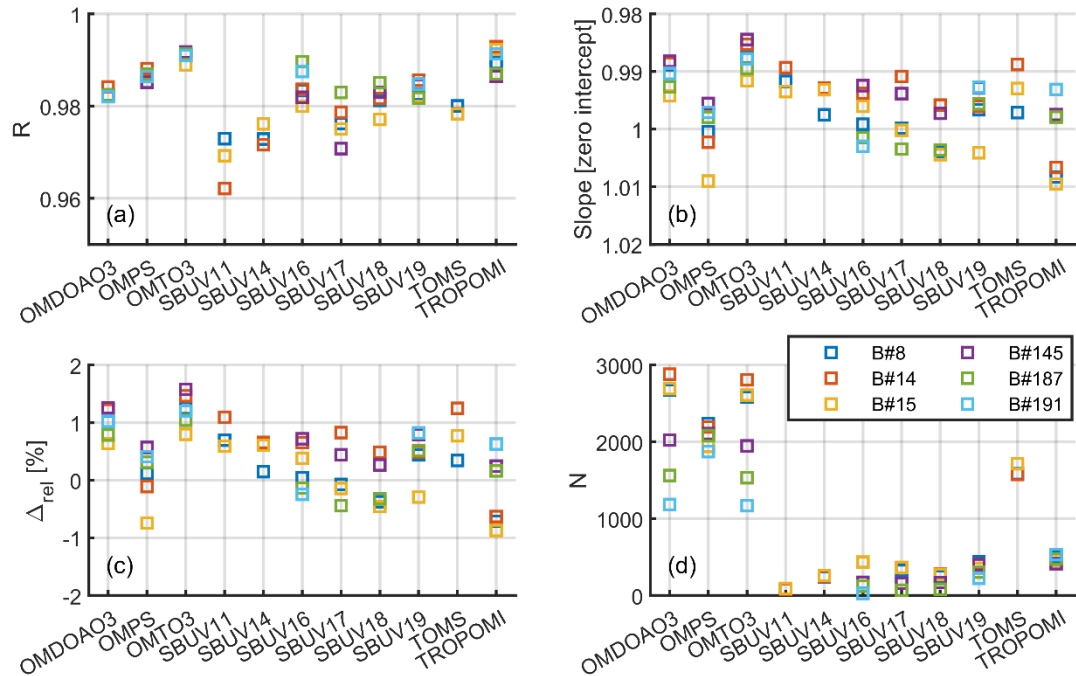


Figure 5. The relative difference between satellites and the world Brewer reference triads (BrT and BrT-D). Each point represents a 3-month average. Brewers and satellite data are paired with the criteria shown in Table 3.



470

Figure 6. Summary of the regression analysis between satellites and the world Brewer reference triads. The four panels represent the (a) correlation coefficient (R) between individual Brewer instruments and different satellites (labelled at the bottom axis), (b) the slope of the zero intercept regression line (multiplicative bias), (c) relative percentage difference (bias), and (d) the total number of coincident observations.

475

In general, the measurements from the individual Brewers have -1 % to 2 % relative difference when compared with all these eleven satellite datasets, with correlation coefficients > 0.96 . For most satellite datasets, the regression with zero intercept (Fig. 6b) also shows that the multiplicative bias between Brewers and satellites are well within ± 1 %. It is known that satellite data also have some biases and drifts (e.g., Antón et al., 2009; Kroon et al., 2008); therefore, the Brewer-satellite difference values alone do not represent the Brewer instrument performance. Comparison with OMI (both versions) shows that besides the 1 % systematic difference between Brewers and satellite data, the spread of biases with individual instruments is also around 1 %.

480

The standard deviation of the Brewer-OMTO3 (OMDOAO3) difference (for 3-month averages) calculated for six instruments is 0.99 % (1.06 %), about 0.5 % higher than Brewers' standard random uncertainties calculated in Section 4.1.1. It is also found that Brewers have lower relative differences compared with OMDOAO3 than OMTO3, which is in agreement with previous researches (e.g., Antón et al., 2009). For high-resolution satellites, such as TROPOMI, the interpretation of the results

485

should be made with extra cautions as the line-of-sight of ground-based and satellite instruments should be accounted for (see more details in Section 6). In general, BrT and BrT-D's stabilities are assessed by using each satellite dataset, via the standard deviations of 3-month Brewer-satellite relative differences, as shown in Table 5. The results show that the BrT-D ($\sigma_{3\text{month}} =$

1.15 %) has a slightly better long-term stability than the BrT ($\sigma_{3\text{month}} = 1.33\%$), which is consistent with the results in Section 4.1.1 that BrT-D has lower random uncertainty than BrT.

Table 5. Mean (Δ_{rel}) and standard deviation ($\sigma_{3\text{month}}$) of the 3-month Brewer-satellite relative differences.

Satellite Dataset	BrT		BrT-D	
	Δ_{rel} [%]	$\sigma_{3\text{month}}$ [%]	Δ_{rel} [%]	$\sigma_{3\text{month}}$ [%]
OMDOAO3	0.84	1.17	0.95	0.86
OMPS	-0.30	1.07	0.39	0.96
OMTO3	1.14	1.08	1.30	0.80
SBUV11	0.93	1.59	N/A	N/A
SBUV14	0.42	1.76	N/A	N/A
SBUV16	0.38	1.59	0.40	1.60
SBUV17	0.26	1.71	0.23	1.75
SBUV18	-0.09	1.60	0.05	1.24
SBUV19	0.21	1.45	0.69	1.26
TOMS	0.82	1.28	N/A	N/A
TROPOMI [*]	-0.84	0.95	0.27	0.73
Mean[#]	0.34	1.33	0.54	1.15

* The comparison includes period when BrT and BrT-D were not collocated (see Section 5 for more details). # Mean: Only include the satellite datasets that have overlap with both BrT and BrT-D. N/A: not applicable.

To compare with the hourly reanalysis data (MERRA-2 column ozone for Toronto), Brewer column ozone data were resampled to hourly mean values. The relative difference in time series is shown in Fig. 7, which demonstrated **similar** long-term stability (i.e., the relative difference within $\pm 2\%$) of the Brewer reference instruments when compared with Pandora or satellite instruments. For example, same as Fig. 3 (comparison with Pandora), Brewer #015 is found to have the lowest column ozone values from 2015 to 2018. In general, the relative differences between Brewers and the reanalysis datasets are within $\pm 2\%$. The inter-instrument differences (i.e., the differences between Brewers) are within $\pm 1\%$ for most of the measurement period.

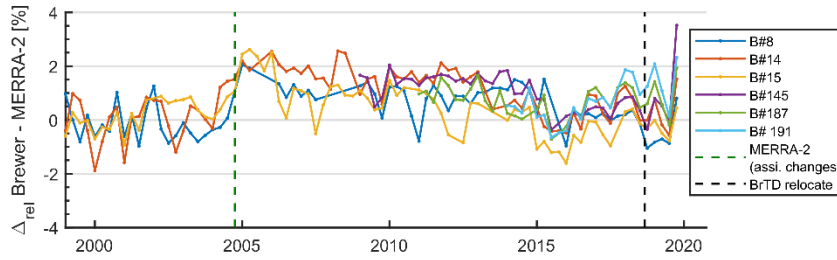


Figure 7. The relative difference between the reference Brewers and MERRA-2 reanalysis. Each point represents a 3-month average. The green dash line represents the time when MERRA-2 changed its assimilation sources from SBUV-2 to MLS/OMI (causing about 2% relative difference). The black dash line represents the time when BrT-D was relocated to Egbert.

505

The shift in relative difference found in 2004 was due to MERRA-2 changing its data assimilation sources (see the green dash line in Fig. 7). MERRA-2 assimilates partial column ozone data from SBUV instruments between January 1980 and September 2004. Starting from October 2004, MERRA-2 assimilates ozone profiles and columns from MLS and OMI instruments (Wargan et al., 2017). For example, the mean Brewer #014 – MERRA-2 relative bias was 0.11 % (Δ_{rel}') for the SBUV-based data assimilation, but it increased to 1.07 % after October 2004, probably due to some bias in OMI data as mentioned previously in Section 4.2. For the MLS/OMI-based assimilation period, the multiplicative biases between individual Brewer instruments and MERRA-2 are from 0.40 % (for Brewer #015) to 1.05 % (for Brewer #014); therefore, the relative biases between Brewers are within 0.65 %. In addition, the standard deviation of the 1-month percentage difference is on average 1.04 % for BrT and 0.87 % for BrT-D. Details of the comparison between Brewer reference instruments and the MERRA-2 reanalysis ozone dataset are summarized in Table 6.

Table 6. Brewer reference instruments vs. MERRA-2 reanalysis ozone dataset.

Brewer serial no.	SBUV-based [1999 – Sep. 2004]			MLS/OMI-based [Oct. 2004 - 2019]		
	Δ_{rel}' [%]	M- Bias* [%]	σ_{1month} [%]	Δ_{rel}' [%]	M- Bias* [%]	σ_{1month} [%]
	#008	-0.13	-0.27	1.14	0.61	0.69
#014	0.11	0.16	1.20	1.07	1.05	1.04
#015	0.21	0.18	1.12	0.39	0.40	1.11
#145	N/A	N/A	N/A	1.01	1.02	0.89
#187	N/A	N/A	N/A	0.79	0.71	0.81
#191	N/A	N/A	N/A	0.76	0.66	0.92

*Multiplicative bias is estimated with the slope of zero intercept linear regression. N/A: not applicable.

5 Discussion

The performance of the European regional reference instruments (i.e., RBCC-E triad) was reported by León-Luis et al., (2018) and compared with the world reference instruments, specifically the BrT. León-Luis et al., (2018) reported that RBCC-E instruments have a mean 3-month standard deviation (δ_{3month}) of 0.27 %, and concluded that the RBCC-E instruments have 36 % lower δ_{3month} when comparing to the world reference instruments (i.e., BrT, 1984-2004 period, $\delta_{3month} = 0.39$ %). However, the comparison was not straightforward. The Model 1 analysis carried out in León-Luis et al., (2018) did not follow the Model 1 design described in Fioletov et al., (2005) and the current work. It is worth noting that the baseline ozone should be the same (except for the offset) for all three RBCC-E instruments. This would be achieved by including the indicator functions described in Section 3.1.1. The 3-month standard deviations of the BrT, BrT-D and RBCC-E instruments (with corresponding data

periods) are summarized in Table 7; however, the results from the RBCC-E instruments should not be directly compared to the ones in Fioletov et al., (2005) or the current work. Moreover, Stübi et al., (2017) examined three Brewer instruments located at Arosa and found a similar performance of short-term variability. They reported that the standard deviation of short-term variability of the Arosa Brewer triad since 1998 was estimated to be about 0.36 % on the scale of a decade. The medium- to long-term stability was estimated to be within ± 0.5 %.

Table 7. World and European regional reference instruments' 3-month standard deviations.

BrT 1999-2019 (1984-2004)		BrT-D 2013-2019		RBCC-E 2005-2016	
Serial no.	$\sigma_{3\text{month}}$ [%]	Serial no.	$\sigma_{3\text{month}}$ [%]	Serial no.	$\sigma_{3\text{month}}$ [%]
#008	0.43 (0.40)	#145	0.44	#157	0.29
#014	0.36 (0.46)	#187	0.26	#183	0.31
#015	0.42 (0.39)	#191	0.33	#185	0.20

*Calculated with a different method.

It is, however, important to understand that there are certain limitations in the Brewer hardware, which explain why the stability below 0.5 % is so difficult to achieve and maintain. For example, it was found that Brewer #015 has a particularly strong temperature dependence where the optical frame was expanding significantly faster than any other Brewer instrument. As a result, the wavelength calibration tests (HG) had to be scheduled more frequently to reduce the impact. However, we should point out that if the time interval between the HG tests is large enough, some measurements can be affected. This issue was fixed in 2017 by replacing the optical frame (details of instrument repair and upgrade history is provided in the supplementary information). A second example is the original configuration of Brewer #145 micrometer was found to have developed wear and became unreliable, causing some wavelength drifts, and as a result, relatively high uncertainties for Brewer #145 as shown in Table 7 (also see larger variations of 3-month deviations from Brewer #145 compared to Brewers #187 and #191 in Fig. 1a). The top and bottom micrometers were fully replaced in 2019, including all the connecting wires of the wire micrometer system.

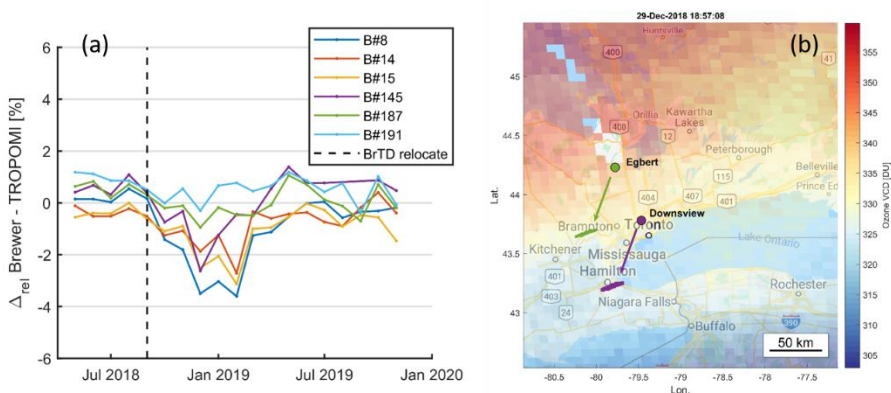
Another example of hardware-related issues with Brewer ozone measurements is the characteristics of the ND filters used to reduce the intensity of incoming radiation (Kerr, 2010). In practice, the filters are not always neutral, but may have some wavelength dependence on their transmittance. The Brewer retrieval algorithm removes effects that are linear as a function of the wavelength, but this offset may not be enough in some cases and a shift of up to a few DU in the retrieved ozone values can occur as a result of a ND filter switch (e.g., from ND filter #1 in the early morning to ND filter #4 in the noon; Savastiouk, 2006). Instruments with ND filters from the same manufactured batch will demonstrate almost identical spectral behaviour.

555 Thus, these instruments may have very similar characteristics, and therefore, demonstrate high precision; however, they all
may be affected by the same or similar hardware-related systematic errors. There are other hardware-related factors that affect
the accuracy and precision of Brewer measurements. For example, a simple replacement of the mercury bulb that is used to
560 ensure the instrument stability could affect total ozone measurements, creating jumps in the data record. The bulb change has
the potential to affect the CalStep (calibration step, the optimal micrometer position found in the “Sun Scan” test; Savastiouk,
2006) of the instrument. If the combined focus of the monochromator mirrors of the instrument (see Savastiouk, 2006 for more
details of instrument’s optical elements) is not optimized and the illuminated filament of the mercury bulb is located in a
significantly different location than the illuminated filament from the original bulb, as much as a 5 micrometer step (one
micrometer step is 0.7 pm) change may be seen. For reference, the effective ozone absorption changes by approximately 1%
every 3 steps, so a 5 steps shift, which is extreme, can give an error of almost 2% in TCO. It is best to change the mercury bulb
565 before it completely fails so that sequential mercury tests can be performed using both bulbs to detect and address any shifts
in the CalStep. It is still recommended to perform the “sun scan” test and verify any potential changes.

The way that the data are processed also affects the results. Siani et al, (2018) concluded that the ozone data processed by
different software agree at the 1 % level; however, some differences can be found depending on the software in use. They also
570 recommended “a rigorous manual data inspection” of the processed data and to be careful with how Standard Lamp (SL) test
results are used. Visual data screening was also used by Stübi et al., (2017b) to eliminate outliers. However, this approach
raises the question of reproducibility of the obtained results and must be carefully documented. For BrT and BrT-D’s data
reprocessing, we recommend using the statistical models developed in relevant studies to help the identifications of potential
hardware or software issues. To keep the integrity of the world reference instruments, data reprocessing could be done only if
575 solid evidence of imperfection of hardware or software have been found and confirmed by Brewer technicians and researchers.

Validation of satellite data is an important application of Brewer measurements and the modern satellite instruments
demonstrated agreement with Brewers within 1 % (e.g. Garane et al., 2019). At the 1 % level, there are many factors that affect
the comparison results. Some of the factors related to ozone absorption cross-sections and their temperature dependence are
580 well established (e.g., Redondas et al., 2014). However, the high spatial resolution of modern satellite instruments such as
TROPOMI brings new challenges. Figure 8 shows that TROPOMI overpass (OVP) data from the Downsview site in Toronto
(centre of ground pixels within 10 km from Downsview) have a better agreement with those of the BrT-D when it was relocated
to Egbert than with those of the Brewers at Toronto. The difference is about 2 %, which is too large to be explained by, for
example, stray light. It is likely related to a difference in viewing geometry. For the Brewer, the light passes through the ozone
585 layer once along the line between the instrument and the sun; for a satellite measurement, the light passes through the ozone
layer in the same way as for ground-based measurements, but then is backscattered by the atmosphere and surface toward the
satellite sensor and passes through the ozone layer again. In the case of a large latitudinal gradient, the thickness of the ozone

layer could be very different (Fig. 8b). As shown by the green and purple lines, the Downsview Brewers were sampling stratospheric ozone over Hamilton, while the Egbert Brewers were sampling stratospheric ozone over west of Brampton (the
 590 Brewers' sampling areas were estimated with viewing geometry of Brewers and MERRA-2 ozone profiles, ground projections of the intersections between the Brewer's line-of-sight and the modelled stratospheric ozone layer; Brampton is about 30 km west of Downsview, Hamilton is about 70 km south-west of Downsview). The previous generations of satellite instruments had spatial resolution in the order of $50 \times 50 \text{ km}^2$ (except for OMI) and the difference in the viewing geometry had only a minor impact. However, for current and future high-resolution satellites, such as TROPOMI and TEMPO (Zoogman et al.,
 595 2014), these sampling effects should be taken into account for future satellite ozone validation works (e.g., Verhoelst et al., 2015). In general, we conclude that all these reference instruments show good long-term stability as well as meet the WMO/GAW requirements.



600 **Figure 8. Example of small scale column ozone field variation. (a) Monthly relative differences between Brewers and TROPOMI total column ozone overpass measurements (for Downsview in Toronto) and (b) TROPOMI total column ozone measured on 29 December 2018 over southern Ontario, masked with Brewers' viewing directions and sampling areas. The base map is from © Google Maps.**

6 Conclusion

This work assessed the long-term performance of the world Brewer reference instruments, maintained by ECCC in Toronto,
 605 Canada, in measuring total column ozone. The last assessment of the BrT was done in 2005 with two decades of ozone data records from 1984 to 2004. This work provides a more recent assessment for the BrT (1999-2019) and reports the first assessment of the BrT-D (2013-2019). It was found that both single and double reference triads met the WMO/GAW ozone monitoring requirements. Using statistical models, both BrT and BrT-D have a better than 0.5 % precision. The 3-month standard deviation of ozone values from the two triads are well within 0.5 %, with BrT-D having slightly better performance
 610 (BrT and BrT-D have mean standard deviations of 0.40 % and 0.34 %, respectively). In addition, the BrT-D has proven to have better performance in low sun conditions (see Appendix A), which provides benefits in ozone monitoring work in the

Polar Regions. Comparison with Pandora total ozone measurements (adjusted for temperature dependence) re-confirmed the high quality of the world Brewer reference instruments. It was found that both BrT and BrT-D have a difference of less than 0.5 %.

615

Further detailed error analysis shows the impacts of ETC and ozone absorption coefficients errors for both reference triads are within ± 2 % when the statistical Model 2 is used. This result is comparable to the BrT findings for data records from 1984 to 2004. When using the Pandora as a reference (Model 3), the ETC and ozone absorption errors from BrT-D are slightly better than the ones from BrT (± 1.5 % and ± 2.0 % for BrT-D and BrT, respectively). It demonstrates that all reference instruments were well-calibrated and maintained in good condition.

620

Differences between the measurements from the individual Brewer triad instruments and eleven satellite datasets are within -1 to +2 %. For most satellite datasets, the multiplicative bias between Brewers and satellites is well within ± 1 %. The viewing geometry (or line-of-sight) of ground-based and satellite instruments should be considered in future high-resolution satellite ozone validation activities. Moreover, a 20-year long-term reanalysis data was compared with the reference Brewers' data record. It shows that the reanalysis data has good quality, with the relative difference between the reference Brewer and the reanalysis datasets being within ± 2 %. However, the changing of assimilation sources will affect the quality of the reanalysis and should be addressed in any ozone trend analysis.

625

630

The **precision** of the Brewer triad instruments are under 0.5 %, while the differences with the best satellite instruments and reanalysis data are close to or slightly lower than 1 %. Further improvement of Brewer total ozone observation precision may be limited by the present Brewer 5-wavelength algorithm and Brewer hardware itself.

635

Data availability. Brewer data are available from WOUDC (<https://woudc.org/>, last accessed: June 2020). Pandora data are available from the Pandonia Global Network (<http://data.pandonia-global-network.org/>, last accessed: June 2020). SBUV data are available from <https://acd-ext.gsfc.nasa.gov/anonftp/toms/sbuV/AGGREGATED/>, last accessed: June 2020. OMI data are available from <https://gs614-avdc1-pz.gsfc.nasa.gov>, last accessed: June 2020. OMPS-NM data are available from doi:10.5067/0WF4HAAZ0VHK, last accessed: June 2020. **TROPOMI data are available from <http://www.tropomi.eu/data-products/total-ozone-column>, last accessed: October 2020.** MERRA-2 data are available from doi:10.5067/VJAFPLI1CSIV, last accessed: June 2020. Any additional data may be obtained from Xiaoyi Zhao (xiaoyi.zhao@canada.ca).

640

645 *Author contributions.* XZ analyzed the data and prepared the manuscript, with significant conceptual input from VF, and
critical feedback from all co-authors. MB, RS, AO, VF and SCL operated and managed the Canadian Brewer
Spectrophotometer Network. IA, MB, AO, VF and XZ performed all Brewer triads' data preparation. VS and MB performed
independent calibrations for the Canadian Brewers. JD, XZ, VF and SCL operated and managed the Canadian Pandora
measurement program. AC, MT, and MM from the Pandonia Global Network team provided critical technical support to the
Canadian Pandora measurement program. MM and MT performed the Pandora ozone harmonization work for the Downsview
650 site in Toronto. DG and CM provided TROPOMI ozone data and supported satellite comparison.

Acknowledgements. We gratefully acknowledge the National Oceanic and Atmospheric Administration (NOAA) and Mauna
Loa Observatory staff for their support in Brewer calibrations. We thank the State Meteorological Agency of Spain
(AEMET) group for hosting Brewer #008 and #145 in the 2008 calibration campaign in Izaña. **We thank James Kerr for his
655 valuable comments and suggestion for this work; Dr. Kerr co-invented the Brewer spectrophotometer when he was a Research
Scientist at ECCC, and is previously a Scientist Emeritus at ECCC.** We thank Nader Abuhassan, Daniel Santana Diaz, Manuel
Gebetsberger, and others from SciGlob and the Pandonia Global Network (PGN) for their technical support of Canadian
Pandora measurements. The PGN is a bilateral project supported with funding from the National Aeronautics and Space
Administration (NASA) and the European Space Agency (ESA). We acknowledge NASA Earth Science Division for
660 providing OMITO3 data. The Sentinel 5 Precursor TROPOMI Level 2 product is developed with funding from the Netherlands
Space Office (NSO) and processed with funding from ESA. We thank NOAA for providing NOAA SBUV data. We
acknowledge the Goddard Earth Sciences Data and Information Services Center (GES DISC) for providing OMPS data. We
thank the Global Modeling and Assimilation Office (GMAO) for providing MERRA-2 data.

References

- 665 Antón, M., López, M., Vilaplana, J. M., Kroon, M., McPeters, R., Bañón, M. and Serrano, A.: Validation of OMI-TOMS and
OMI-DOAS total ozone column using five Brewer spectroradiometers at the Iberian peninsula, *J. Geophys. Res.*, 114,
doi:10.1029/2009jd012003, 2009.
- Bais, A. F., Zerefos, C. S. and McElroy, Ct.: Solar UVB measurements with the double- and single-monochromator Brewer
ozone spectrophotometers, *Geophys. Res. Lett.*, 23, 833–836, 1996.
- 670 Bass, A. M. and Paur, R. J.: The ultraviolet cross-sections of ozone: I. The measurements, in *Atmospheric Ozone*, pp. 606–
610, Springer, Germany., 1985.
- Bhartia, P. K. and Wellemeyer, C. W.: OMI TOMS-V8 Total O₃ algorithm, algorithm theoretical baseline document: OMI
ozone products, NASA Goddard Space Flight Center, Greenbelt, Md., 2002.

Brewer, A. W.: A replacement for the Dobson spectrophotometer?, *Pure Appl. Geophys.*, 106, 919–927, 1973.

675 Cede, A.: Manual for Blick Software Suite 1.6, 2019.

Cede, A., Herman, J., Richter, A., Krotkov, N. and Burrows, J.: Measurements of nitrogen dioxide total column amounts using a Brewer double spectrophotometer in direct Sun mode, *J. Geophys. Res.*, 111(D5), D05304, doi:10.1029/2005JD006585, 2006.

680 Cede, A., Tiefengraber, M., Gebetsberger, M. and Kreuter, M.: TN on PGN products “correct use” guidelines, Pandonia Global Network. [online] Available from: https://www.pandonia-global-network.org/wp-content/uploads/2020/01/LuftBlick_FRM4AQ_PGUserGuidelines_RP_2019009_v1.pdf (Accessed 13 November 2020), 2019.

685 Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N. and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, *Q. J. Roy. Meteor. Soc.*, 137, 553–597, doi:10.1002/qj.828, 2011.

690 Dobson, G. M. B.: Forty Years’ Research on Atmospheric Ozone at Oxford: a History, *Appl. Optics.*, 7, 387–405, doi:10.1364/ao.7.000387, 1968.

Farman, J. C., Gardiner, B. G. and Shanklin, J. D.: Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction, *Nature*, 315, 207–210, doi:10.1038/315207a0, 1985.

Fioletov, V. E. and Ogyu, A.: Brewer Processing Software, [online] Available from: ftp://exp-studies.tor.ec.gc.ca/pub/Brewer_Processing_Software/brewer_processing_software.pdf (Accessed 1 June 2020), 2008.

695 Fioletov, V. E., Griffioen, E., Kerr, J. B., Wardle, D. I. and Uchino, O.: Influence of volcanic sulfur dioxide on spectral UV irradiance as measured by Brewer Spectrophotometers, *Geophys. Res. Lett.*, 25(10), 1665–1668, doi:10.1029/98GL51305, 1998.

700 Fioletov, V. E., Kerr, J. B., Hare, E. W., Labow, G. J. and McPeters, R. D.: An assessment of the world ground-based total ozone network performance from the comparison with satellite data, *J. Geophys. Res.*, 104(D1), 1737–1747, doi:10.1029/1998JD100046, 1999.

Fioletov, V. E., Kerr, J. B., Wardle, D. I. and Wu, E.: Correction of stray light for the Brewer single monochromator, in *Proceedings of the Quadrennial Ozone Symposium*, p. 37, Sapporo, Japan., 2000.

705 Fioletov, V. E., Kerr, J. B., Wardle, D. I., Krotkov, N. A. and Herman, J. R.: Comparison of Brewer ultraviolet irradiance measurements with total ozone mapping spectrometer satellite retrievals, *Opt. Eng.*, 41(12), 3051–3062, doi:10.1117/1.1516818, 2002.

Fioletov, V. E., Kerr, J. B., McElroy, C. T., Wardle, D. I., Savastiouk, V. and Grajnar, T. S.: The Brewer reference triad, *Geophys. Res. Lett.*, 32, L20805, doi:10.1029/2005GL024244., 2005.

- Fioletov, V. E., McLinden, C. A., McElroy, C. T. and Savastiouk, V.: New method for deriving total ozone from Brewer zenith sky observations, *J. Geophys. Res.*, 116, D08301, doi:10.1029/2010JD015399, 2011.
- 710 Flynn, L., Long, C., Wu, X., Evans, R., Beck, C. T., Petropavlovskikh, I., McConville, G., Yu, W., Zhang, Z., Niu, J., Beach, E., Hao, Y., Pan, C., Sen, B., Novicki, M., Zhou, S. and Seftor, C.: Performance of the Ozone Mapping and Profiler Suite (OMPS) products, *J. Geophys. Res.*, 6181–6195, doi:10.1002/2013JD020467, 2014.
- 715 Garane, K., Koukouli, M.-E., Verhoelst, T., Lerot, C., Heue, K.-P., Fioletov, V., Balis, D., Bais, A., Bazureau, A., Dehn, A., Goutail, F., Granville, J., Griffin, D., Hubert, D., Keppens, A., Lambert, J.-C., Loyola, D., McLinden, C., Pazmino, A., Pommereau, J.-P., Redondas, A., Romahn, F., Valks, P., Roozendael, M. V., Xu, J., Zehner, C., Zerefos, C. and Zimmer, W.: TROPOMI/S5P total ozone column data: global ground-based validation and consistency with other satellite missions, *Atmos. Meas. Tech.*, 12(10), 5263–5287, doi:10.5194/amt-12-5263-2019, 2019.
- Gueymard, C. A.: The sun's total and spectral irradiance for solar energy applications and solar radiation models, *Sol. Energy*, 76(4), 423–453, doi:10.1016/j.solener.2003.08.039, 2004.
- 720 Herman, J., Cede, A., Spinei, E., Mount, G., Tzortziou, M. and Abuhassan, N.: NO₂ column amounts from ground-based Pandora and MFDOAS spectrometers using the direct-sun DOAS technique: Intercomparisons and application to OMI validation, *J. Geophys. Res.*, 114(D13), D13307, doi:10.1029/2009JD011848, 2009.
- 725 Herman, J., Evans, R., Cede, A., Abuhassan, N., Petropavlovskikh, I. and McConville, G.: Comparison of ozone retrievals from the Pandora spectrometer system and Dobson spectrophotometer in Boulder, Colorado, *Atmos. Meas. Tech.*, 8, 3407–3418, doi:10.5194/amt-8-3407-2015, 2015.
- J. P. Veefkind, J. F. de Haan, E. J. Brinksma, M. Kroon and P. F. Levelt: Total ozone from the ozone monitoring instrument (OMI) using the DOAS technique, *IEEE Trans. Geosci. Remote Sens.*, 44, 1239–1244, doi:10.1109/tgrs.2006.871204, 2006.
- 730 Karppinen, T., Redondas, A., García, R. D., Lakkala, K., McElroy, C. T. and Kyrö, E.: Compensating for the Effects of Stray Light in Single-Monochromator Brewer Spectrophotometer Ozone Retrieval, *Atmos.-Ocean*, 53(1), 66–73, doi:10.1080/07055900.2013.871499, 2015.
- Kazadzis, S., Bais, A., Kouremeti, N., Gerasopoulos, E., Garane, K., Blumthaler, M., Schallhart, B. and Cede, A.: Direct spectral measurements with a Brewer spectroradiometer: absolute calibration and aerosol optical depth retrieval, *Appl. Opt.*, 44(9), 1681–1690, doi:10.1364/AO.44.001681, 2005.
- 735 Kerr, J. B.: New methodology for deriving total ozone and other atmospheric variables from Brewer spectrophotometer direct sun spectra, *J. Geophys. Res.*, 107(D23), 4731, doi:10.1029/2001JD001227, 2002.
- Kerr, J. B.: The Brewer Spectrophotometer, in *UV Radiation in Global Climate Change: Measurements, Modeling and Effects on Ecosystems*, edited by W. Gao, J. R. Slusser, and D. L. Schmoldt, pp. 160–191, Springer, Berlin, Heidelberg., 2010.
- Kerr, J. B., McElroy, C. T. and Olafson, R. A.: Measurements of ozone with the Brewer ozone spectrophotometer, in *Proceedings of the Quadrennial Ozone Symposium*, pp. 74–79, Boulder, U.S.A., 1981.
- 740 Kerr, J. B., Evans, W. F. J. and Asbridge, I. A.: Recalibration of Dobson Field Spectrophotometers with a Travelling Brewer Spectrophotometer Standard, in *Atmospheric Ozone*, edited by C. S. Zerefos and A. Ghazi, pp. 381–386, Springer Netherlands, Dordrecht., 1985.

- Kerr, J. B., Asbridge, I. A. and Evans, W. F. J.: Intercomparison of total ozone measured by the Brewer and Dobson spectrophotometers at Toronto, *J. Geophys. Res.*, 93, 11129–11140, doi:10.1029/JD093iD09p11129, 1988.
- 745 Kerr, J. B., McElroy, C. T. and Wardle, D. I.: The Brewer instrument calibration center 1984-1996, pp. 915–918., 1998.
- Kramarova, N. A., Nash, E. R., Newman, P. A., Bhartia, P. K., McPeters, R. D., Rault, D. F., Seftor, C. J., Xu, P. Q. and Labow, G. J.: Measuring the Antarctic ozone hole with the new Ozone Mapping and Profiler Suite (OMPS), *Atmos. Meas. Tech.*, 14(5), 2353–2361, doi:10.5194/acp-14-2353-2014, 2014.
- 750 Kroon, M., Veefkind, J. P., Sneep, M., McPeters, R. D., Bhartia, P. K. and Levelt, P. F.: Comparing OMI-TOMS and OMI-DOAS total ozone column data, *J. Geophys. Res.*, 113, doi:10.1029/2007jd008798, 2008.
- Kurucz, R. L.: New atlases for solar flux, irradiance, central intensity, and limb intensity, *Mem. S.A.It. Suppl.*, 8, 189, 2005.
- Kyrö, E.: Intercomparison of total ozone data from Nimbus 7 TOMS, the Brewer UV Spectrophotometer and SAOZ UV-Visible Spectrophotometer at High Latitudes Observatory, Sodankylä, *Geophys. Res. Lett.*, 20(7), 571–574, doi:10.1029/93GL00806, 1993.
- 755 Labow, G. J., McPeters, R. D., Bhartia, P. K. and Kramarova, N.: A comparison of 40 years of SBUV measurements of column ozone with data from the Dobson/Brewer network, *J. Geophys. Res.*, 118(13), 7370–7378, doi:10.1002/jgrd.50503, 2013.
- León-Luis, S. F., Redondas, A., Carreño, V., López-Solano, J., Berjón, A., Hernández-Cruz, B. and Santana-Díaz, D.: Internal consistency of the Regional Brewer Calibration Centre for Europe triad during the period 2005–2016, *Atmos. Meas. Tech.*, 11(7), 4059–4072, doi:10.5194/amt-11-4059-2018, 2018.
- 760 Marengo, F., Sarra, A. di and Luisi, J. D.: Methodology for determining aerosol optical depth from brewer 300–320-nm ozone measurements, *Appl. Opt.*, 41(9), 1805–1814, doi:10.1364/AO.41.001805, 2002.
- McPeters, R. D., Bhartia, P. K., Krueger, A. J., Herman, J. R., Wellemeyer, C. G., Seftor, C. J., Jaross, G., Torres, O., Moy, L., Labow, G., Byerly, W., Taylor, S. L., Swissler, T. and Cebula, R. P.: Earth Probe Total Ozone Mapping Spectrometer (TOMS) Data Products User’s Guide, 1998.
- 765 Ramaswamy, V., Schwarzkopf, M. D. and Shine, K. P.: Radiative forcing of climate from halocarbon-induced global stratospheric ozone loss, *Nature*, 355, 810–812, 1992.
- Redondas, A., Evans, R., Stuebi, R., Köhler, U. and Weber, M.: Evaluation of the use of five laboratory-determined ozone absorption cross sections in Brewer and Dobson retrieval algorithms, *Atmos. Chem. Phys.*, 14, 1635–1648, doi:10.5194/acp-14-1635-2014, 2014.
- 770 Redondas, A., Carreño, V., León-Luis, S. F., Hernández-Cruz, B., López-Solano, J., Rodríguez-Franco, J. J., Vilaplana, J. M., Gröbner, J., Rimmer, J., Bais, A. F., Savastiouk, V., Moreta, J. R., Boulkelia, L., Jepsen, N., Wilson, K. M., Shiroto, V. and Karpinen, T.: EUBREWNET RBCC-E Huelva 2015 Ozone Brewer Intercomparison, *Atmos. Chem. Phys.*, 18(13), 9441–9455, doi:10.5194/acp-18-9441-2018, 2018.
- 775 Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L. and Kim, G.-K.: MERRA: NASA’s modern-era retrospective analysis for research and applications, *J. Climate.*, 24, 3624–3648, 2011.

Rimmer, J. S., Redondas, A. and Karppinen, T.: EuBrewNet – A European Brewer network (COST Action ES1207), an overview, *Atmos. Chem. Phys.*, 18(14), 10347–10353, doi:10.5194/acp-18-10347-2018, 2018.

780 Savastiouk, V.: Improvements to the direct-sun ozone observations taken with the Brewer spectrophotometer, Ph.D Thesis, York University, Canada., 2006.

Serdyuchenko, A., Gorshelev, V., Weber, M., Chehade, W. and Burrows, J. P.: High spectral resolution ozone absorption cross-sections - Part 2: Temperature dependence, *Atmos. Meas. Tech.*, 7, 625–636, doi:10.5194/amt-7-625-2014, 2014.

785 Siani, A. M., Frasca, F., Scarlatti, F., Religi, A., Diémoz, H., Casale, G. R., Pedone, M. and Savastiouk, V.: Examination on total ozone column retrievals by Brewer spectrophotometry using different processing software, *Atmos. Meas. Tech.*, 11(9), 5105–5123, doi:10.5194/amt-11-5105-2018, 2018.

Solomon, S., Garcia, R. R., Rowland, F. S. and Wuebbles, D. J.: On the depletion of Antarctic ozone, *Nature*, 321, 755–758, doi:10.1038/321755a0, 1986.

790 Staehelin, J., Renaud, A., Bader, J., McPeters, R., Viatte, P., Hoegger, B., Bugnion, V., Giroud, M. and Schill, H.: Total ozone series at Arosa (Switzerland): Homogenization and data comparison, *J. Geophys. Res.*, 103(D5), 5827–5841, doi:10.1029/97JD02402, 1998.

Stolarski, R. S., Krueger, A. J., Schoeberl, M. R., McPeters, R. D., Newman, P. A. and Alpert, J. C.: Nimbus 7 satellite measurements of the springtime Antarctic ozone decrease, *Nature*, 322, 808–811, doi:10.1038/322808a0, 1986.

Stolarski, R. S., Bloomfield, P., McPeters, R. D. and Herman, J. R.: Total Ozone trends deduced from Nimbus 7 Toms data, *Geophys. Res. Lett.*, 18, 1015–1018, doi:10.1029/91gl01302, 1991.

795 Stübi, R., Schill, H., Klausen, J., Vuilleumier, L., Gröbner, J., Egli, L. and Ruffieux, D.: On the compatibility of Brewer total column ozone measurements in two adjacent valleys (Arosa and Davos) in the Swiss Alps, *Atmos. Meas. Tech.*, 10(11), 4479–4490, doi:10.5194/amt-10-4479-2017, 2017a.

Stübi, R., Schill, H., Klausen, J., Vuilleumier, L. and Ruffieux, D.: Reproducibility of total ozone column monitoring by the Arosa Brewer spectrophotometer triad, *J. Geophys. Res.*, 122(8), 4735–4745, doi:10.1002/2016JD025735, 2017b.

800 Thuillier, G., Floyd, L., Woods, T. N., Cebula, R., Hilsenrath, E., Hersé, M. and Labs, D.: Solar irradiance reference spectra for two solar active levels, *Adv. Space Res.*, 34(2), 256–261, doi:10.1016/j.asr.2002.12.004, 2004.

Tzortziou, M., Herman, J. R., Cede, A. and Abuhassan, N.: High precision, absolute total column ozone measurements from the Pandora spectrometer system: Comparisons with data from a Brewer double monochromator and Aura OMI, *J. Geophys. Res.*, 117, D16303, doi:10.1029/2012JD017814, 2012.

805 Van Hoosier, M. E.: Solar ultraviolet spectral irradiance data with increased wavelength and irradiance accuracy, in *Proceedings of Ultraviolet Atmospheric and Space Remote Sensing: Methods and Instrumentation*, vol. 2831, pp. 57–64, International Society for Optics and Photonics, Denver, CO, United States., 1996.

810 Verhoelst, T., Granville, J., Hendrick, F., Köhler, U., Lerot, C., Pommereau, J.-P., Redondas, A., Van Roozendaal, M. and Lambert, J.-C.: Metrology of ground-based satellite validation: co-location mismatch and smoothing issues of total ozone comparisons, *Atmos. Meas. Tech.*, 8(12), 5039–5062, doi:10.5194/amt-8-5039-2015, 2015.

Wardle, D. I., McElroy, C. T., Kerr, J. B., Wu, E. and Lamb, K.: Laboratory tests on the double Brewer spectrophotometer, in Proceedings of the Quadrennial Ozone Symposium, pp. 997–1000, L'Aquila, Italy., 1996.

Wargan, K., Labow, G., Frith, S., Pawson, S., Livesey, N. and Partyka, G.: Evaluation of the Ozone Fields in NASA's MERRA-2 Reanalysis, *J. Climate*, 30(8), 2961–2988, doi:10.1175/JCLI-D-16-0699.1, 2017.

815 WMO: World Meteorological Organization Global Atmosphere Watch - Global Atmosphere Watch Measurements Guide, WMO TD No. 1073, [online] Available from: https://library.wmo.int/doc_num.php?explnum_id=4632 (Accessed 24 November 2020), 2001.

WMO: Scientific assessment of ozone depletion: 2018, Global Ozone Research and Monitoring Project, World Meteorological Organization, Geneva, Switzerland., 2018.

820 Zanjani, Z. V., Moeini, O., McElroy, T., Barton, D. and Savastouk, V.: A calibration procedure which accounts for non-linearity in single-monochromator Brewer ozone spectrophotometer measurements, *Atmos. Meas. Tech.*, 12(1), 271–279, doi:10.5194/amt-12-271-2019, 2019.

825 Zerefos, C. S., Eleftheratos, K., Kapsomenakis, J., Solomos, S., Inness, A., Balis, D., Redondas, A., Eskes, H., Allaart, M., Amiridis, V., Dahlback, A., De Bock, V., Diémoz, H., Engelmann, R., Eriksen, P., Fioletov, V., Gröbner, J., Heikkilä, A., Petropavlovskikh, I., Jaroslowski, J., Josefsson, W., Karppinen, T., Köhler, U., Meleti, C., Repapis, C., Rimmer, J., Savinykh, V., Shirov, V., Siani, A. M., Smedley, A. R. D., Stanek, M. and Stübi, R.: Detecting volcanic sulfur dioxide plumes in the Northern Hemisphere using the Brewer spectrophotometers, other networks, and satellite observations, *Atmos. Chem. Phys.*, 17(1), 551–574, doi:10.5194/acp-17-551-2017, 2017.

830 Zhao, X., Fioletov, V., Cede, A., Davies, J. and Strong, K.: Accuracy, precision, and temperature dependence of Pandora total ozone measurements estimated from a comparison with the Brewer triad in Toronto, *Atmos. Meas. Tech.*, 9, 5747–5761, doi:10.5194/amt-9-5747-2016, 2016.

Zhao, X., Weaver, D., Bognar, K., Manney, G., Millán, L., Yang, X., Eloranta, E., Schneider, M. and Strong, K.: Cyclone-induced surface ozone and HDO depletion in the Arctic, *Atmos. Chem. Phys.*, 17(24), 14955–14974, doi:10.5194/acp-17-14955-2017, 2017.

835 Zhao, X., Bognar, K., Fioletov, V., Pazmino, A., Goutail, F., Millán, L., Manney, G., Adams, C. and Strong, K.: Assessing the impact of clouds on ground-based UV–visible total column ozone measurements in the high Arctic, *Atmos. Meas. Tech.*, 12(4), 2463–2483, doi:10.5194/amt-12-2463-2019, 2019.

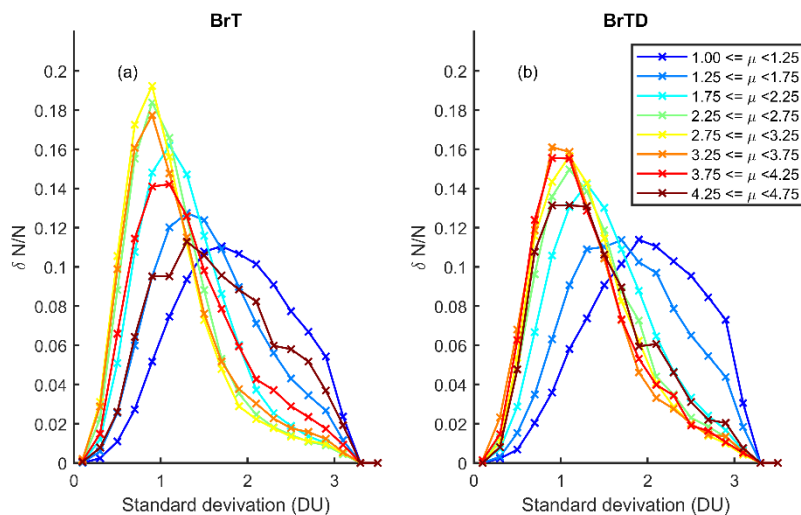
840 Zoogman, P., Jacob, D. J., Chance, K., Liu, X., Lin, M., Fiore, A. and Travis, K.: Monitoring high-ozone events in the US Intermountain West using TEMPO geostationary satellite observations, *Atmos. Chem. Phys.*, 14(12), 6261–6271, doi:10.5194/acp-14-6261-2014, 2014.

Appendices

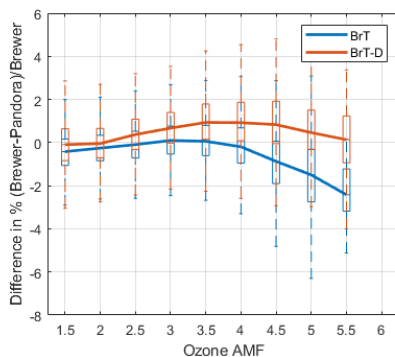
A. Distribution of standard deviations of individual DS measurements

845 Figure A1 shows the distribution of the measurement standard deviation (δ_M), which is used to determine the acceptability of each DS ozone data point in the Brewer data processing algorithm. For Brewers, each final DS ozone data point is a mean of five individual measurements (performed within 3 minutes), and the δ_M is the standard deviation of these five measurements. Typically, the total column ozone values are assumed to be stable within the time of these five measurements. Thus, any DS ozone data with $\delta_M > 3$ DU will be removed. Figure 3a in Fioletov et al., (2005) shows the distribution of δ_M for BrT with $\mu \leq 3.25$. Since the δ_M is proportional to the measured quality F divided by μ , the variability of F (among five measured F) is also
850 influenced by μ . For example, in the $1.00 \leq \mu \leq 1.25$ range, δ_M of BrT has a peak value of about 1.8 DU. However, in a higher range of $2.75 \leq \mu \leq 3.25$, δ_M of BrT has a peak value of about 1 DU.

Typically, Brewer DS ozone data are reported only when $\mu \leq 3.5$ (Note: Except this section, all Brewer DS ozone data used in this study have $\mu \leq 3.5$). This is because, for single spectrometer Brewers, measurements at high μ values are strongly affected
855 by the stray light (Bais et al., 1996; Fioletov et al., 2000; Wardle et al., 1996). The double Brewers were designed to have low stray light (i.e., internal stray light fraction of 10^{-7} and 10^{-5} for double and single Brewers, respectively) and showed good performance when $\mu > 3.5$ (e.g., Zhao et al., 2016). To demonstrate the benefits of low stray light in double Brewer instruments and make a direct comparison between BrT and BrT-D, the μ range is extended to higher values ($\mu \leq 4.75$) in this analysis. Figure A1 shows that for typical $\mu \leq 3.25$ conditions, BrT-D has similar performance as BrT. Whereas, for low solar zenith
860 angle (SZA) conditions (e.g., $4.25 \leq \mu \leq 4.75$), double Brewers still have similar distributions at moderate SZA conditions. Please note that since BrT only reports ozone data with $\mu \leq 3.5$, to make sure the comparison and assessment provided in this work is comparable to Fioletov et al., (2005), both BrT and BrT-D data used in any other sections are filtered with the $\mu \leq 3.5$ criteria. However, the capability of measuring ozone value in low sun conditions is very important for the ozone monitoring in Polar Regions where the SZA is large in early springtime. This stray light effect is further illustrated in Fig. A2, in which
865 the percentage difference between Pandora and BrT(BrT-D) are binned by ozone air mass factors. Figure A2 indicates that in low air mass conditions ($AMF < 3.5$), BrT and BrT-D have similar air mass dependence, which is consistent with the results reported by Tzortziou et al. (2012) and Zhao et al. (2016). Note that Fig. A2 is similar to Figure 15 in Zhao et al. (2016), but with an extended dataset (2013-2015 in Zhao et al. (2016), 2013-2019 in this work). It is found that the air mass dependencies of BrT and BrT-D are consistent within these two periods. Further information on relative difference between BrT and BrT-D, in terms of air mass factor and slant column ozone are provided in Fig. S2.
870



875 **Figure A1.** The distribution of the standard deviations of individual DS measurements as a function of air mass value. (a) shows the Brewer reference triad (BrT) data (1999-2019), (b) shows the double Brewer reference triad (BrT-D) data (2013-2019). Data from all three Brewers for each triad were used for this plot.



880 **Figure A2.** The percentage difference between Pandora and Brewers (grouped as BrT and BrT-D) as a function of ozone air mass factor. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points not consider outliers.

B. Model 3 analysis improvement examples

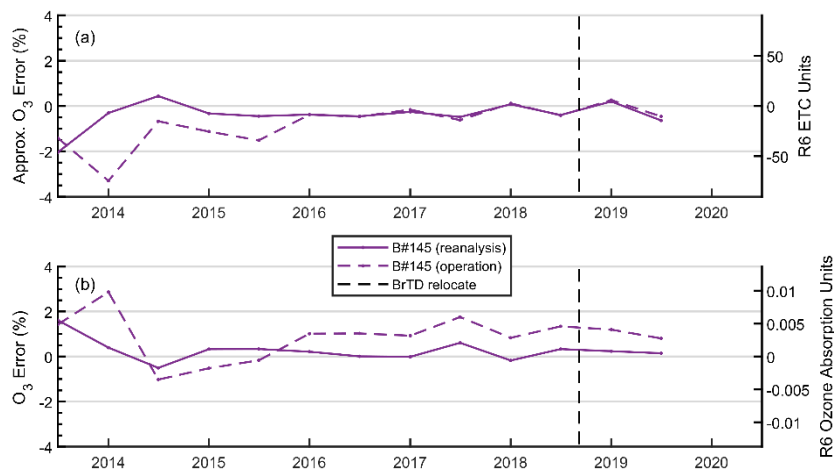


Figure B1. Comparison between reprocessed and operational data from Brewer #145. (a) Relative systematic uncertainties in ETCs and (b) ozone absorption coefficients estimated using Model 3. Description of y-axes in Fig. 2. Each point on the graph represents a 6-month average. The black dash line represents the time when BrT-D was relocated to Egbert.

885

The early operational processing run of the Brewer triad data, when reviewed through Model 3, indicated that there were some errors in the ETC and absorption values, but were compensating each other when ozone values were calculated. As a result, the used configuration produced a reasonable daily average ozone, but not individual values. For example, Fig. B1 shows that the ETC error in early 2014 was as large as 4 % and the ozone absorption error was about 3 % in the operational processing version. After this observation, the data were scrutinized to find that a calibration step had inadvertently been changed by 5 steps from what was intended. An artificial offset in ozone absorption was introduced in an equal offset to the change in the calibration step to correct for this error. The solid line in Fig. B1 indicates the improvement made.

890

A. Brewer Single and Double Triad Instrument Repair and Upgrade Review: 1999 – 2019

Brewer #008

Single monochromator, multi-board electronics Brewer c/w pushrod micrometer design.

- 5 • May 2000: Zenith positioning issues were noted. Some zenith sky measurements, due to poor zenith pointing, may be incorrect, and depending on the filter criteria and other factors, the incorrect data may or may not be omitted. For example, incorrect data may be uniform enough to clear filters and be considered a good measurement. The zenith drive bearing was replaced.
- 10 • August 2001: Degrading Run/Stop tests noted. Some data, due to intermittent shutter movement quality, may be incorrect and be omitted. Shutter motor replaced Aug. 22 – 24. Testing after the motor replacement resulted in a dead time change from 49 ns to 45 ns.
- January 2002: Secondary power supply failure and replacement. Board replaced and board settings adjusted to replicate the original voltage settings. Measurement failures occurred until deficiency was addressed.
- 15 • March 2002: Dead Time tests poor and Standard Lamp (SL) voltage too low. Not able to set to historic values - replaced faulty Secondary Power Supply (SPS).
- January 2004: Intermittent power issues – cable replaced. Intermittent power issues - power switch faulty. Some zenith sky measurements, due to poor azimuth tracking, may be incorrect and depending on the filter criteria and other factors, the incorrect data may or may not be omitted. For example, incorrect data may be uniform enough to clear filters and be considered a
- 20 good measurement.
- January 2005: New heater system. Original design attached to instrument structure and had potential to affect data by producing temperature gradients through the spectrometer frame.
- March 2005: Power issues caused by a faulty power harness. Harness replaced. Measurements missed or failed during intermittent problems.
- 25 • July 2007: Grating arm “pin flag” replaced with a Delrin flag. The flag change necessitated a micrometer offset change equal to the change in flag position on the grating arm, leaving the optical reference to mercury emission line distance unchanged. Occurrences of the flag being missed by the optical sensor and jamming were eliminated.
- March 2008: Jumps seen within scan measurements were attributed to binding between the pushrod and the pointed grating arm cones during micrometer movement. The pointed grating arm cone was replaced with a ball ended grating arm cone to ensure uniform movement.
- 30 • April 2008: Installed a viewing window cover. Visible light entering the viewing ports could be adding to the stray light measured by single monochromator Brewers. Adding the cover eliminated this possibility.
- 35 • January 2009: Replaced braided line driven iris drive with a worm motor driven iris drive. Poor data caused by periodic breakage in the braided line resolved.
- January 2010: Proportional heater control and humidity sensor installed. Environmental stability reduced the necessity of diagnostics tests. More time for atmospheric measurements.
- 40 • March 2011: Installed a rotary connector into Brewer #008’s tracker. Data lost due to cable damage and azimuth tracker failures resolved.
- October 2012: Replaced A/D (Analog-to-Digital) board. Diagnostics returned to normal. Measurements not affected.
- May 2015: Adjusted shutter motor height; increased SL voltage and lowered dead time during instrument review in preparation for the calibration at Mauna Loa Observatory (MLO).
- 45 • March 2016: Zenith motor failure. Some zenith sky measurements, due to poor zenith pointing may be incorrect, and depending on the filter criteria and other factors, the incorrect data may or may not be omitted. For example, incorrect data may be uniform enough to clear filters and be considered a good measurement. The zenith motor was replaced.

Brewer #014

50 Single monochomator, multi-board electronics Brewer c/w pushrod micrometer design.

- September 2000: Replaced braided line driven iris drive with a worm motor driven iris drive. Poor data caused by periodic breakage in the braided line resolved.
- September 2000: Replaced micrometer ribbon cable, I/O board. Removed, cleaned, lubricated and repositioned micrometer. Intermittent positioning errors would affect data quality.
55 Deficiency addressed.
- October 2002: Azimuth tracker failure. Reference flag missed. Cables unwound and azimuth reference re-established. All measurements stopped until deficiency was addressed.
- March 2004: HG (mercury-line wavelength calibration) lamp failure traced to SPS board failure. Board replaced and board settings adjusted to replicate the original voltage settings.
60 Measurement failures occurred until deficiency was addressed.
- July 2004: Primary power supply failure. Power supply replaced and the 5 volt setting adjusted to replicate the original voltage setting. All measurements stopped until deficiency was addressed.
- February 2008: New heater system. Original design attached to instrument structure and had potential to affect data by producing temperature gradients through the spectrometer frame.
65
- April 2008: Installed a viewing window cover. Visible light entering the viewing ports could be adding to the stray light measured by single monochromator Brewers. Adding the cover eliminated this possibility.
- November 2008: Replaced braided line driven iris drive with a worm motor driven iris drive. Poor data caused by periodic breakage in the braided line resolved.
70
- November 2008: Fore-optics lenses optimized for focus and alignment. Adjustment improved intensity and tracking, improving measurements.
- June 2009: The iris constant was found set to 75 steps and should have been set to 250 steps for new iris that was installed. Field of view on the sun and sky would have been reduced, affecting the measurements and likely reducing the attenuation needed from the neutral density filters during the measurements.
75
- March 2010: Proportional heater control and humidity sensor installed. Environmental stability reduced the necessity of diagnostics tests. More time for atmospheric measurements.
- September 2010: Frequent sighting changes, suggesting tracking issues. Early morning or late afternoon tracking may have been affected. Re-levelled instrument.
80
- February 2011: Azimuth tracker failure. Reference flag missed. Cables unwound and azimuth reference re-established. All measurements stopped until deficiency was addressed.
- January 2012: Standard Lamp burnt out. Lamp replaced. Standard lamp test data interpolation needed over this time period.
85
- May 2011: Frequent sighting changes, suggesting tracking issues. Zenith prism alignment performed and instrument re-levelled.
- August 2013: Standard Lamp burnt out. Lamp replaced. Standard lamp test data interpolation needed over this time period.
- October 2013: Installed a rotary connector into Brewer #014's tracker. Data lost due to cable damage and azimuth tracker failures resolved.
90
- October 2013: Installed longer zenith window. Now able to measure closer to the horizon.

Brewer #015

Single monochomator, multi-board electronics Brewer c/w pushrod micrometer design.

- August 2000: Azimuth tracker failure. Reference flag missed. Cables unwound and azimuth reference re-established. All measurements stopped until deficiency was addressed.
95

- December 2000: Micrometer jam. Removed, cleaned, lubricated and repositioned micrometer. Intermittent positioning errors would affect data quality. Deficiency addressed.
- December 2000: Blown tracker fuse. Fuse replaced. Measurements ceased until deficiency addressed.
- 100 • April 2001: Intermittent azimuth and micrometer failures persisted. BA-E501A I/O board replaced. Board replacement addressed intermittent errors. Measurements missed or fail until deficiency was addressed.
- June 2001: Frequent sighting changes, suggesting tracking issues. Early morning or late afternoon tracking may have been affected. Re-levelled instrument.
- 105 • October 2001: Multi-step corrections between observations noted. Suspected the cause to be micrometer wear. New micrometer installed.
- 2002: Multiple Azimuth tracker failures throughout the year. Reference flag missed. Cables unwound and azimuth reference re-established. All measurements stopped until deficiency was addressed.
- 110 • January 2003: Micrometer positioning discrepancies. Micrometer removed, cleaned and re-installed. Intermittent positioning errors would affect data quality. Deficiency addressed.
- April 2003: Zenith positioning errors. Zenith gear worn and binding, replaced Brass gear c/w shaft and split gear. Intermittent positioning errors would affect data quality. Deficiency addressed.
- 115 • March 2004: Secondary power supply board failure. Board replaced and board settings adjusted to replicate the original voltage settings. Measurement failures occurred until deficiency was addressed.
- July 2004: Primary power supply failure. Power supply replaced and the 5 volt setting adjusted to replicate the original voltage setting. All measurements stopped until deficiency was addressed.
- 120 • January 2005: New heater system. Original design attached to instrument structure and had potential to affect data by producing temperature gradients through the spectrometer frame.
- April 2008: Installed a viewing window cover. Visible light entering the viewing ports could be adding to the stray light measured by single monochromator Brewers. Adding the cover eliminated this possibility.
- 125 • January 2009: Secondary power supply board failure. Board replaced and board settings adjusted to replicate the original voltage settings. Measurement failures occurred until deficiency was addressed. Pointed grating arm cone replaced with ball ended grating arm cone to ensure uniform movement. Jumps seen within scan measurements attributed to binding on the pointed cones. Replaced braided line driven iris drive with a worm motor driven iris drive. Poor data caused by periodic breakage in the braided line resolved.
- 130 • March 2010: Proportional heater control and humidity sensor installed. Environmental stability reduced the necessity of diagnostics tests. More time for atmospheric measurements.
- June 2011: Major failure. Replaced BA-E501A I/O board; microprocessor board; clock board and the shutter motor. All measurements suspended during failure and repair.
- 135 • October 2011: Measurement discrepancy traced to a damaged NiSO₄/UG11 combo filter. Filter replaced. Measurements would be drifting as the filter degraded.
- March 2012: Installed a rotary connector into Brewer #015's tracker. Data lost due to cable damage and azimuth tracker failures resolved.
- 140 • September 2012: Thermister 1 and humidity getting intermittent erroneous readings. Erroneous temperature affects and causes erroneous ozone readings. Not investigated, as the problem seemed to be solved by reseating the instrument electronics boards in the card cage.
- August 2013: Thermister 1 and humidity getting erroneous readings. Not investigated, as the problem seemed to be solved by reseating the instrument electronics boards in the card cage.
- 145 • September 2013: Filter Wheel #2 Assembly (FW2) replacement - non-uniform spacing between filters corrected; Packed up for shipping to MLO.

- October 2013: Installed longer zenith window. Now able to measure closer to the horizon.
- October 2013: It was noted that the aluminum optical frame was expanding 90-100 parts per million per degree Celsius faster than the aluminum pushrod was expanding. The spectrum was drifting with temperature faster than is compensated for by the mercury (Hg) bulb tests. This instrument characteristic will cause the measurements to have greater than normal variability.
- November 2014: Zenith motor failure. Some zenith sky measurements, due to poor zenith pointing may be incorrect, and depending on the filter criteria and other factors, the incorrect data may or may not be omitted. For example, incorrect data may be uniform enough to clear filters and be considered a good measurement. The zenith motor was replaced.
- December 2016: Micrometer positioning discrepancies. Micrometer motor replaced. Intermittent positioning errors would affect data quality. Deficiency addressed.
- October 2017: Replaced optical frame to address the temperature dependence and replaced the main wire harness to address the unknown source of the voltage spikes. Changes made increased the dependability of the measurements.

Brewer #145

Double monochromator, multi-board electronics Brewer c/w wire connected micrometer design.

- November 2013: Change in FW2 position from position 1 to 0 to increase standard lamp intensity. The increased counts had no effect on SL ratios or atmospheric measurements but did improve diagnostics test results.
- November 2014: Top micrometer jammed. Top micrometer shaft replaced, micrometer spindle serviced all micrometer wire attachments unchanged. Intermittent micrometer positioning issues causing inaccurate wavelength selection addressed, with minimal disruption to the instrument.
- August 2015: Installed a rotary connector into Brewer #145's tracker. Data lost due to cable damage and azimuth tracker failures resolved.
- April 2016: Mistakenly, it was thought that the Cal Step was accidentally left at 962 at October 2015 calibration at MLO. Sun scans seen in April indicated that the Cal Step should have been 969, as did the processed MLO data. The 969 Cal step was applied. Later it was discovered that the slit mask was manufactured such that the offset in Cal step was both intentional and necessary. Adjustments were made to the calibration for processing data collected in this configuration.
- August 2017: Mercury bulb test repeats were increasing. Counts were too high at 1M. Changed FW2 to Pos1 in the configuration file for HG measurements, to reduce the counts. Now about 300K counts at the peak. This reduced test failures, increased the bulb life and allowed more time for atmospheric measurements.
- September 2017 Micrometer 1 jammed, and found to be stiff. Top micrometer replaced. Wire connections left undisturbed. Intermittent micrometer positioning issues causing inaccurate wavelength selection addressed, with minimal disruption to the instrument.
- September 2018: Instrument moved temporarily to Egbert, Canada (44.230° N, -79.780° W) in preparation for construction work proposed on the building in Toronto.
- August 2019: Fully replaced the top and bottom micrometers including all the connecting wires of the wire micrometer system. This preventative maintenance work was done in preparation of the absolute calibration in October and to ensure reliable micrometer movement in both micrometers in future measurements.

Brewer #187

Double monochromator, multi-board electronics Brewer c/w pushrod micrometer design.

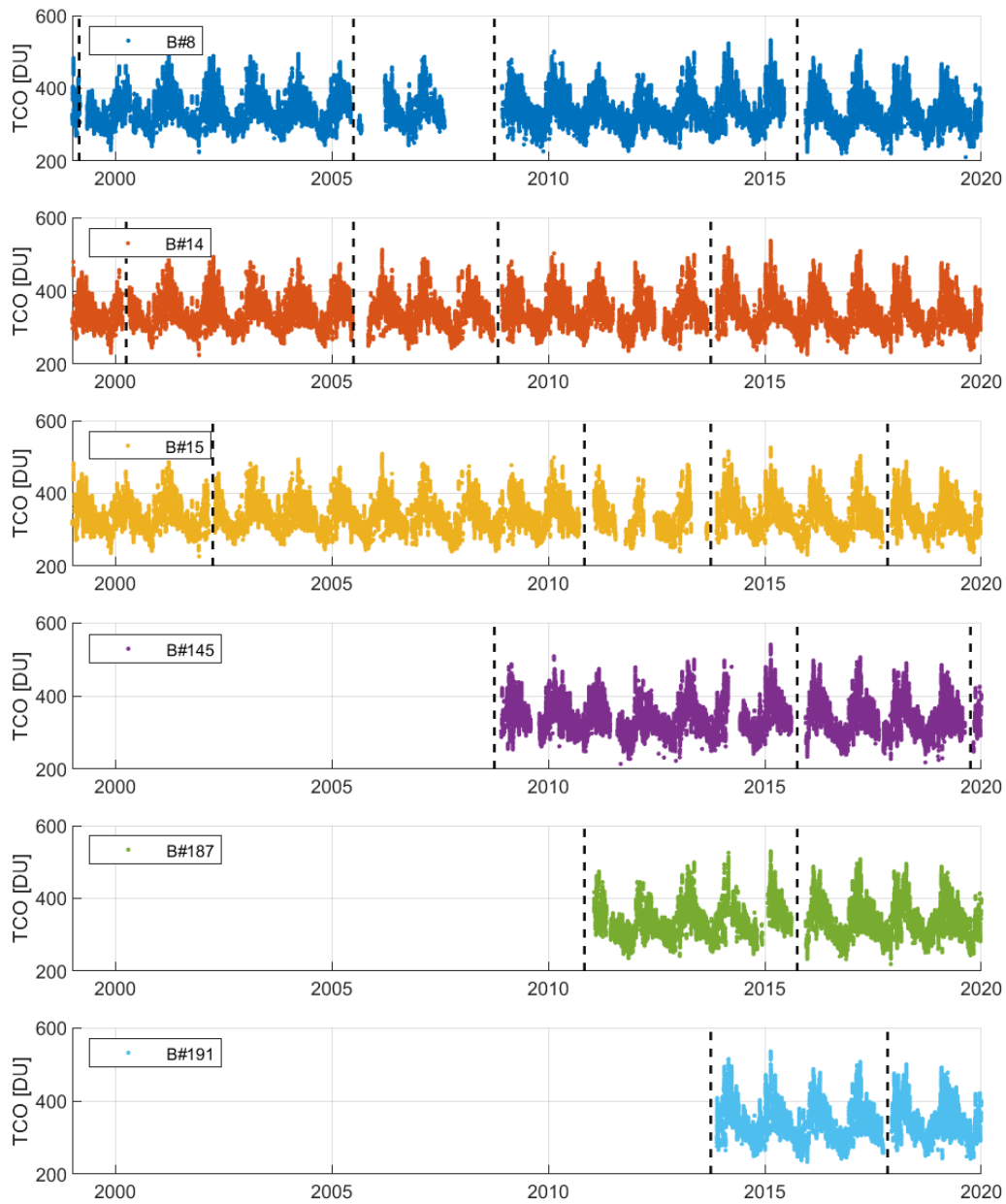
- 195 • April 2013: Multiple mercury test failures even though bulb counts were good; service micrometers with no improvement. Intermittent positioning errors would affect data quality.
- April 2013: Increased the voltage from the primary power supply to aid power to the micrometer motors. Intermittent positioning errors would affect data quality.
- July 2013: Changed the mercury bulb in an attempt to minimize data loss due to mercury test failures. Failures reduce the amount of data collected.
- 200 • November 2013: Reporting micrometer#2 jammed at maximum end - cycled power with no change - checked micrometer, not jammed. Suspected main board; board configuration or power problems.
- November 2013: Micrometer #2 issues still persisted even after firmware and configuration file was reloaded - new 110W PS installed. Data collection interrupted while problems persisted.
- 205 • November 2013: Micrometer #2 grating arm position adjusted at grating mount to allow Micro#2 to follow Micrometer #1 without crashing at the high wavelength end of the travel. Adjusted minimum motor position in configuration file for Micrometer #2 motor. The original instrument setup reached physical ends during some measurements. Adjustments made to avoid physical limitations of the instrument.
- January 2014: Changed Micrometer #2 minimum boundary from -9546 to -9808. The configuration needed adjustment to reflect the physical changes made to the instrument.
- 210 • February: Testing for instrument dark count improvement done. Returned discriminator value back to 6.7 mV and HV (High Voltage) back to 1132 V. #187 Photomultiplier (PMT) noisy - DC presently at 2200. The signal to noise ratio is acceptable but this PMT is poor when compared with others within the network.
- 215 • March 2014: Micrometer jammed. Measurements interrupted until operator intervention.
- September 2014: Brewer indoors for PMT replacement.
- November 2014: Top micrometer having trouble moving - Replaced main board - bottom micrometer appeared to be moving well but some inconsistencies showing up. Intermittent positioning errors would affect data quality.
- 220 • December 2014: Micrometers continuing to fail. Converting to multi-board electronics to decipher if the problems originated from board control or from motor quality.
- January 2015: Brewer #187 optics now running with Brewer #020 electronics.
- January 2015: Found PMT Shield Ground wire broken - replaced wire.
- 225 • January 2015: Micrometer #1 jammed at the minimum end, caused by building power outages and poor recovery by the instrument. Measurements interrupted until micrometer was repositioned by hand and the schedule restarted.
- January 2015: Testing with #020 PMT installed.
- March 2015: Found uneven torsion of vertical support wires on bottom grating - released stress reassembled and established new slope and intercept values.
- 230 • August 2015: Replaced all pins on photon counter DB - found 2 broken but held in place by silicone. Changed the SL filter position from 1 to 0 and the HG position from 2 to 1. Packed up for MLO.
- February 2017: SL Bulb replacement required. All offsets affected by the change in SL reference position were addressed. Zenith offset from 20 to 32, and changed horizon correction from -5 to 7 and UV offset from 2226 to 2238.
- 235 • September 2018: Instrument moved temporarily to Egbert, Canada in preparation for construction work proposed on the building in Toronto.
- November 2019: During UV Data review, a noted intensity drop was seen after the instrument transport to Egbert, Canada. The most likely cause is the bottom mirror positioning screw is

240 loose and the mirror tilted out of position during transport. Relative measurements (Ozone) will not be affected, but absolute measurements (UV) will be affected. Yet to be addressed.

Brewer #191

Double monochromator, single-board electronics Brewer c/w pushrod micrometer design.

- 245 • August 2013: Brought inside for pre-MLO check-up. SL burned out, replaced. Standard lamp test data interpolation needed over this time period.
- September 2013: Brought in for FW2 replacement - non-uniform spacing between filters corrected. Packed for MLO. Uniform spacing between optical densities will reduce oscillations between filters and increase time for measurements.
- 250 • January 2014: Stuck in the loop while running the RL (prints the firmware error log) command - log full - restarted schedule - working with K&Z too find a solution. Received code to clear the Brewer internal log. Developed code to read, record then delete the internal Brewer logs. The development of log handling reduced the possibility of interruption caused by the log features within the single board electronics.
- 255 • February 2015: SL bulb failure - bulb changed, now using long life (2000-hour rating) bulbs. Modified configuration file to change zenith offset from 4 to 11, changed horizon correction from -32 to -38 and UV position from 2221 to 2214. Adjusted to ensure offsets reflect the standard lamp reference position change caused by the bulb replacement.
- March 2017: HG failed, b1 (turns on the HG lamp) no light, changed Hg bulb still no light. Continued to troubleshoot to find the lamp board Hg potentiometer too high. Adjusted HG 260 voltage from ~15V to ~10. Brewer mercury tests now working normally.
- August 2017: Instrument review and maintenance in preparation of calibration at MLO. Bulbs changed. Changed zenith origin from 4 to 11, and changed horizon correction from -31 to -24 and UV offset from 2214 to 2207. Adjusted to ensure offsets reflect the standard lamp reference position change caused by the bulb replacement.
- 265 • September 2018: Instrument moved temporarily to Egbert, Canada in preparation for construction work proposed on the building in Toronto.



270 **Figure S1. Timeseries of Brewer TCO observations in Toronto. Vertical black dash lines indicate the time of primary calibrations as shown in Table 2.**

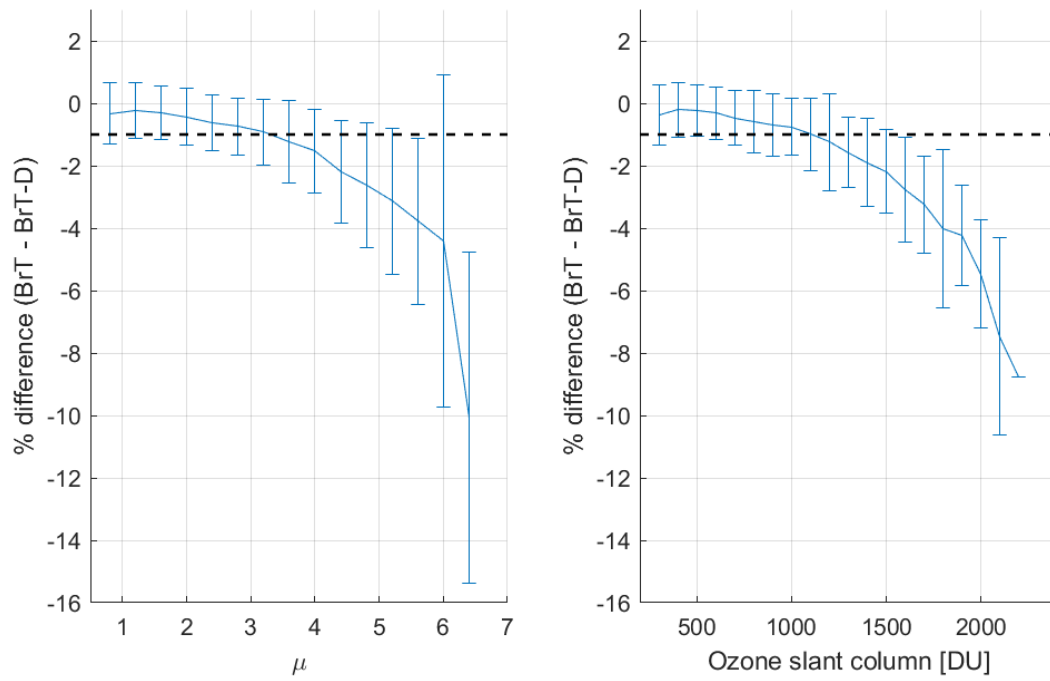


Figure S2. The relative difference between BrT and BrT-D, in terms of air mass factor (μ) and slant column ozone. The error bars represent 1σ of the relative difference values. The black dash lines show the -1 % relative difference.