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).

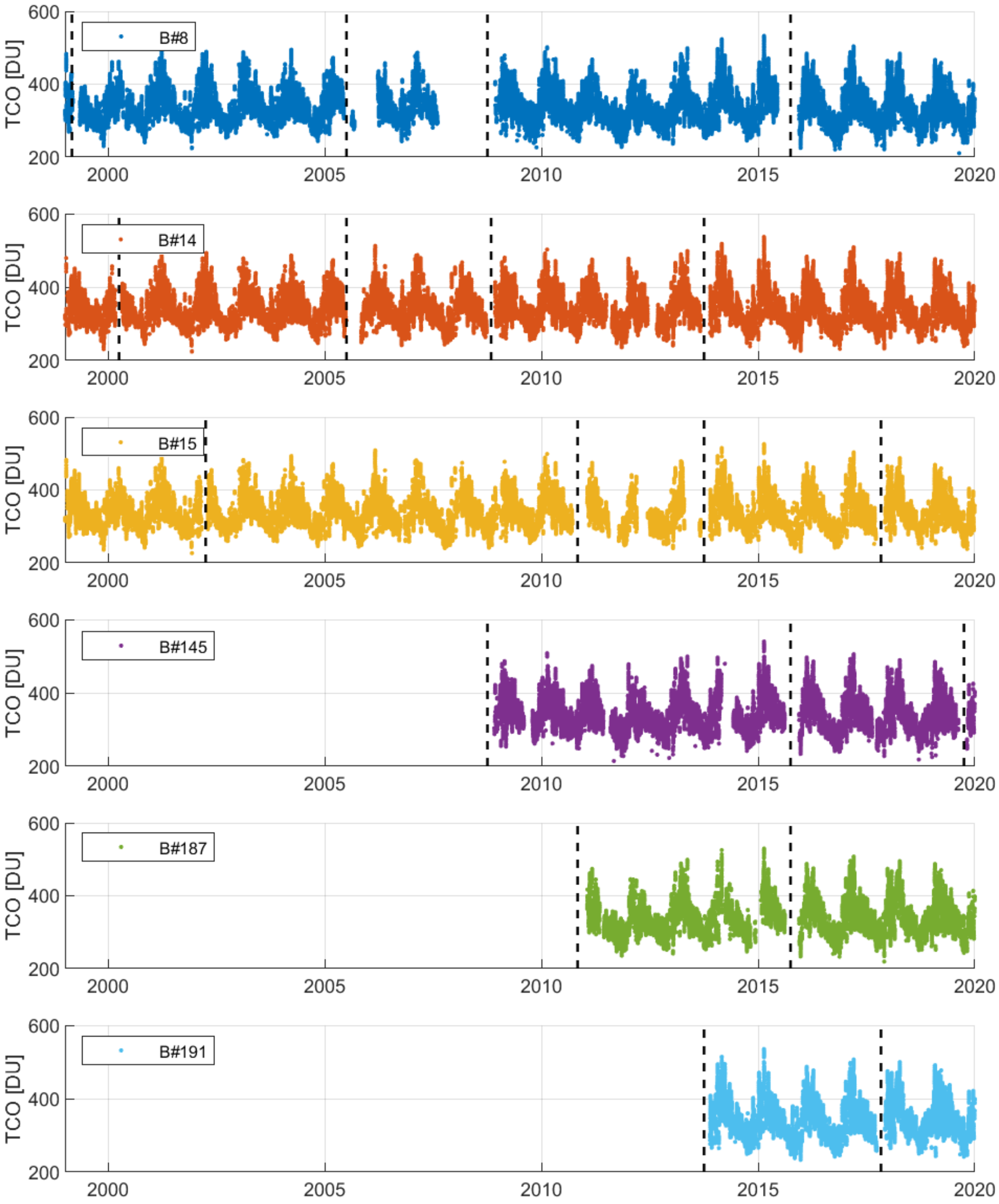

Figure S1. Time series of Brewer TCO observations in Toronto. Vertical black dash lines indicate the time of primary calibrations as shown in Table 2.

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 TCO ≤ 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.

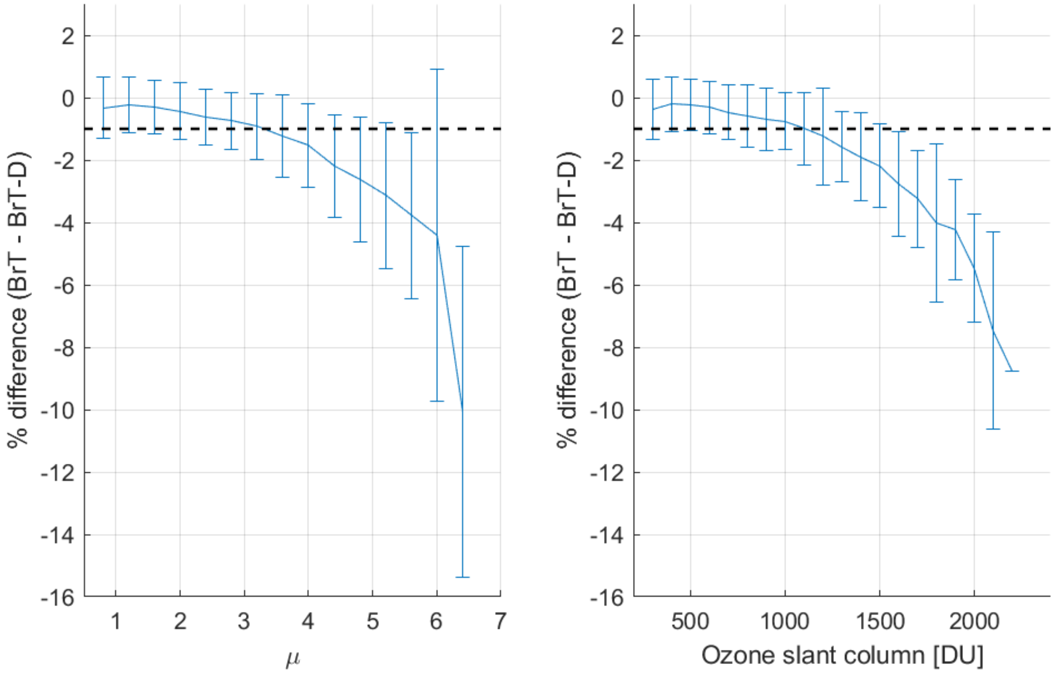

Figure R3. The relative difference between BrT and BrT-D, in terms of air mass factor (μ) and slant column ozone. The error bars represent 1σ of the relative difference values. The black dash lines show the -1 % relative difference.

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 μ > 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 (μ = 4 and slant column = 1200 DU) will be calculated as TCO = 250 DU (μ = 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.

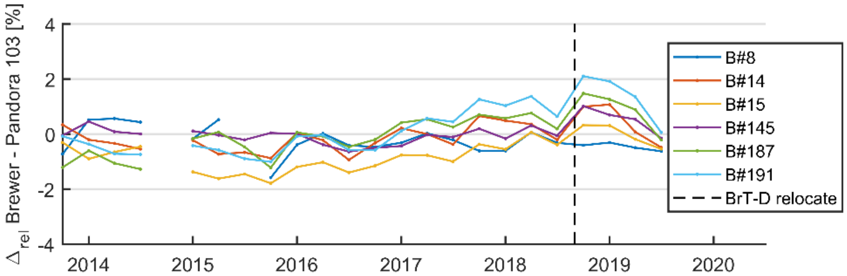

Figure 3. 3-month relative differences between Brewers and Pandora total column ozone. 3-month averages are calculated if there are at least ten coincident measurements between Brewer and Pandora for that period. The black dash line represents the time when BrT-D was relocated to Egbert, i.e., Pandora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

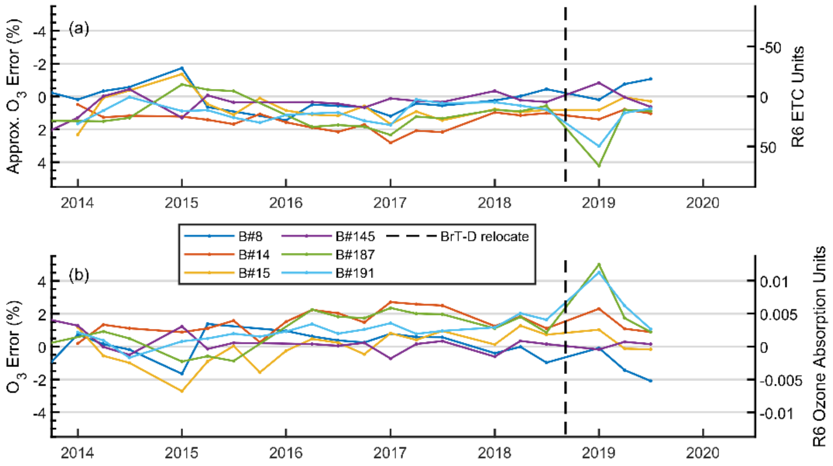

Figure 4. Relative systematic uncertainties in ETCs and effective ozone absorption coefficients estimated using Model 3. Description of y-axes is in Fig. 2. Each point on the graph represents a 3-month average. The black dash line represents the time when BrT-D was relocated to Egbert.

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 (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)).

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 (SO_2; e.g., Fioletov et al., 1998; Zerefos et al., 2017), column nitrogen dioxide (NO_2, 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 ($F + \Delta\beta m$) 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.

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 Pandonia 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.

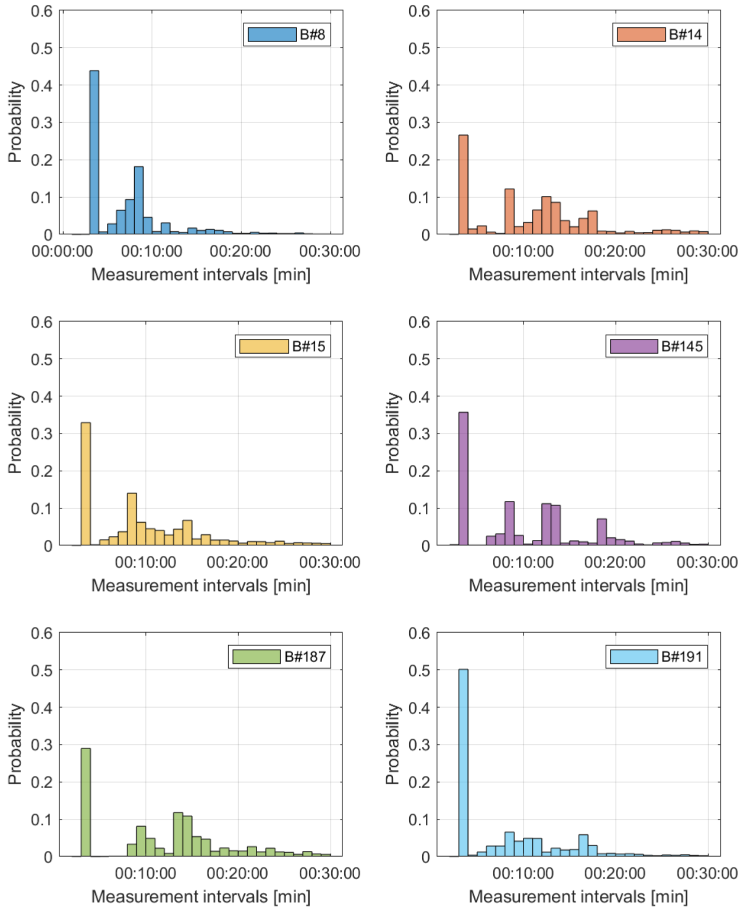

Figure R7. Probability of Brewer reference instruments' direct-sun TCO observation interval.

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.

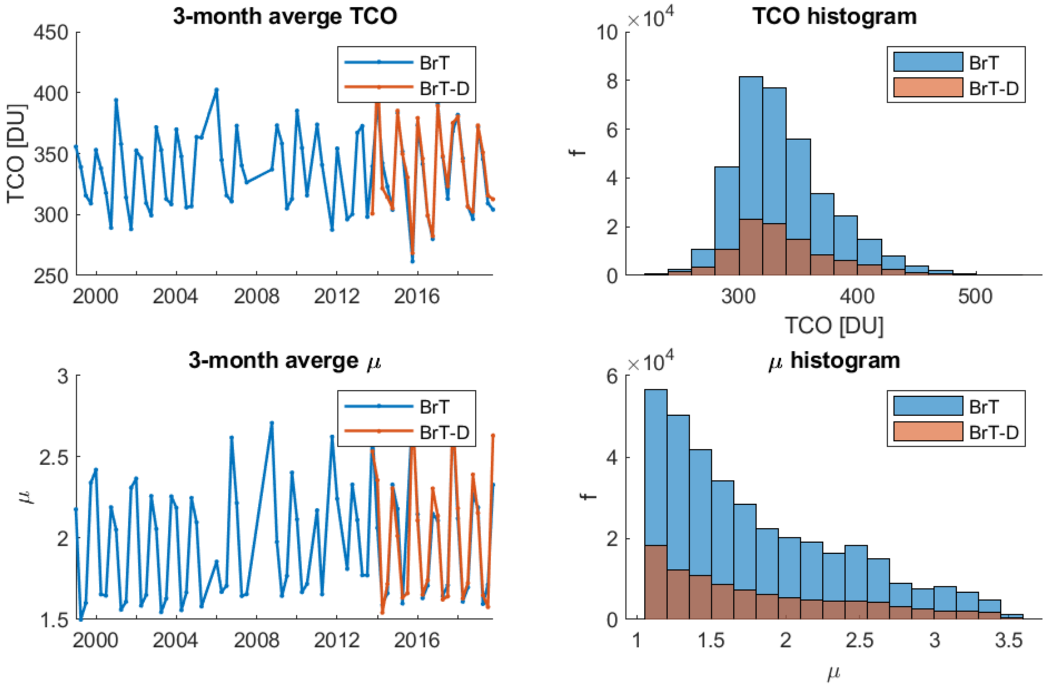

Figure R6. Time series and histogram of TCO and air mass factor (μ) in Toronto.

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 ±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×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., Ω = 330 DU, $\Delta\alpha = 0.34$, and μ = 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 μ = 1 (worst case for error), we can calculate that a 5 unit uncertainty in $ETCO_3$ 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.

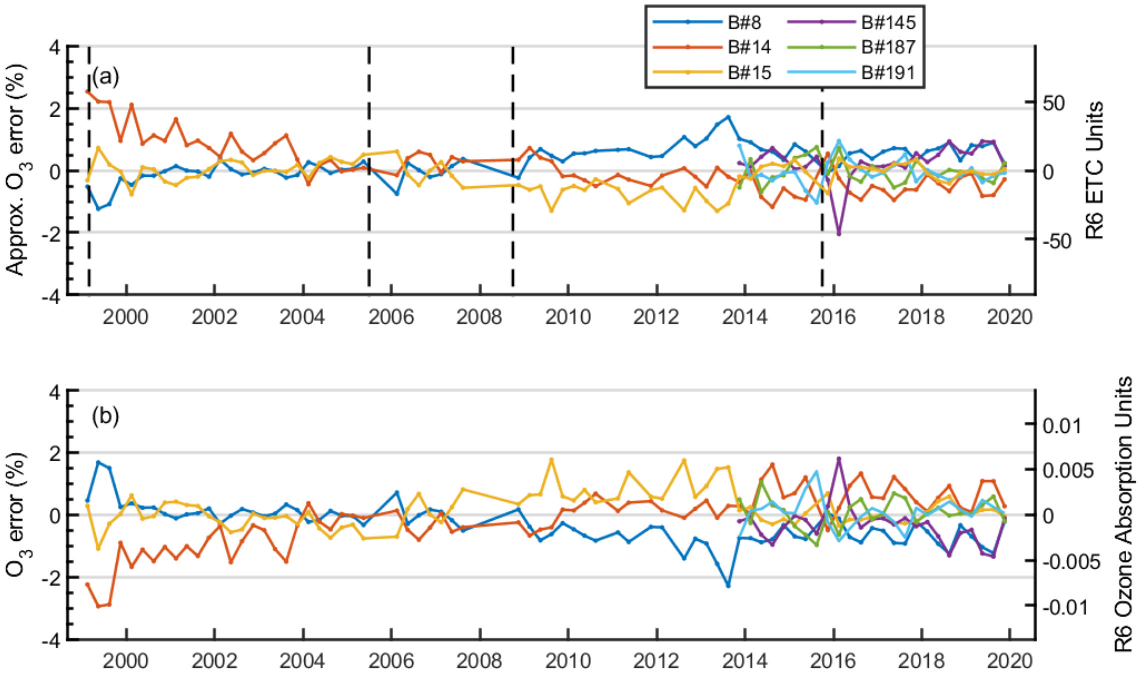

Figure R8. Modified Fig. 2, with primary calibration dates for Brewer #008 indicated on panel (a) by vertical dash lines.

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.

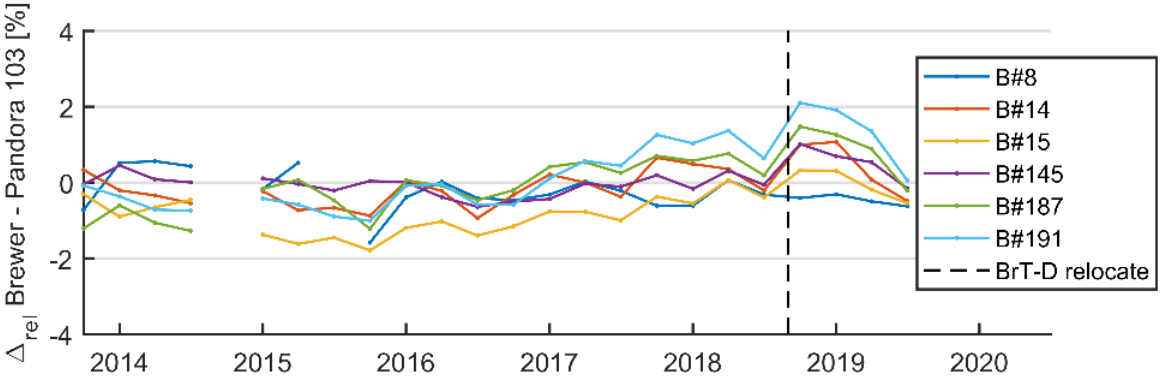

Figure 3. 3-month relative differences between Brewers and Pandora total column ozone. 3-month averages are calculated if there are at least ten coincident measurements between Brewer and Pandora for that period. The black dash line represents the time when BrT-D was relocated to Egbert, i.e., Pandora 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.

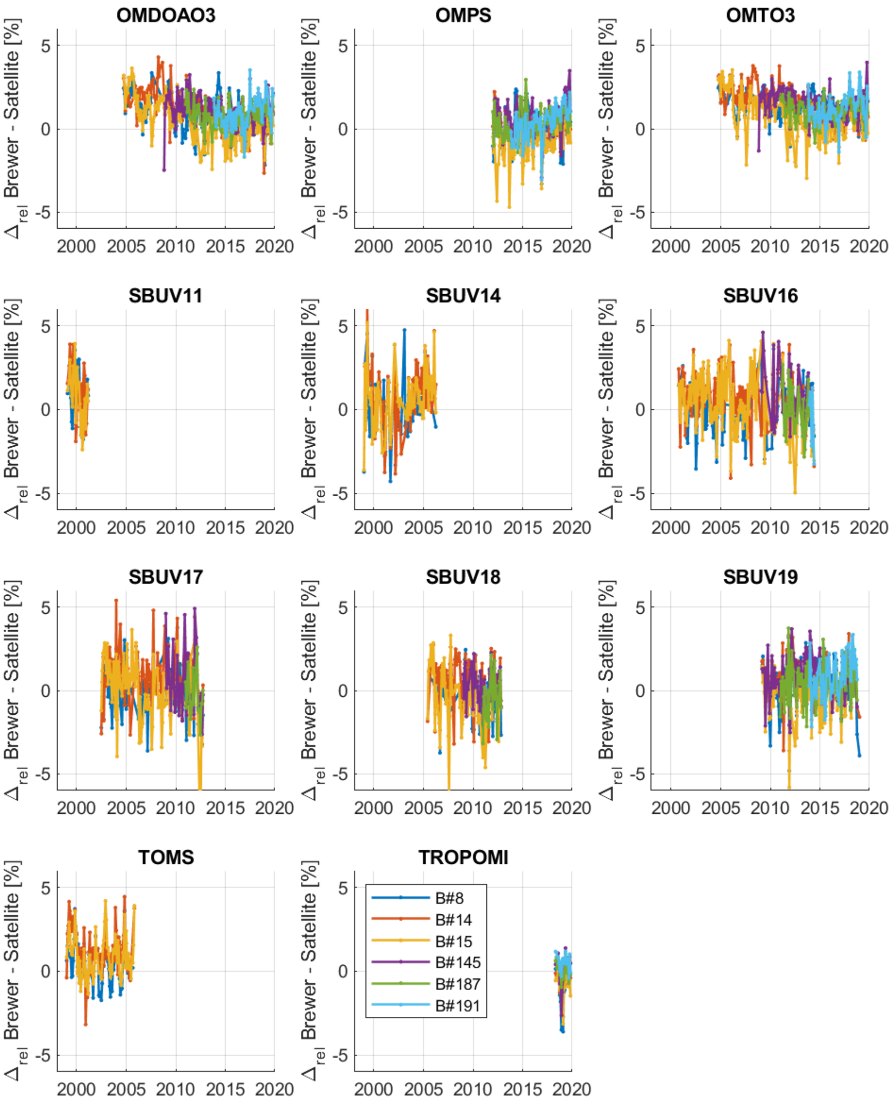

Figure R9. The relative difference between satellites and the world Brewer reference triads (BrT and BrT-D). Each point represents a 3-month average. Brewers and satellite data are paired with the criteria shown in Table 3.

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 μ = 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.*~~