Reply to Referee #2:

Many thanks to referee #2 for the positive comments on our work. We appreciate the suggestions for improvement provided by the referee. Our replies are inserted into the comments (in blue). Any references to the changes made in the manuscripts are italicized and shown in red.

Review of Zhao et al. " Measurement report: Global Total Ozone Records – part 1: ground-based monitoring networks performance assessment and status review"

General comments

The manuscript provides a very welcome survey and assessment of the current and historical global ground-based total column ozone observing system. In fact the survey itself would be worth publishing in my opinion even without the quality assessment.

Building on techniques previously used in older work by Fiolotev et al., five different instrument networks (Brewer, Dobson, Filter, FTIR, UV/Vis and Pandora) are compared to multiple satellite sensors and also reanalysis datasets. Such a comprehensive approach is very strongly commended.

I believe the work to be a valuable contribution and worth of publication once some issues below have been considered, which are mostly to do with the presentation and fairly easily addressed.

Specific comments

My first comment is a purely pragmatic one but is crucial. Many readers will only look at figures 6 to 11, which together form the main result of the manuscript. These figures need to make much more obvious that the quality criteria have been tuned separately for each network. This needs to be made clear in the caption of each figure and also the heading for each plot. The captions at the moment say "Same as ... " but this is not actually correct.

Done. The captions for Figures 6 to 11 have been updated to include such information, such as:

Figure 6. *Heatmap of data quality for Dobson sites assessed using satellite measurements. The assessment results are colour-coded for each tile and labelled on the legend; assessment criteria are determined based on Dobson network records (which only reflect Dobson networks' internal performance). The stations' WOUDC ID number and name are*

printed on x-axis ticks. Sites are sorted by latitude from north to south. Vertical dashed lines separate the sites into four regions: Arctic, Northern Hemisphere, Southern Hemisphere, and Antarctic.

Figure 8. *Same as Figure 6, but the results are for the Brewer network and assessed using satellite measurements. Note that the assessment criteria are determined based on Brewer network records (which only reflect Brewer networks' internal performance).*

Secondly, in general the text needs to include more explanation. Several points need more explicit treatment, and some details are skipped over. (I felt when reading the manuscript that the authors seemed to be in too much of a rush to get to the results).

The whole approach relies on assumptions which are never clearly stated. Satellite total ozone datasets have been extensively validated with ground observations and are frequently reprocessed with corrections which may be latitude or zenith angle dependent. They are also extensively compared with each other. At times though the discussion here implies that satellite observations can serve as independent reference data for ground measurements. Presumably the assumption is along the lines that departures from the general pattern of relation between stations and a given instrument indicate a problem with the station, but changes in the general pattern indicate a problem with the satellite? (As is evident in Figure 2 for example). It is clear from figures such as Figure 2 and Figure A1 that the satellites and reanalyses only agree with each other to within a broad band of roughly plus or minus 2%. This must put a limit on your approach.

We fully agree with the referee that we need to better clarify the method. As described in Section 3.1 (Initial quality control), we identified the limitations of the satellites and reanalysis, such as the fact that we have to remove OMDOAS records from the assessment work. The method we employed is not because we have better trust in satellites or reanalysis than in ground-based observations. In fact, we trust the ground-based networks more. However, we have the assumption that the entire satellite data group and the reanalysis data group can be trusted to a certain level, so that we can use them as benchmarks to identify outliers within the ground-based observations.

In general, to make this happen, we can only identify the sites that have worse performance than those benchmarks. We have included such information in Section 2.4.

The obtained ΔTCOs were then analyzed to determine the criteria that have been used in the statistical model. Other technical details of the statistical criteria can be found in Fioletov et al. (1999). This analysis assumes that satellite records and reanalysis data are accurate and precise enough to help identify problematic ground-based records.

Conversely, a change in the overall pattern of differences between a satellite/reanalysis data set and ground-based observations may indicate an issue with that set.

The value of reanalysis data also is not made out (other than just convenience). The authors need to explain what is the benefit of reanalysis when the satellite overpasses are available from the satellites that have been assimilated.

I am especially unconvinced about the validity of using reanalysis prior to reliable satellite total ozone data. The onus is on the authors to explain why this is valid.

This is not the first time we have tested using reanalysis data to facilitate ground-based observations (e.g., we tested using MERRA-2 to simulate calibration conditions for Brewer, which shows good agreement with observation records (Zhao et al., 2023)). Note that in Fioletov et al. (2008), only satellite records are used as the comparison targets. So, this work is also a proof of concept that modern reanalysis data is already good enough for the site survey work. We are glad that the referee agrees with us that using reanalysis shows similar results to when we use satellite.

We have to clarify that the convenience of using reanalysis includes: easy extraction of records from any location, no need to collect 20 satellite records, and no need to perform necessary quality control for each of the satellites (i.e., those 20 satellites are not sharing the same pixel size and/or sampling areas, we have to make sensitivity test to use a uniform criteria to "harmonize" them).

Regarding the results for the pre-satellite era, we have the same concerns. Without assimilation sources, it is very clear that many reanalyses show larger differences compared to Dobson networks. Given that those reanalyses records are all publicly available, we believe this is the first work that also sheds some light on such issues. More details about the performance of reanalysis models will be the topic for Part II of this work (the next paper).

We have included some justification and benefits of using reanalysis data in the manuscripts.

A detailed site-by-site and year-by-year review is strongly recommended when using these pre-satellite era records for ozone trend analysis. Lastly, this is the first time reanalysis data have been used for the ground-based network performance survey. Utilizing reanalysis data in such work reduces the impacts of potential issues with individual satellite records. Besides, reanalysis records can be easily extracted for any location and date. However, note that the quality of the reanalysis records is not always uniform, as there are clear quality changes before and after the satellite total ozone observations became available in the 1970s.

Many of the plots are too cluttered and it is difficult to discern a clear message from them (Figures 1 and 2 in particular are very cluttered and figure 5 could be improved).

Unfortunately, we have too many stations to be included in Figure 1 (the map). For example, just in the 2020-2024 period, there are 342 stations from those five networks. The purpose of the figure is more of an illustration of the distribution of each network. For example, the Filter network provides some critical coverage in Eurasia, whereas the Dobson and Brewer networks truly provide better global coverage.

Similarly, Figure 2 (Dobson and Brewer records vs. satellites) has 20 satellites and four reanalysis records (24 lines). However, this figure is also more of an illustration of datasets than being included in this work. More detailed results on the performance of each individual satellite and reanalysis will be provided in Part II of this work.

Figure 5 has three goals. It shows the difference between using network-dependent and uniform statistical criteria (a and c). It also shows the performance of using reanalysis as the benchmark (b). Last, panel c is needed to show the changes in the network's capacities. We fully understand the figure is very busy, and we could cut it into two figures. However, we believe this information is needed to support our arguments. Also, we already have 11 figures in the manuscript and six figures in the appendix; we would prefer not to split this figure into two.

We have updated Figure 1 to separate the reference networks and other networks. For more details, see our replies in the later part of the answers.

The details of the quality tests are referred to but never explained systematically. In particular there is only scattered comments about what the tests are really looking for and how they are assessed (eg lines 88-89). Taking for example, the amplitude of the seasonal component of the difference, I would expect that a given instrument type at a certain latitude compared to a given satellite instrument would have an expected seasonal cycle of differences due to differences in the respective retrievals. You need to explain clearly how you then turn this into a quality check. I would suggest that, at least in the supplement, you plot an example of each of these tests to help the reader understand what their purpose is.

The details of these five statistical criteria were included in the previous work by Fitoletov et al. (1999 and 2008), to avoid repeating information, those details were not included in the previous manuscript. Given the concern from the referee, we have included more

description in Appendix D for the seasonal difference fitting and provided one example. The other four criteria were simple statistical calculations, and the results were included in Appendix A.

The obtained ΔTCOs were then analyzed to determine the criteria that have been used in the statistical model. For the daily mean and daily standard deviation criteria, the daily data points within a 5-year time bin must be larger than 100; for monthly mean criteria, the minimum requirement is 15 months; for annual range and amplitude, the minimum requirement is two years. These limits are to ensure that a sufficient amount of data and sites will be included to estimate the filtering criteria. The amplitude of the seasonal difference was estimated for each station, and each bin from the best fit of monthly differences using local regression (see Appendix D). Other technical details of the statistical criteria can be found in Fioletov et al. (1999). In general, this work is done based on the assumption that the group of satellites records and reanalysis data have good enough accuracy and precision, which is better than certain problematic ground-based records.

D. ΔTCO seasonal amplitude fitting

To account for the potential changing amplitude of the seasonal cycle in ΔTCO, the data have been decomposed into linear and changing amplitude seasonal components.

$$\Omega = \beta t + \alpha(t) + \sum_{i=1}^{2} I\left(b_{1i}\, sin\left(\frac{2i\pi t}{12}\right) + b_{2i}\, cos\left(\frac{2i\pi t}{12}\right)\right) + N \qquad (D1)$$

Here, Ω is the ΔTCO amount; t represents time (in months); βt is the linear trend term (to account for potential drift between the comparison and target datasets); α is the time-dependent seasonal offset; the summation operator term is the seasonal signal with a local regression indicator (I); b_{1i} and b_{2i} are the constant coefficients to be determined from the fitting; and N is the residual. The local regression indicator (I) fits the data with a time window of ±6 months for each year (i.e., each year is fitted by its own data plus data from half a year before and half a year after). One example of this fitting for Brewer at Toronto (STN065) against OMTO3 records is shown in Figure D1. The clear seasonal structure in ΔTCO is caused by satellite records having larger sensitivities to effective ozone temperature changes than the Brewer spectrophotometer. Different network instruments have different seasonal features; thus it is critical to make sure the ΔTCO seasonal amplitude criteria used in the statistical model are derived for each network by only using their own observations.

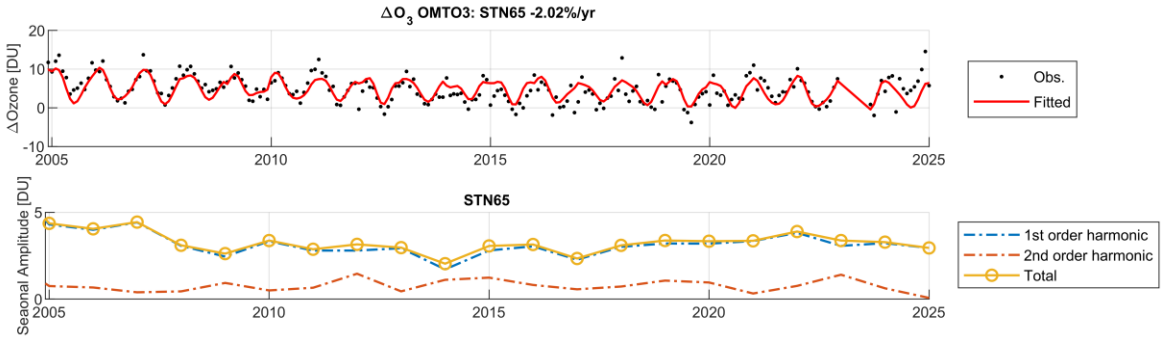


Figure D1. *Example of ΔTCO (Brewer vs. OMTO3) seasonal amplitude fitting (local regression) for Brewer spectrophotometer records in Toronto (STN065). Observations (ΔTCO) are plotted on the first panel in black dots, with the final fitting shown in a red line. Fitted first- and second-order harmonics and the combined seasonal amplitudes are plotted on the bottom panel.*

Similarly the method of combining all the satellite instruments together is not explained.

Done. More explanation has been included.

The "Satellites Mean" column is the results for using satellites' mean results as the comparison target; the "Satellites" column is the results for using each individual satellite as the comparison target and then combining their results. For example, when using each satellite as an individual comparison target, there were N = 3060 data points that were used in calculating the criteria for the daily mean of ΔTCO. When we calculate the mean of all satellites and then use it as the comparison target, we only have N = 570 data points, which is close to the number of data points when we use reanalysis as the comparison target (e.g., when using ERA5, N = 589; the small difference came from quality controls in satellite data, missing overpass records, etc.).

I appreciate that the current work is seen as an update to the older Fioletov papers but even so it needs to be reasonably self-contained and the reader shouldn't have to read the earlier work to understand the details of your approach.

(I am also mildly curious about the fact that no Chinese satellite total ozone datasets have been included?)

Done. We have updated the description of the methods as suggested (see previous replies). Regarding the Chinese satellites (e.g., EMI), we did not have access to their records.

Line 26 "global monitoring" – perhaps "systematic" is more the point you're trying to make, rather than "global"? The first sites were distributed in different places around the world, and arguably we still don't have a "global" network.

Done.

Total column ozone (TCO) has been observed since the 1920s, with ~~global~~ a systematic monitoring network established during the International Geophysical Year (1957–1958).

Lines 41-43 I found this sentence confusing, who is "they"? Do you mean users of the data or operators of the ground sites?

Done. The sentence has been modified to make this clear.

By providing station-level flags and thresholds, this assessment helps network operators prioritize calibration and reprocessing efforts and helps users identify robust records, ~~prioritizing calibration and reprocessing, which ultimately strengthens their~~ thereby improving confidence in long-term ozone trend detection and satellite validation.

Line 67 "reference network" needs defining – why exactly do you call Brewers "reference" but not Pandoras?

Done.

Multiple types of instruments contribute to the global TCO observation, each with distinct measurement principles, retrieval assumptions, and calibration practices. The World Meteorological Organization/Global Atmosphere Watch (WMO/GAW) establishes reference measurements and guidelines to ensure global consistency in atmospheric composition monitoring in key areas (such as TCO, UV radiation, greenhouse gases, aerosols, etc.) (WMO/GAW, 2017, 2026). For total ozone, ~~The~~ the Dobson and Brewer spectrophotometers are the WMO/GAW reference instruments ~~for TCO~~, ~~with~~ based on their demonstrated precision and stability when appropriately maintained (Fioletov et al., 2008; Gröbner et al., 2021; Zhao et al., 2021).

Line 69 "excellent spatial coverage and redundancy" – I can see a lot of white space in the lower panel of figure 1. How do you know that the spatial coverage and redundancy is sufficient outside of western Europe?

Done.

While the variety of monitoring instruments and networks provides ~~excellent~~ good spatial coverage and redundancy (except in regions such as Africa and part of South America), this diversity also introduces heterogeneity in measurement characteristics, calibration chains, and data processing.

Line 78 "networks have expanded or emerged" - Dobsons and Brewers have decreased though.

Done. In fact, Brewers have expanded compared to the previous assessment by Fioletov in 1999 and 2008.

In addition, ~~the~~ some ground-based networks have expanded or emerged (e.g., Brewer and PGN).

Lines 78-79 This sentence doesn't make sense to me – how can the networks "offer a more comprehensive" evaluation of themselves?

Done. The paragraph has been modified.

In addition, some ground-based networks have expanded or emerged (e.g., Brewer and PGN). Furthermore, reanalyses methods have matured and expanded in temporal scales, as seen in datasets like ERA5, MERRA-2, JRA-3Q, and MSR2 (van der A et al., 2015; Hersbach et al., 2020; Kosaka et al., 2024; Wargan et al., 2017). ~~In addition, some ground-based networks have expanded or emerged (e.g., Brewer and PGN).~~ These developments offer a possibility to perform a more comprehensive and updated evaluation, extending into the pre-satellite era, but also introducing new challenges.

Lines 103-104 This sentence seems over-stated. "Quality assured" and "traceable" have specific meanings in measurement which don't apply here.

Done.

Importantly, by documenting both strengths and limitations across networks and time, this work supports the continued integrity of the global ozone monitoring system and helps ensure that research about long-term ozone changes is based on well-characterized and good-quality ~~-assured, and traceable~~ observations.

Figure 1 – This figure is much too cluttered and I would say needs a complete redesign. It's very hard to see anything clearly or draw any sort of message from it. I leave this to the authors, but perhaps one suggestion would be to include six panels for the six instrument types.

Done. The map is now plotted for reference networks and other networks individually.

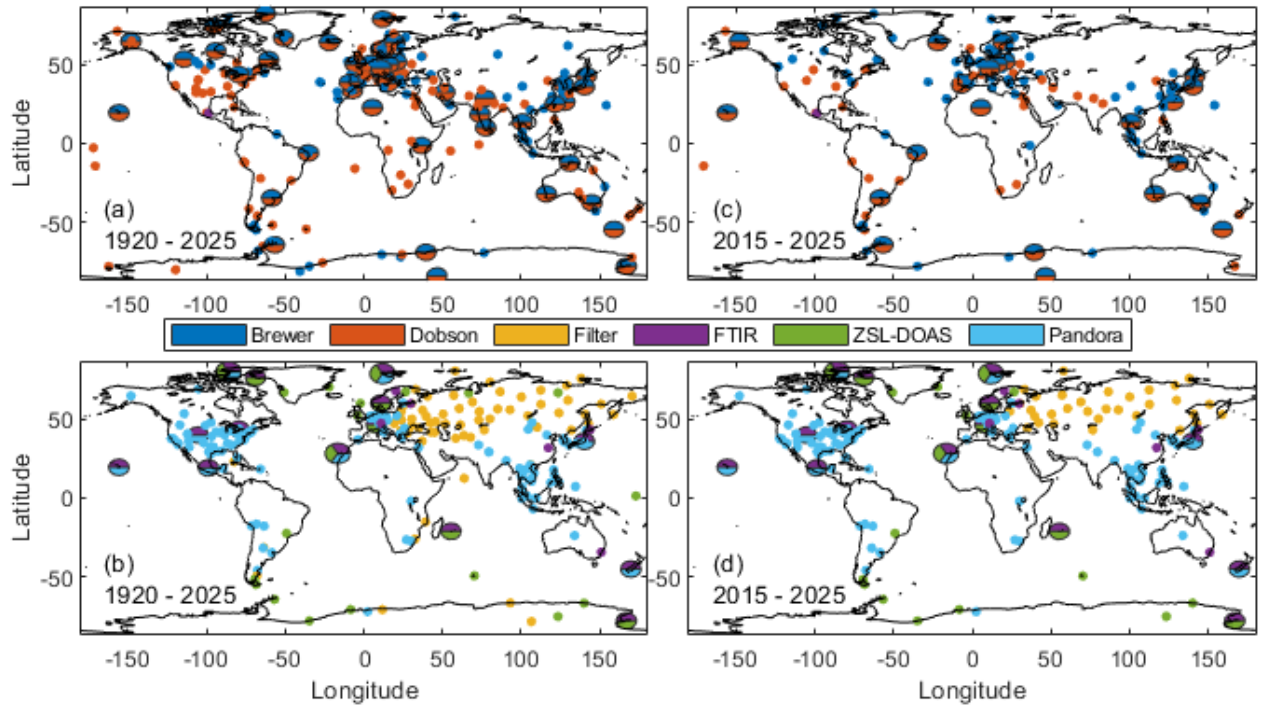


Figure 1. *Map of ground-based TCO observation networks, the panels a and b show all sites that have ever (1940 to 2024) submitted data to WOUDC, NDACC, EUBrewnet, or PGN, the panels c and d show the sites that have submitted data in the past decade. The panels a and c show Dobson and Brewer sites, while panels b and d show sites equipped with other instruments. The larger pies correspond to sites with multiple instruments.*

Line 130 Why do you say "one of "? What would beat it?

Done.

The Dobson network has the longest global operation, and provides ~~one of~~ the most continuous and reliable records of atmospheric ozone (Dobson, 1968).

Line 132 I suggest adding some words to this sentence making clearer that the stations are frequently calibrated by the regional standard instruments and not necessarily the world standard. The later wording in lines 141-142 is about Brewers.

Done.

This standard is used to define instrumental constants for all Dobson observational sites (every 4 years for Regional standards and every 6 years for stations; regional standards can then be used to calibrate station instruments), thus ensuring a single, coherent observational scale across the entire network.

As the WMO/GAW TCO reference instruments, both the Dobson and Brewer spectrophotometer has an accuracy better than 1% and are traceable via the hierarchical calibration chain (e.g., from the world primary standard to a travelling standard instrument or regional standards, and then to the field instruments).

Line 152-154 It's great that you found this problem and it was able to be corrected.

Thanks for the comments.

Lines 155-156 This paragraph reads a little bit strangely to me. It starts talking about UV/Vis in general but then some descriptions of the SAOZ specifically looks like it has been pasted in which is not totally relevant. For example, the abbreviation "CDP" isn't used again .

Done. SAOZ is a subnetwork within the NDACC UVVIS network. Unlike most research-grade UVVIS instruments (developed by institutions or universities), the SAOZ subnetwork has centralized data processing, and the instrument itself is commercially available. We have included a bit more information.

All UVVIS sites follow the NDACC UVVIS total ozone retrieval recommendation (Hendrick et al., 2011), but are mostly processed by the individual PI of the site. Within the current NDACC UVVIS network, the SAOZ instrument comprises a large portion of it (as a subnetwork), especially in the polar regions (Pazmiño et al., 2023; Sarkissian et al., 1997).

Line 169 Is this figure for SAOZ specifically? I believe the SAOZ instrument has changed over the years (mini-SAOZ).

We think the referee is talking about the 5% accuracy of total ozone. This accuracy is not just for SAOZ but for all UVVIS instruments. The sentence has been modified to make this clear.

However, as the cost, UVVIS instruments rely on radiative transfer modelling to calculate the effective air mass factor (e.g., in Hendrick et al., 2011, using TOMS v8 climatology ozone and temperature profiles).~~;~~ For example, the combined uncertainty in UVVIS scattering sunlight ozone air mass factor is 3.6%, whereas the uncertainty in direct-sun ozone air mass factor is less than 0.5% when SZA<80°. ~~the~~ The typical accuracy for UVVIS total ozone is on the level of 5%.

Lines 177-178 "As expected" – is the implication then that instruments measuring in the UV are better?

Yes, we think the existing UV (Huggins band) results have better accuracy, given the stronger features in ozone absorption and high signal-to-noise ratio. For example, the accuracy of FTIR ozone is estimated to be at a 3% level (Björklund et al., 2024), where the accuracy of Brewer ozone is estimated to be at the 1% level (Siani et al., 2018; Zhao et al., 2021). However, FTIR measurements have their own advantages (such as they can retrieve various species and even vertical profiles).

Lines 156, 171, 185 What is the message you are wanting to convey by listing out the different species able to be measured? Is it that only the Dobson and Brewer were specifically designed to measure ozone? Does the order of the lists matter, meaning does it relate to the priority of the design?

We agree with the referee on this point. Dobson and Brewer were designed for total ozone observations and are the WMO/GAW reference instruments. Unlike Dobson and Brewer, other networks (UVVIS, FTIR, and Pandora) are not specifically designed for total ozone monitoring. So, we have to give the reader the general background information to avoid misinterpretation (such as, since Dobson and Brewer are great, why does anyone else need other types of instruments). We believe the total ozone observations from other networks are still a very welcome complement to the global ozone monitoring efforts.

Line 189 Where does the 1% figure come from?

Done. A citation is provided.

In this work, the latest official retrieval version of PGN total ozone (rout2p1-8) has been used, which was found to have a 1% level agreement with Brewer (Zhao et al., 2025).

Lines 190-191 "Upcoming versions ... will address" - it would be more factual to say "are planned to " rather than "will", or "planned upcoming versions" or words to that effect – this all might never happen.

Done.

Upcoming versions of the PGN total ozone retrieval ~~will~~ are planned to address both issues.

Lines 193-201 You need to give details of the overpass criteria that you've used. Are they consistent for all of these numerous instruments? You don't say anything here about drifts or about changes in crossing times.

The overpass criteria were provided in Section 2.4, introduced together with the coincident criteria for reanalysis. This work is based on daily total ozone; thus, the satellite observations' crossing time differences are not accounted for in the analysis (note that in this work, no GEO satellite is included). This information has been updated in the manuscript.

For satellites, the coincident overpass criterion is selected to be <300 km (closest observation to the station). Note that this work is for daily TCO; thus, the satellite observations' overpass time differences are not accounted for in the analysis. For reanalysis data, the model grid that covers the site is selected.

Lines 212-213 I can't understand what this sentence means. It doesn't sound like the ozone value can be considered in any way reliable.

Correct. The pre-satellite period of ozone from ERA-5 is less reliable than later results. Note that the extension of ERA-5 records to the 1940s has already been done (Bell et al., 2021) and is publicly available. So, this work is the first effort to our knowledge that performed an assessment on those records using ground-based network observation. More details of the evaluation for the reanalysis will be covered in Part II of this work in future. We have updated the description to make this clear.

Before 1970 when satellite ozone measurements were not available, ERA5 ozone analysis is only indirectly influenced by observations that provide information on upper-air temperature, wind and humidity (Bell et al., 2021). These pre-satellite era records are expected to have lower accuracy compared to the results since the 1970s.

Lines 221-222 Why is this relevant?

The sentence the referee referred to is “The ozone distributions that are used in the JRA-3Q forecast model and provided to the public as the JRA-3Q ozone data were produced separately from the JRA-3Q data assimilation system (see Section 4.3 of Kosaka et al., 2024). ” We feel that we had better note these technical details for JRA-3Q ozone because the ozone treatment in JRA-3Q is very different from that in ERA5 and MERRA-2.

Lines 225 How is the bias correction done, are the satellite values corrected to match ground observations?

The reviewer is correct. The satellite data were corrected against the Dobson and Brewer measurements. The details are described by Naoe et al. (2020). The sentence has been updated:

A chemistry transport model is driven by wind data from the previous-version reanalysis from JMA, JRA-55 (Kobayashi et al., 2015) for the period starting from 1958 and from a JRA-3Q preliminary experiment before 1958, with bias-corrected satellite level-2 total ozone data (corrected against the Dobson and Brewer measurements. The details are described by Naoe et al. (2020)) nudged to the model after 1979.

Lines 226-227 Again, this is difficult to make sense of, but it sounds like the pre-satellite ozone values shouldn't be used in the way you're using them?

The pre-satellite reanalysis ozone is published by those institutions together with the satellite-era results. So, one goal of this work is also to check their performance and at least provide some information to the total ozone community on this development. The way we implemented it, i.e., using those pre-satellite reanalyses as a comparison target, is only a small part of the work. We have shown that those parts of the records have larger differences compared to the Dobson network (Figure 2). In the ground-based sites assessment, these new attempts only reflect the results in Figure 7, which extended the assessment from the 1940s to the 1970s. As described in lines 465 to 473, we emphasize the limitations of the assessment in this period. This paragraph has been updated to make motivations and results clearer.

Another feature of Fig. 7 is that more "not assured" tiles are found in the pre-satellite era. Note that the "not assured" category needs to be interpreted as some level of concerns have been identified by using the comparison data for a five-year period. Thus, the results are only as good as the accuracy the comparison data can provide and are only a general mark for that period. A detailed site-by-site and year-by-year review is strongly recommended when using these pre-satellite era records for ozone trend analysis. Lastly, this is the first time reanalysis data have been used for the ground-based network performance survey. Utilizing reanalysis data in such work reduces the impacts of potential issues with individual satellite records. Besides, reanalysis records can be easily extracted for any location and date. However, note that the quality of the reanalysis records is not always uniform, as there are clear quality changes before and after the satellite total ozone observations became available in the 1970s.

Lines 232-235 If the satellite data is corrected in this way, can it still be used to assess ground data? In the pre-1979 era the Dobson network was very sparse in many parts of the world.

Those biases in satellite records are well-known issues, and assimilation of problematic results in the reanalysis will not help generate accurate total ozone records. We believe performing corrections for satellite records before assimilation is a more logical and reasonable way. However, this approach hasn't been adopted by all reanalyses yet. One of the goals for Part II of this work (next paper) is to further evaluate the differences caused by these updates.

The ground-based assessment is based on statistical thresholds defined for each network based on its own observations. The quality of comparison targets plays a role in the results but is not the most critical one in the detection of problematic sites/records. Note that the value we evaluated is ΔTCO. I.e., even if the quality of the comparison target is low, then statistically, the problematic sites will still have larger differences than the good sites. However, there could be an exception if the comparison target's quality is so low that all ΔTCO signals have higher noise levels than the errors within ground-based observations; then the statistical method will not work. Fortunately, those modern reanalyses seem to have good enough accuracy. More detailed discussions on reanalysis data's performance will be the topic for Part II of this work.

The sparseness of the Dobson network in the pre-1979 era apparently affected MSR2 results, which assimilated those records. However, we would prefer to have a more in-

depth discussion on this topic in Part II. Section 3.1 has been updated to include some of the information.

For example, from the 1940s to the 1960s, ERA-5 and JRA3Q show up to -5 % bias compared to the Dobson network. It is worth noting that the excellent agreement (differences close to 0%) between MSR2 and the Dobson network from the 1960s to 1979 is largely attributable to the assimilation of Dobson observations into MSR2 as an additional satellite-like data source.

Lines 243-244 Are you using daily averages for all ground-based data? (It seems to say so a few lines later but it would be clearer to state it here.) Did you consider finer temporal resolutions where it was available?

Yes, we are using daily averages for all ground-based data. This was stated in the manuscript.

These calculations were done using daily total ozone values from ground-based, satellites, and reanalysis data.

This work and previous work (Fioletov et al., 2008) did not include any GEO satellites, which provide better temporal resolution for TCO. Thus, daily resolution from the ground-based network is good enough. However, in Part II of this work, we plan to investigate diurnal agreements between reanalysis and ground-based observations (which need an hourly resolution).

Line 249 This seems an important point in your analysis but it's rushed over. From what you're saying, the percentiles used later on are calculated with bad data already excluded? Doesn't this affect the use you make of the percentiles later on? (For example, being outside of the 90th percentile range is not so worrying if the 90th percentile is based on only good data). This should be explained. How did you identify the "low quality" data at this stage?

The low-quality data that were removed from the analysis are the network “reported” ones. I.e., if the network’s data has a quality flag provided, we will use it to filter out those self-claimed bad results. Note that many networks only submit quality-assured results, such as NDACC (UVVIS and FTIR). However, such as PGN, they provide all results with different quality flags. For ozone (version rout2p-18), it has an L2 quality flag as:

Column 36: L2 data quality flag for ozone, 0=assured high quality, 1=assured medium quality, 2=assured low quality, 10=not-assured high quality, 11=not-assured medium quality, 12=not-assured low quality, 20=unusable high quality, 21=unusable medium quality, 22=unusable low quality

We will remove observations with this L2 data quality flag 2, 12, and above. We have made this point clear.

Here, "Ground" is the daily mean value from reported observation data. Low-quality data, if labelled by each network's quality flag, has been excluded. For example, PGN records have an L2 data quality flag with labels such as "assured low quality". Those records will be excluded from further analysis. In comparison, the Dobson network only submits quality-assured data, and this pre-screening is not needed.

Lines 246-250 You've defined the daily difference but none of the other statistics. It's not sufficient to refer back to Fioletov et al. (1999).

Done. We have updated this paragraph to include more information and added more details about the seasonal amplitude fittings in Appendix D (see our replies to Referee no.2's previous questions).

Line 261 Define 'median difference' – do you average each station for a year and then take the median of all the different stations? Or is it the median of all station days? Or something else?

Done. It is the "average of each station for a year and then take the median of all the different stations". The analysis is done for yearly data, as stated in the first sentence.

Before using all the satellite and reanalysis data to calculate the criteria for the statistical model, we examined the yearly percentage differences between network observations and these comparison datasets. Figure 2 shows examples of these analyses (the median difference of yearly data in percent) for the Dobson and Brewer networks.

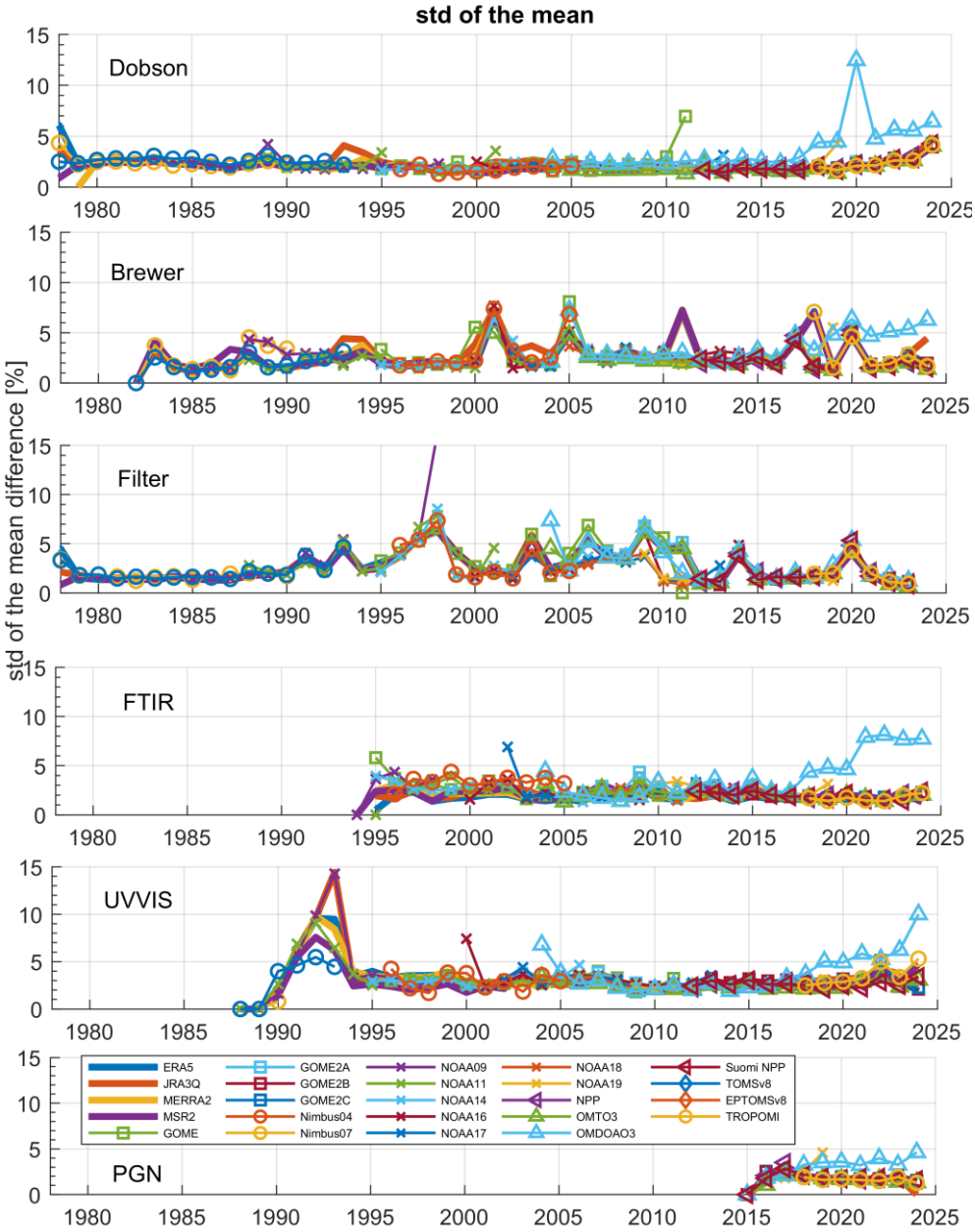
Line 269 Again, interquartile range of what exactly? Stations' daily differences?

As stated in the previous sections, the data we collected are on a daily resolution. However, this initial quality control is done based on yearly averaged data, which is mainly to identify and remove problematic comparison data (such as OMDOAS in this case).

The bottom panels of Figure 2 show the interquartile range (75th to 25th percentile) of the annual mean percentage difference in yearly averaged data (i.e., each station's daily ΔTCO will be averaged into yearly ΔTCO).

Lines 272-274 This seems like an important point but it is skipped over and hard to follow.

The bottom panels in Figure 2 show the percentile range, which helps to identify the outliers in the comparison dataset (i.e., satellites and/or reanalysis). Another criterion we can use is the standard deviation of all those yearly ΔTCO from all stations within a network, which will reflect the internal agreement within a network on yearly resolution. Below are the results for all networks:



Those spikes in Brewer's std of the mean plot are simply due to more stations submitting problematic records in those years. Similar issues can be seen in the Filter network and the early years of the UVVIS network. Given the paper already has too much information, we decided not to include extra figures but only mention the results.

Figure 2 – I find the figure much too cluttered. What is the message here? The overall agreement seems very good when completely averaged out but if the satellites have been reprocessed to match the ground networks and each other it's not so surprising.

Yes, the results came from 20 satellites and four reanalyses. The lines are all "cluttered," which is a good sign of agreement between the data. As stated in the section, the main goal of the figure is to illustrate the general agreement between ground-based and satellite/reanalysis data (i.e., Fig 2b and c) and also to find the obvious outliers in the satellite record, i.e., OMDOAO3 (i.e., Fig 2e and f).

Note that satellite data are reprocessed not to "match" with ground-based records, but to correct validation-identified issues (such as wrong surface albedo, cloud identification, or improved a priori, etc.). The good agreement between satellite and ground-based records is not guaranteed. Removing bad satellite records is a key step before we build up the trustworthy statistical criteria.

Line 283-285 Doesn't it call into question the appropriateness of using reanalysis before the modern satellites were operational?

The line 283-285 in the manuscript was:

Despite such glitches in the reanalysis, Fig. 3a still shows some important results, such as the general agreement of all networks being within a 2-3% level for most of the records, with the agreement between the two reference networks (Dobson and Brewer) mostly within 1%.

This is in the discussion of Figure 3, which shows the agreement between all networks with MERRA-2. Note that MERRA-2 only have records for the satellite era. If the referee is talking about Fig 2a, we believe that we have already answered the concerns about applying the reanalysis in the data survey in previous questions. In short, this is the first attempt to show the performance of reanalysis in that pre-satellite era. As stated in the manuscript, we are not expecting the results to be flawless. In contrast, we hope this work will shed light on the issues and give end users a heads-up if they are using those records.

Lines 286-296 A lot of this repeats what has already been said earlier (lines 175-184).

Done. The lines 175-184 are the statement of previous work, and lines 286-296 explain our results shown in Figure 3. We removed the following sentence from lines 286-296.

~~*This is thanks to the use of the HITRAN 2020 database (instead of HITRAN 2008 in the old datasets), in which ozone spectroscopic parameters in the FTIR retrieval region have been improved (Gordon et al., 2022).*~~

Lines 304-306 Again, some of the different tests are listed ("such as ... ") but none of them are ever concretely defined, or their purpose explained.

As requested by the referee, we have already addressed this in Section 2.4 and Appendix D.

Lines 338-345 This section is very good but I would like to see it expanded and some plots be included to illustrate what the tests are showing.

The statistics of those tests were included in Appendix B as Figures B1 to B4 and also summarized in Appendix C.

Line 356 I can't see that the threshold criteria were defined in section 3.2.

As we have five thresholds as criteria, Section 3.2 only provides one of them as an example. But, as stated in the manuscript, the results for the remaining four are provided in Appendix B and summarized in Appendix C.

Line 370 "models" – what do you mean here by "model"?

Done. Reanalysis data.

This feature also reflects the fact that some reanalysis data ~~models~~ have less agreement with ground-based networks than others.

Lines 388-391 When you're counting up these issues, do you count "satellites" as just one, or is each specific sensor counted?

For counting these issues, the satellites are considered as one group. I.e., for each site, we used all satellites as individual comparison targets to calculate ΔTCO to determine the general statistical criteria (this is to ensure we have a large enough number of data points to derive reliable results). In the next step, i.e., station evaluation, we averaged all good satellite records and then used them as the comparison target to calculate ΔTCO. This ΔTCO is assessed by the statistical criteria defined in the previous step.

We tested another method, i.e., using each individual satellite as a comparison target and calculating ΔTCO, which is then used to determine the quality of the data (i.e., each satellite alone will be considered as a referee). The problem is we will have 20 referees who provided assessments for different periods of time. Merging those results into one is tricky (even by using the DS method). Another test we did was to use one-vote voting, i.e., if any satellite labelled a site in a period as “suspect”, that data point would be labelled as “suspect”. This method works to a certain level, but it highly relies on the quality of all satellites (i.e., we can’t have any satellite that has any “bad” period). In short, the simplest and most robust way to use the satellites is just to average them as one comparison target. But, for the statistics filter threshold determination, it is better to treat the satellites as individual targets (this is important, especially for regions where observations are more

limited). We have included these details in the paper, including a table for the categorization used for this step.

Next, following Fioletov et al. (2008), for every 5 years, we identify a site as one with minor issues (data quality as medium) if its record in this time bin has either 1) one to three "suspect" characteristics or 2) one "outlier" with zero or one "suspect" characteristics. If a site in this time period has a greater number of "suspect" or "outlier" characteristics, it will be labelled with major issues (data quality as "not assured"). This categorization is summarized in Table 1. Note that, different from the determination of statistical criteria in Section 3.2, where we treated each individual satellite as an independent comparison target (to increase the number of data points for reliable statistics; minimize the impact of problematic satellite records), all satellite records are averaged to generate the comparison target in the step of statistical model analysis. One could use an individual satellite record as a comparison target in the statistical model analysis; however, that approach would lead to 20 satellite assessments. Reasonably merging all satellite assessments into one result is challenging, given that these assessments cover different periods of time and potentially have different qualities. Using methods such as "one-vote vote" (i.e., if any satellite labels a site in a period as "suspect", we will label that data as suspect), majority vote, or the Dempster–Shafer (DeSh) method (Hall and Llinas, 1997; Shafer, 1976) (will be introduced and discussed in Section 4.1) all have their own issues and then are avoided here.

Table 1. Data quality categorization based on the statistical characteristics

	Number of Suspects	*Number of Outliers*
No issues/good quality	*0*	*0*
Minor issues/medium quality	*1-3*	*0*
	0-1	*1*
Major issues/not assured	*any*	*≥1*
	≥4	*any*

Lines 401-403 It seems to me using NASA AVDC is a great convenience but it does leave you at their mercy if they haven't included a site or have the wrong position for it (as you mention again in line 516). Did you ask them for the missing sites? "Reprocessing" is not the right word here, you would only be extracting overpass values from gridded data.

Yes, we have communicated the issues with the NASA team; however, we did not get any updates on whether extracting some historical satellite datasets will be made available in the near future.

Figure 5 I can't say I like this figure very much at all. What most stands out to the eye is the solid bar. The different rows are too hard to compare. I suggest a redesign.

This is a redesign we made. The solid bars are the sites with minor issues. We made that solid bar most visible on purpose, as the top of the solid bars is the % number of sites that have no or only minor issues (which we considered to be good enough for most applications). The bottom of the solid bars is the % number of sites that have no issues (which is the best record of ground-based observations). We found we had typos in the description and have updated the paragraph to make the message from the figure clearer.

The bar plot is colour-coded for each network, with ~~solid~~ shaded parts (in light colours) representing the percentage of sites that have high-quality data records and no issues were detected by the statistical model. The ~~shaded~~ solid parts are stacked on top, representing the sites that have medium-quality data records for that period with some minor issues. I.e., the top of the solid bars shows the percentage of observations that have no or minor issues, which can be used for most research applications (such as satellite validation and trend analysis). The bottom of the solid bars shows the percentage of observations that have no issues, which can be used for studies which need more assured results (such as validation or verification of high-precision ground-based instruments).

Line 424 "trace" – I don't know what you mean by "trace" here? Do you mean just "date", or do you need to uncover them somehow?

Done.

On WOUDC, the early Dobson total ozone observations ~~trace~~ date back to 1924.

Line 426 IGY not IPY and 1957-1958.

Done.

The rapid increase of Dobson instruments started with the formal formation of the Dobson network in IGY1957-1958.

Line 439 I don't think you really mean "assured" here. Quality Assurance has a definition. GAW has a Quality Assurance system, eg Quality Assurance

We thank the referee for pointing this out. We believe our method is also a quality assurance process, i.e., the records from any sites that passed our exam are carefully checked; they meet the defined statistical criteria. However, we agree that this is different from the GAW Quality Assurance system. We have made this clearer. In fact, we think that, in the long-term, it will be useful to include this statistical assurance system in WOUDC.

Red tile means the model identifies some "major issues" and the data quality can not be assured by our statistical model (i.e., the end user should exercise extra caution when using that data).

Line 440 What is the advantage of using five year bins?

The use of five-year bins is to ensure each assessment (per period per site) has enough data points to draw reliable conclusions. But, theoretically, we can make the temporal resolution higher (no less than 1 year, as some statistical criteria are based on yearly values). The cost will be that we're likely to see more "gray" tiles, for which we don't have enough data points.

Line 458 Do you mean "models" here, rather than "reanalyses"?

Done.

As the ~~models~~ reanalysis may have systematic issues (e.g., see Fig. 2c, where ERA-5 and MERRA-2 show a 3% bias in recent decades), they are used as independent comparison benchmarks, and their assessments are then merged into one using the Dempster–Shafer (DeSh) method (Hall and Llinas, 1997; Shafer, 1976).

Line 483 "stronger" should be "greater" or just "larger"

Done.

For Figs. 6 to 8, as the statistical criteria are generated by the entire network, statistical filters such as the range of annual mean or even monthly standard deviation are stricter for polar regions (where the natural variability of total ozone is much ~~stronger~~ larger) than they are for mid-latitude or low-latitude regions.

Figure 8 Why don't the stations "SIN", "NAT', "BBN" and "HBT" have actual names? Are these Eubrewnet identifiers? You should give the place-name as well to help the reader.

Correct, those are EUBrewnet sites that don't have a WOUDC number ID. Done. Figure 8 has been updated.

Figure 8 caption – you can't say "same as" because the criteria for green is different.

Done.

Figure 8. *Same as Figure 6, but the results are for the Brewer network and assessed using satellite measurements. Note that the assessment criteria are determined based on Brewer network records (which only reflect Brewer networks' internal performance).*

Lines 505-511 The discussion repeats what has already been said previously, but I think the more important thing is that the figures need to make clear that the criteria for each network are different, for someone who only looks at the figures and doesn't read all the text. The captions needs to make this explicit.

Done. We have included the suggested changes in the figure captions to clarify the statistical criteria that have been used. We have exactly the same worries that the reader might misinterpret our results. So, the "repeats" here are to prevent the reader from missing this critical information in Section 3 and also to provide a bit more discussion based on the results just shown here (in Section 4).

Line 516 I hope you checked the others. It's disappointing that the mistakes were there but good that you found them.

Yes, we have some communication with the NASA team, hopefully, such issues will be resolved in future.

Lines 518-519 I don't follow the logic of the sentence. Newer sites have not been included in AVDC. Isn't the problem that you are relying on an overpass service that (I believe) isn't being maintained anymore. It's not an issue of satellites per se.

Some satellites still produce OVP files, such as OMI. However, given that they did not update the station list, we were not able to find OVP files in AVDC for new Pandora sites. We fully agree that this is not a satellite issue, but a data extraction need for end users.

Figure C1 I can't see any the circles or crosses other than in the first panel? It is hard to take a clear message from this figure.

As answered in previous questions, we have 24 lines (20 satellites and 4 reanalysis data) jammed together. The fact that we can't "distinguish" them is good for us, as there are no major issues. A more detailed check for each individual satellite or reanalysis will be shown in Part II of this work in future, which will better illustrate their performance. The purpose of this Figure C1 is only to illustrate what has been done in our initial quality control step and to show the general agreement of all satellites and reanalysis data.

Technical corrections

Line 129 Insert "The" before "Dobson network"

Done.

Line 134 Insert "the" before "1970s".

Done.

Line 136 Insert "the" before "Brewer"

Done.

Line 175 "since many years" – please reword

Done.

The ozone retrieval settings are harmonized within the FTIR NDACC network since ~~many years~~ 2008 (Vigouroux et al., 2008, 2015).

Line 230 Insert "a" before Kalman

Done.

Line 264 Reword "have been found suffering from"

Done.

OMDOAS (OMI data processed via the DOAS method) overpass data have been found ~~suffering from~~ to be deteriorating due to drifting issues in the last decade, and have been removed from the inputs for the statistical model.

Line 283 "glitches" – please change to a different word, perhaps "discontinuities"

Done.

Despite such ~~glitches~~ discontinuities in the reanalysis, Fig. 3a still shows some important results, such as the general agreement of all networks being within a 2-3% level for most of the records, with the agreement between the two reference networks (Dobson and Brewer) mostly within 1%.

Line 286 – "spectra" should be "spectral"

Done.

Line 311 "but not included in this work" shouldn't be in brackets

Done.

Line 324 Delete "were" before "changed"

Done.

Line 437 Insert "the" before "Dark green"

Done.

Line 438 Insert "a" before "Red".

Done.

Line 550 "decent" is not the right word here

Done.

The sites host~~ing~~ world and regional reference instruments show~~ing~~ ~~decent~~ good long-term records.

Reference

van der A, R. J., Allaart, M. a. F., and Eskes, H. J.: Extended and refined multi sensor reanalysis of total ozone for the period 1970–2012, Atmos. Meas. Tech., 8, 3021–3035, https://doi.org/10.5194/amt-8-3021-2015, 2015.

Bell, B., Hersbach, H., Simmons, A., Berrisford, P., Dahlgren, P., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Radu, R., Schepers, D., Soci, C., Villaume, S., Bidlot, J.-R., Haimberger, L., Woollen, J., Buontempo, C., and Thépaut, J.-N.: The ERA5 global reanalysis: Preliminary extension to 1950, Q. J. Roy. Meteor. Soc., 147, 4186–4227, https://doi.org/10.1002/qj.4174, 2021.

Björklund, R., Vigouroux, C., Effertz, P., García, O. E., Geddes, A., Hannigan, J., Miyagawa, K., Kotkamp, M., Langerock, B., Nedoluha, G., Ortega, I., Petropavlovskikh, I., Poyraz, D., Querel, R., Robinson, J., Shiona, H., Smale, D., Smale, P., Van Malderen, R., and De Mazière, M.: Intercomparison of long-term ground-based measurements of total, tropospheric, and stratospheric ozone at Lauder, New Zealand, Atmos. Meas. Tech., 17, 6819–6849, https://doi.org/10.5194/amt-17-6819-2024, 2024.

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

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, 1737–1747, https://doi.org/10.1029/1998JD100046, 1999.

Fioletov, V. E., Labow, G., Evans, R., Hare, E. W., Köhler, U., McElroy, C. T., Miyagawa, K., Redondas, A., Savastiouk, V., Shalamyansky, A. M., Staehelin, J., Vanicek, K., and Weber, M.: Performance of the ground-based total ozone network assessed using satellite data, J. Geophys. Res. Atmos., 113, https://doi.org/10.1029/2008JD009809, 2008.

Gordon, I. E., Rothman, L. S., Hargreaves, R. J., Hashemi, R., Karlovets, E. V., Skinner, F. M., Conway, E. K., Hill, C., Kochanov, R. V., Tan, Y., Wcisło, P., Finenko, A. A., Nelson, K., Bernath, P. F., Birk, M., Boudon, V., Campargue, A., Chance, K. V., Coustenis, A., Drouin, B. J., Flaud, J. –M., Gamache, R. R., Hodges, J. T., Jacquemart, D., Mlawer, E. J., Nikitin, A. V., Perevalov, V. I., Rotger, M., Tennyson, J., Toon, G. C., Tran, H., Tyuterev, V. G., Adkins, E. M., Baker, A., Barbe, A., Canè, E., Császár, A. G., Dudaryonok, A., Egorov, O., Fleisher, A. J., Fleurbaey, H., Foltynowicz, A., Furtenbacher, T., Harrison, J. J., Hartmann, J. –M., Horneman, V. –M., Huang, X., Karman, T., Karns, J., Kassi, S., Kleiner, I., Kofman, V., Kwabia–Tchana, F., Lavrentieva, N. N., Lee, T. J., Long, D. A., Lukashevskaya, A. A., Lyulin, O. M., Makhnev, V. Yu., Matt, W., Massie, S. T., Melosso, M., Mikhailenko, S. N., Mondelain, D., Müller, H. S. P., Naumenko, O. V., Perrin, A., Polyansky, O. L., Raddaoui, E., Raston, P. L., Reed, Z. D., Rey, M., Richard, C., Tóbiás, R., Sadiek, I., Schwenke, D. W., Starikova, E., Sung, K., Tamassia, F., Tashkun, S. A., Vander Auwera, J., Vasilenko, I. A., Vigasin, A. A., Villanueva, G. L., Vispoel, B., Wagner, G., Yachmenev, A., and Yurchenko, S. N.: The HITRAN2020 molecular spectroscopic database, J Quant Spectrosc Radiat Transf, 277, 107949, https://doi.org/10.1016/j.jqsrt.2021.107949, 2022.

Gröbner, J., Schill, H., Egli, L., and Stübi, R.: Consistency of total column ozone measurements between the Brewer and Dobson spectroradiometers of the LKO Arosa and PMOD/WRC Davos, Atmos. Meas. Tech., 14, 3319–3331, https://doi.org/10.5194/amt-14-3319-2021, 2021.

Hall, D. L. and Llinas, J.: An introduction to multisensor data fusion, Proceedings of the IEEE, 85, 6–23, https://doi.org/10.1109/5.554205, 1997.

Hendrick, F., Pommereau, J. P., Goutail, F., Evans, R. D., Ionov, D., Pazmino, A., Kyrö, E., Held, G., Eriksen, P., Dorokhov, V., Gil, M., and Van Roozendael, M.: NDACC/SAOZ UV-visible total ozone measurements: improved retrieval and comparison with correlative ground-based and satellite observations, Atmos. Chem. Phys., 11, 5975–5995, https://doi.org/10.5194/acp-11-5975-2011, 2011.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Kosaka, Y., Kobayashi, S., Harada, Y., Kobayashi, C., Naoe, H., Yoshimoto, K., Harada, M., Goto, N., Chiba, J., Miyaoka, K., Sekiguchi, R., Deushi, M., Kamahori, H., Nakaegawa, T., Tanaka, T. Y., Tokuhiro, T., Sato, Y., Matsushita, Y., and Onogi, K.: The JRA-3Q Reanalysis, J. Meteorol. Soc. Jpn., 102, 49–109, https://doi.org/10.2151/jmsj.2024-004, 2024.

Naoe, H., Matsumoto, T., Ueno, K., Maki, T., Deushi, M., and Takeuchi, A.: Bias Correction of Multi-sensor Total Column Ozone Satellite Data for 1978–2017, Journal of the Meteorological Society of Japan. Ser. II, 98, 353–377, https://doi.org/10.2151/jmsj.2020-019, 2020.

Pazmiño, A., Goutail, F., Godin-Beekmann, S., Hauchecorne, A., Pommereau, J.-P., Chipperfield, M. P., Feng, W., Lefèvre, F., Lecouffe, A., Van Roozendael, M., Jepsen, N., Hansen, G., Kivi, R., Strong, K., and Walker, K. A.: Trends in polar ozone loss since 1989: potential sign of recovery in the Arctic ozone column, Atmos. Chem. Phys., 23, 15655–15670, https://doi.org/10.5194/acp-23-15655-2023, 2023.

Sarkissian, A., Vaughan, G., Roscoe, H. K., Bartlett, L. M., O'Connor, F. M., Drew, D. G., Hughes, P. A., and Moore, D. M.: Accuracy of measurements of total ozone by a SAOZ ground-based zenith sky visible spectrometer, J. Geophys. Res., 102, 1379–1390, https://doi.org/10.1029/95JD03836, 1997.

Shafer, G.: A Mathematical Theory of Evidence, Princeton University Press, https://doi.org/10.2307/j.ctv10vm1qb, 1976.

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, 5105–5123, https://doi.org/10.5194/amt-11-5105-2018, 2018.

Vigouroux, C., De Mazière, M., Demoulin, P., Servais, C., Hase, F., Blumenstock, T., Kramer, I., Schneider, M., Mellqvist, J., Strandberg, A., Velazco, V., Notholt, J., Sussmann, R., Stremme, W., Rockmann, A., Gardiner, T., Coleman, M., and Woods, P.: Evaluation of tropospheric and stratospheric ozone trends over Western Europe from ground-based FTIR network observations, Atmos. Chem. Phys., 8, 6865–6886, https://doi.org/10.5194/acp-8-6865-2008, 2008.

Vigouroux, C., Blumenstock, T., Coffey, M., Errera, Q., García, O., Jones, N. B., Hannigan, J. W., Hase, F., Liley, B., Mahieu, E., Mellqvist, J., Notholt, J., Palm, M., Persson, G., Schneider, M., Servais, C., Smale, D., Thölix, L., and De Mazière, M.: Trends of ozone total columns and vertical distribution from FTIR observations at eight NDACC stations around the globe, Atmos. Chem. Phys., 15, 2915–2933, https://doi.org/10.5194/acp-15-2915-2015, 2015.

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, 2961–2988, https://doi.org/10.1175/JCLI-D-16-0699.1, 2017.

WMO/GAW: WMO Global Atmosphere Watch (GAW) Implementation Plan: 2016-2023, 2017.

WMO/GAW: Global Atmosphere Watch Central Facilities Activity Report (2010–2023), WMO, Geneva, Switzerland, 2026.

Zhao, X., Fioletov, V., Brohart, M., Savastiouk, V., Abboud, I., Ogyu, A., Davies, J., Sit, R., Lee, S. C., Cede, A., Tiefengraber, M., Müller, M., Griffin, D., and McLinden, C.: The world Brewer reference triad – updated performance assessment and new double triad, Atmos. Meas. Tech., 14, 2261–2283, https://doi.org/10.5194/amt-14-2261-2021, 2021.

Zhao, X., Fioletov, V., Redondas, A., Gröbner, J., Egli, L., Zeilinger, F., López-Solano, J., Arroyo, A. B., Kerr, J., Maillard Barras, E., Smit, H., Brohart, M., Sit, R., Ogyu, A., Abboud, I., and Lee, S. C.: The site-specific primary calibration conditions for the Brewer spectrophotometer, Atmos. Meas. Tech., 16, 2273–2295, https://doi.org/10.5194/amt-16-2273-2023, 2023.

Zhao, X., Griffin, D., Fioletov, V., McLinden, C., Liu, X., Park, J., Petropavlovskikh, I., Hanisco, T. F., Szykman, J., Valin, L., Baumann, E., Cede, A., Tiefengraber, M., Gebetsberger, M., Uesato, I., Zheng, X., Ahn, S., Chang, L., Lee, W.-J., Kim, J. H., Lee, H., Baek, K., Redondas, A., Fujiwara, M., Wang, T., Grutter, M., Houck, J. C., Haffner, D., and Lee, S. C.: Geostationary Satellites Total Ozone Observations: First Results on Ground-Based Networks Validation Efforts for TEMPO and GEMS, Geophysical Research Letters, 52, e2025GL114768, https://doi.org/10.1029/2025GL114768, 2025.