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 ECCC, 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.

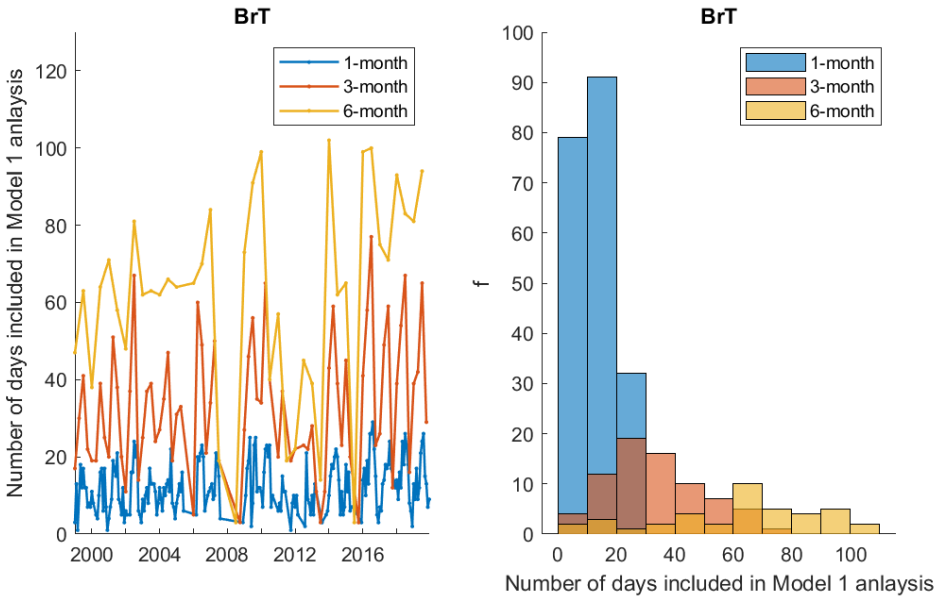

Figure R1. The number of days included in Model 1 analysis for Brewer reference Triad (BrT).

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.

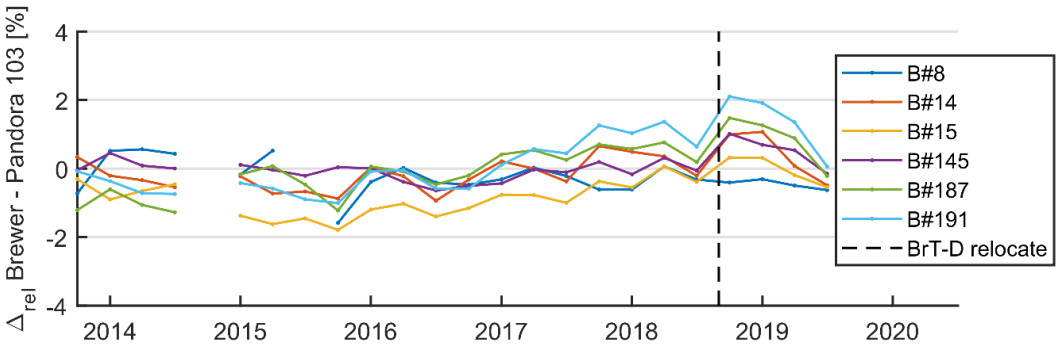

Figure 3. 3-month relative differences between Brewers and Pandora total column ozone. 3-month averages are calculated if there are at least ten coincident measurements between Brewer and Pandora for that period. The black dash line represents the time when BrT-D was relocated to Egbert, i.e., Pandora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

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

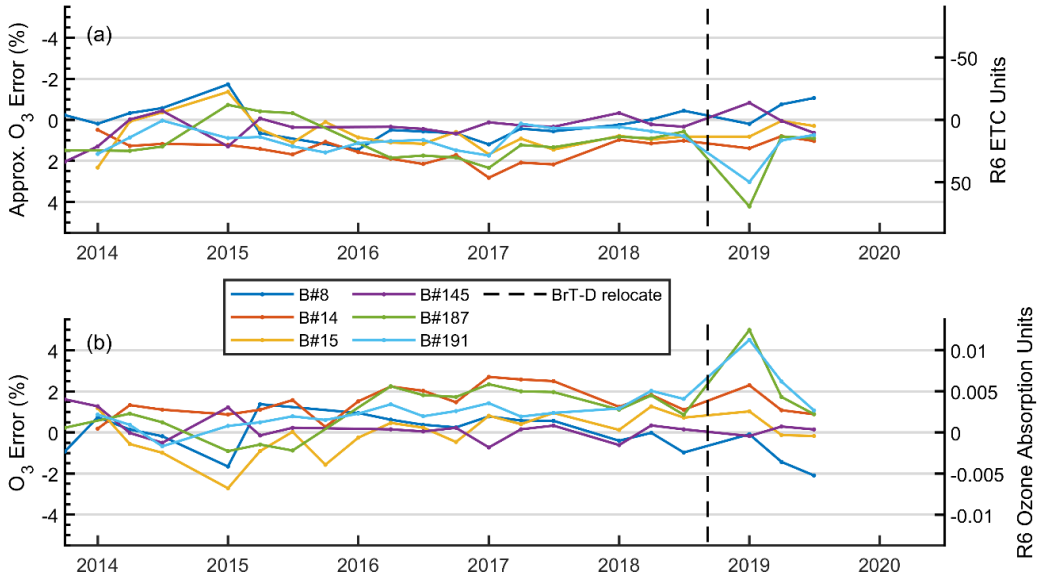

Figure 4. Relative systematic uncertainties in ETCs and effective ozone absorption coefficients estimated using Model 3. Description of y-axes is in Fig. 2. Each point on the graph represents a 3-month average. The black dash line represents the time when BrT-D was relocated to Egbert.

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.

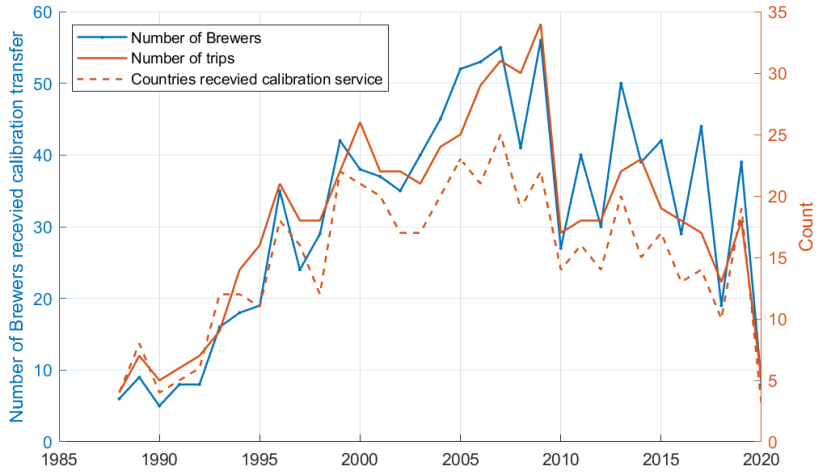

Figure R2. Time series of the calibration transfers done by IOS from 1988 to 2020.

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 (μ= 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.

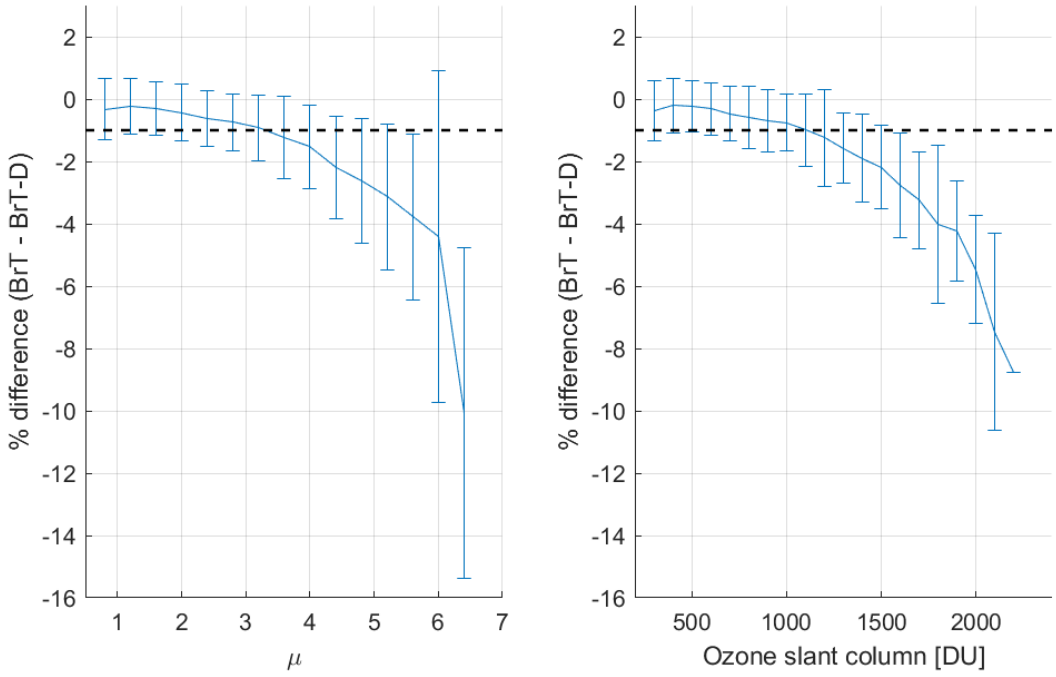

Figure R3. The relative difference between BrT and BrT-D, in terms of air mass factor (μ) and slant column ozone. The error bars represent 1σ of the relative difference values. The black dash lines show the -1 % relative difference.

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) Line181, 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 (±2° in latitude and ±20° 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 ±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).

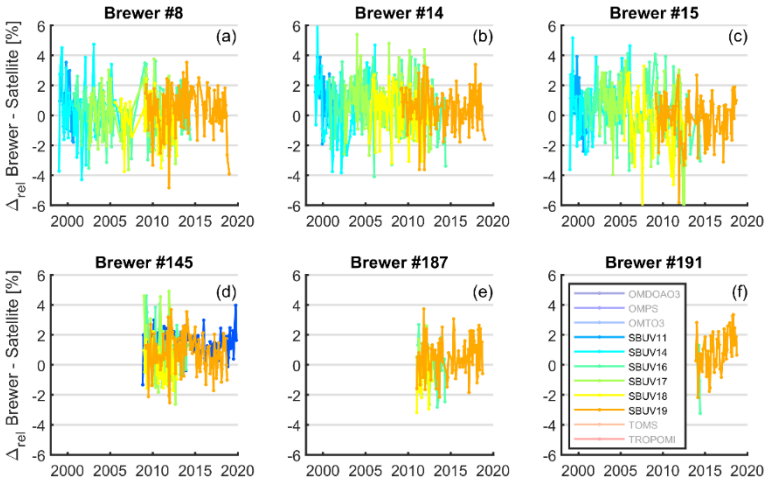

Figure R4. The relative difference between satellites and the world Brewer reference triads (BrT and BrT-D). Same as Fig. 5, but only shows SBUV satellites.

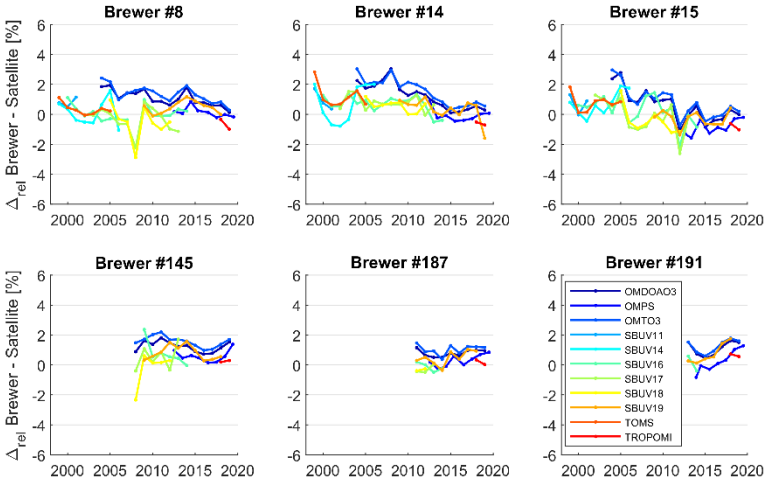

Figure R5. The yearly relative difference between satellites and the world Brewer reference triads (BrT and BrT-D).

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 μ ≤ 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 (±2° in latitude and ±20° 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 OMDOAO3 and OMTO3 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/).

The offline (OFFL v010107) total ozone column data (Garane et al., 2019) are used in this work (only L2 data with qa ≥ 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 μ = 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 μ = 2 represents the mid-point of TCO seasonal (3-month) air mass variations from 1.5 to 2.5.

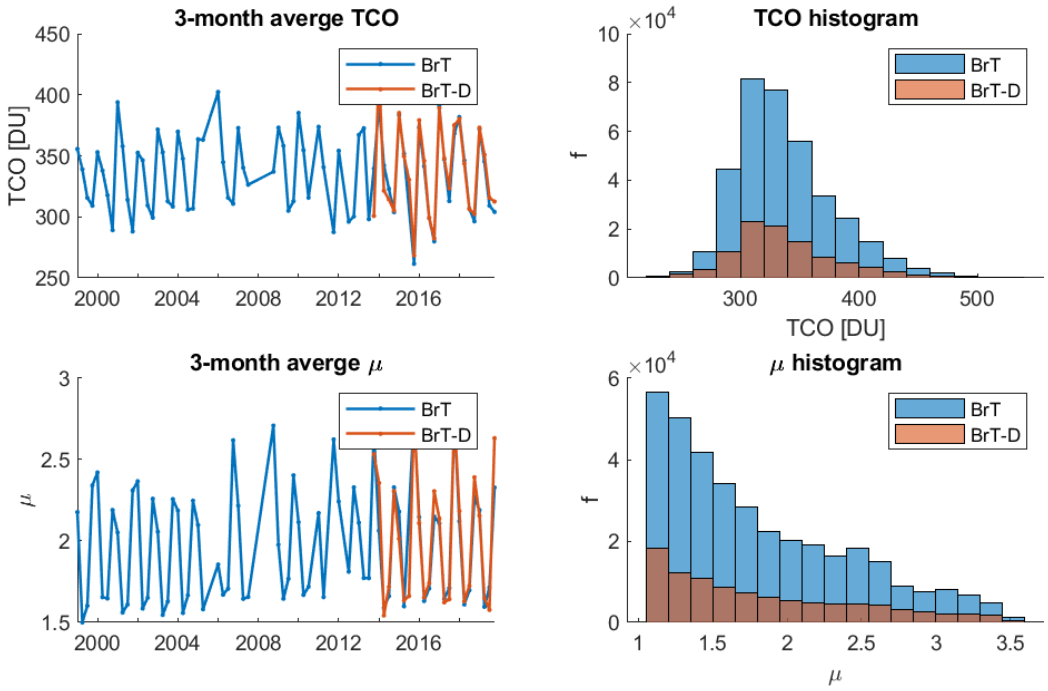

Figure R6. Time series and histogram of TCO and air mass factor (μ) in Toronto.

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 ±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 if 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 OMITO3 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 TRPOMI 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 ±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 ±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.