Reply to Referee #1:

Many thanks to referee #1 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.

The paper is of high relevance for the scope of ACP. It assesses the quality of the largest ground-based total column ozone observation networks in a time-dependent manner and allows to identify problematic stations and times. It also provides intercomparisons between the ground-based records and the major satellite and re-analysis products.

The paper is based on a sound methodology and well written. I recommend it for publication after minor revisions.

specific comments:

- l. 37-39: differences to what, re-analysis? what is the internal consistency for Dobsons?

Done. Details and precise numbers were provided in Section 3.2. The number for Dobson is 2.04.

Network-wide distributions of daily mean differences to those from reanalysis and satellite data indicate comparable internal consistency for Brewer, Dobson, FTIR, and Pandora (standard deviations generally around 2%), while Filter and UVVIS exhibit slightly broader spreads (<3%), especially at high latitudes.

- l. 65/191: you might also want to put this into context with the upcoming activities with the BTS array spectrometers with WMO / SAG ozone mandating some evaluation of its capacity as potential 3rd generation ozone reference instrument as an information to the reader. But as there is no network associated to it (at least for now) I leave it to you if you to decide if you think such a statement would be sensible.

Done. A new sentence is added at the end of Section 2.1.

It is worth mentioning that the WMO Scientific Advisory Group on Ozone and UV is working to evaluate the capabilities of other new technologies as a potential third-generation ozone reference instrument, including Pandora and the BTS-solar instrument (Zuber et al., 2021).

- l. 80-81: what do you exactly mean by "assimilation-driven shifts in the reanalysis datasets"? Some clarification and also the mention which datasets are (not) assimilated in which re-analyses will help to better understand not only this sentence but the whole paper.

Done. Due to changes in assimilation sources (different satellites), the reanalysis dataset could have an artificial "step function" change. The assimilated satellite datasets were provided in Section 2.3.

When comparing these latest satellites and reanalysis datasets with ground-based observations requires accounting for assimilation-source driven shifts in the reanalyses datasets and latitude- or season-dependent biases in satellite records.

- l. 91: similar to the above comment. Can you explain what you mean by time-dependent shifts in data assimilation

Done. Even for the period when assimilation sources are not changed (i.e., the same group of satellites), some satellite datasets themselves also suffer from shifts caused by many reasons, such as degradation of the optics or detector. Evidence was found in this work, but it will be more extensively discussed in Part II of this work, which will focus on the evaluation of satellites and reanalysis performance.

By avoiding a reliance on a single reference, it mitigates biases caused by specific reanalysis artifacts and time-dependent shifts (such as some satellite datasets could drift over the years due to degradation of optics and/or detector) in data assimilation.

- l. 113: "from their entire 85-year records" -> as you mention in the introduction some sites started measuring in the 1920ies already, there are active stations with > 85 years of data.

Done.

Figure 1a shows the map for these networks from their ~~entire~~ 85-year records (1940 to 2024; pre-1940 records were not included).

- Fig 1. Please specify in the caption what the size of the circle means. Without this information the figure is not clear to me. Additionally a small cosmetic comment: the figure you provide seems to have lost its aspect ratio in the sense that the "circles" are actually ellipses (higher than wide)

We used different sizes of circles only to make those “super sites” with multiple instruments more visible. If the same size for all, it will be very difficult to see the different types of instruments for those ones. Caption has been updated to clarify this. The plot has also been updated to make the projection of circles look better. Following the suggestions from another referee, the map has been updated too (four panels to show the reference network and other network individually).

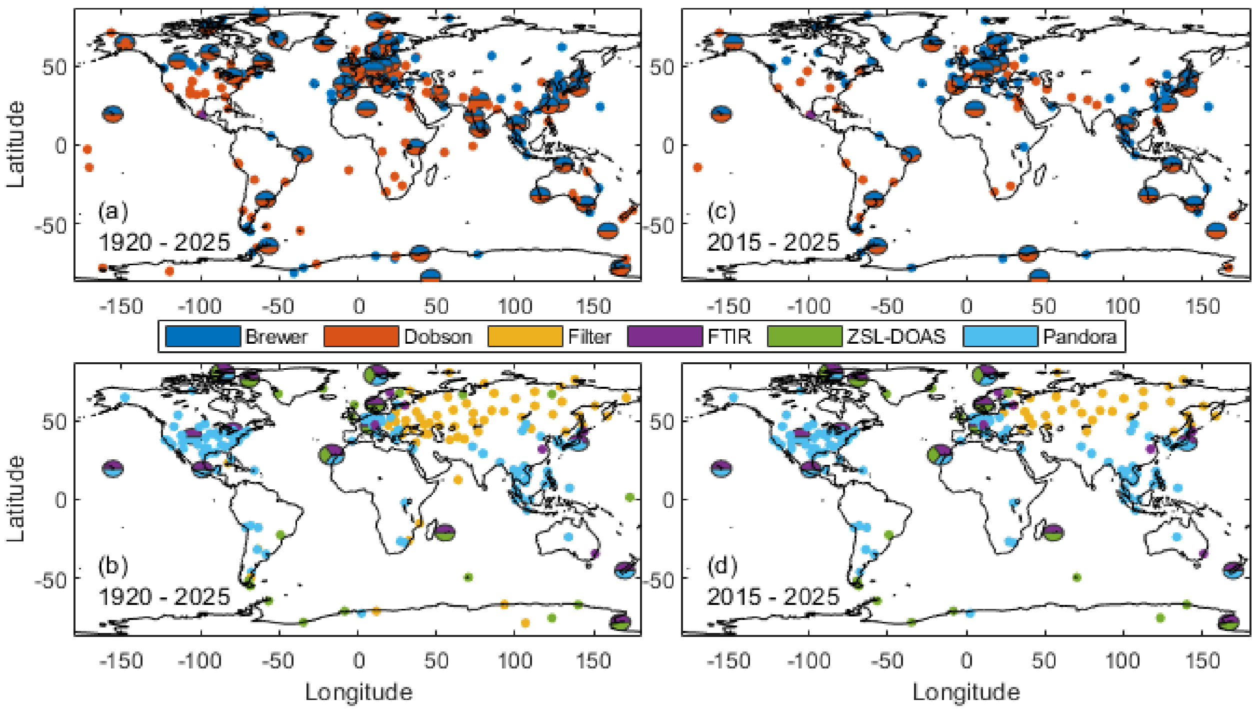

Figure 1. *Map of ground-based TCO observation networks, the panels a and b show all sites that have ever (1940 to 2024) submitted data to WOUDC, NDACC, EUBrewnet, or PGN, the panels c and d show the sites that have submitted data in the past decade (2015 to 2025). The panels a and c show Dobson and Brewer sites, while panels b and d show sites equipped with other instruments. The larger pies correspond to sites with multiple instruments.*

- l.125-126: "it employs a double monochromator and optical wedge system to provide high measurement precision" --> you might also want mention the long optical path within the instrument which is an advantage in terms of stray light suppression

Done.

Operating typically in the 305–345 nm spectral range, it employs a double monochromator (with a long optical path within the instrument to reduce stray light) and optical wedge system to provide high measurement precision.

- l. 125: "e.g. zenith sky" --> also mention Umkehr

Done.

There are other types of observations (e.g., zenith-sky total ozone and Umkehr vertical profile measurements), however, in this work, only direct-sun observations are included, given that it is the primary data products from Dobson and have the highest quality for this type of instrument (e.g., Gröbner et al., 2021).

- l. 129: I would argue that "IGY" would be a more adequate and widely used abbreviation for the International Geophysical Year.

Done.

Dobson measurements started in the 1920s, with global-scale monitoring established during the International Geophysical Year (IGY), 1957-1958 (Brönnimann et al., 2003).

- l. 129: "Dobson network" --> "The Dobson network"

Done.

- l. 131: Mention that D083 is itself calibrated using Langley plots under the excellent atmospheric conditions at Mauna Loa

Done.

The Dobson network is homogenized by using Dobson 083 as the primary WMO GAW World Calibration Standard, which receives independent absolute calibration at the Mauna Loa Observatory.

- l. 125-133: please specify which ozone cross-sections and ozone effective temperatures were used in the retrieval of the data in the actual paper. This is especially relevant for Dobson spectrophotometers where the often-used Bass and Paur cross-sections with constant temperature are known to introduce a artificial seasonal cycle (and an offset at polar sites) which may interfere with your seasonality analysis. Please also make a statement in your analysis how this change would affect it.

Done. New information has been provided in Appendix B, where we had the figures for the criterion of seasonal amplitude.

Note that some ground-based networks have larger seasonal biases than others. For example, the Dobson network data used in this work is processed with the Bass and Paur cross-section, which induced a well-known artificial seasonal bias (e.g., 0% in summer and up to 2.5% in winter in some midlatitude sites). Thus, it is critical for each network to use its own statistics. An ongoing effort in improving the Dobson network's records on WOUDC is to re-process the data to the new Serdyuchenko cross-section scale (Voglmeier et al., 2024), to minimize the seasonal bias.

- l. 136-137: why do you capitalize "World" and "Regional" but not "calibration centres". This looks strange to me.

Done.

- l. 167-168: I do not understand why using radiative transfer for the airmass is a "cost" in itself. In my eyes, it would rather be the input data to the model that can introduce uncertainty, but I would argue that you have these assumptions almost everywhere. Dobsons and Brewers calculate an airmass based on geometrical considerations and atmospheric profiles.

Yes, we have assumptions for other types of measurements. However, the major issue is that, to calculate the ozone air mass factor for scattering sunlight observations, the RTM needs inputs such as cloud conditions, aerosol profiles, temperature and pressure profiles, surface albedo, etc. All these inputs mostly come from coarse climatology data, which will inevitably induce higher uncertainties. In fact, instead of running RTM for each observation, the UVVIS network also typically uses pre-constructed look-up tables. This is very different from the direct-sun air mass factor calculation. The RTM itself is estimated to have 0.7% uncertainty; however, the combined uncertainty in the RTM calculated ozone air mass factor will be 3.6% (Hendrick et al., 2011). This is much higher than the uncertainty in the direct-sun ozone air mass factor (<0.5% when SZA < 80 degrees). Some of the explanations have been provided in the manuscript.

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 typical accuracy for UVVIS total ozone is on the level of 5%.*

- l. 206: please make clear from the beginning that (almost) no ground-based data is assimilated in any of the re-analyses, so that these can be considered as independent validation data for the ground-based observations, except for MSR2 before 1979 (and to some degree maybe also afterwards due to the correction you mention)

Done.

Most reanalysis datasets do not assimilate ground-based observations (except MSR2; details are provided later), and thus can be considered as independent comparison targets for the ground-based observations.

- l. 266-269, Fig. 2: A comment on the almost zero pre-satellite bias of MSR2 with Dobson assimilated might be insightful (to guide the reader in first place). If additionally you can think of a way to assess whether this is only due to quite local tearing of the model towards the assimilated observations which are also used for verification or by how much this is a real general improvement of the model quality (i.e. if there are non-assimilated ground-based observations for verification, e.g. filter ozonometer), this would be a very interesting finding.

MSR2 have used the BUV data, with a fitted offset and SZA dependent correction based on the available ground data, which are the Dobson stations and Filter instruments. This process includes a visual check of the time series of each ground station; suspicious records are removed for the entire year from that ground station in the fit. That means that not all data has been used. For all years before 1979, the ground data are also assimilated as if it were an extra satellite. This explains why there is almost no difference between MSR and Dobson data. A new sentence has been included in the manuscript for clarification.

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.

- Fig. 3: please make more clear that in panel b) you are showing the std over all stations' annual means (at least that is what I understood/guessed from your plot).

Done.

The bottom panel shows the standard deviation of the annual mean difference over all stations.

- l. 304-306. This is almost the same sentence as at the beginning of section 2.4. Please rephrase and/or consider avoiding duplication.

Done. We removed the duplicated parts.

After the initial quality control, the five statistical characteristics were calculated for each site and instrument type.

- l. 330-331: this is a bit like comparing peers to apples and might advantage the dataset with the shortest records, because satellite and re-analysis quality could be expected to be best in the more recent years. Could you please indicate the same values for a coincident time period of all instruments as well?

We fully agree with the referee on this point. The total ozone dataset from each network we evaluated has two dimensions: the length in time and the size of the network. As a result, the size of the dataset from each network is very different. In the 1978-2024 period, the largest dataset is from Brewer (N=3060, i.e., one satellite-ground paired station's data in one 5-year bin will be considered as one data point), and the 2nd place is Dobson (N=2373). The smallest dataset is, in fact, from FTIR (N=312), not the youngest network PGN (N = 1193). Thus, the length in the time dimension might affect the results, but the size of the network will also contribute. Note that here, the statistics we generated are used to perform an overall evaluation of the total ozone datasets produced by different networks. Limiting it to only recent data of a network to derive such criteria will not help this "whole dataset" evaluation process. In general, we believe most networks show very good internal agreements on a 2-3% level. Small differences mostly come from those low-quality data points from some problematic period of certain sites.

We still processed the statistics, limited to only using data since 2010.

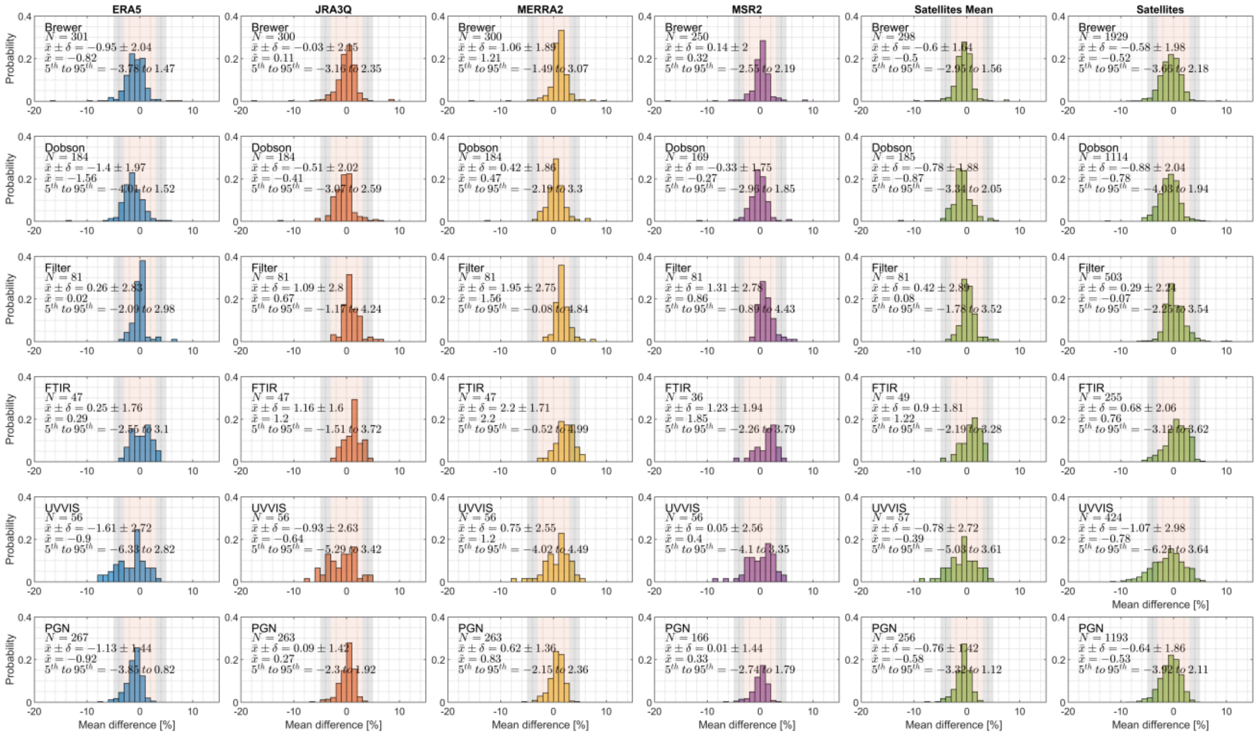

Fig R1. Same as Fig. 4, but only using data since 2010.

As one can see, the standard deviation of the daily mean for the Dobson network is still 2.04 (while the N decreased from 2373 to 1114; more satellite observations in the most recent decades). The values for the Brewer, FTIR, UVVIS, and PGN all have minor changes, except the Filter network, which changed from 2.63 to 2.06.

- l. 337: From l. 228-235 I had understood that MSR2 just assimilated Dobson in the pre-satellite era, so I am confused with the statement "This is because MSR2 assimilated the Brewer and Dobson observations from some sites" made here. Please clarify this and the reason for the better agreement.

Yes, the Dobson data for the pre-satellite period were directly assimilated. For the satellite period, Brewer and Dobson data are still used in the assimilation process, but via providing the corrections (e.g., SZA bias) to the satellite datasets. This has been clarified:

In the pre-satellite era before 1979, MSR2 assimilated ground-based measurements from Dobson instruments. Before assimilating the satellite data, some basic corrections as a function of solar zenith angle, viewing angle, and atmospheric temperature are applied to *the satellite data. These corrections are based on a comparison of the satellite data with the ~~ground~~ Brewer and Dobson network (van der A et al., 2015).*

This is because MSR2's assimilation process included satellite data corrections based on the Brewer and Dobson observations from some sites.

- Fig C1: The fact of showing lines in this graph is confusing as it suggests that the different datasets are somehow related in a series. Why not showing this as stacked boxplots or something the like?

We thank the referee for this suggestion. We have tested many options to show the results better and be absorbed by the reader easily. However, please note that there are too many data points on these figures. For example, the first panel has 20 lines, which in fact contain 120 data points; six networks, each network has results against five comparison targets (four reanalysis and grouped satellite data), each comparison target shows four numbers (95^{th}, 90^{th}, 10^{th} and 5^{th} percentile).

So, we grouped the results into visible shapes/types of lines and the colour of lines to better illustrate the difference between different networks. Stacked boxplots will lead to more jamming of the figure, or we will need to plot different networks on different figures. We have cleared this in the description.

The percentile threshold derived for each network by using different comparison targets is summarized and illustrated in Fig. C1. Note that the results from different networks are not related in a series; plotting them together in different types of lines is only for better illustration.

l. 369-381: you tend to use "will" here, but I assume you are rather speculating on what "would" happen if...

Done.

- l.389: I don't understand your condition 2). Can you describe it a little more, so that the reader is not obliged to check Fioletov et al. just to understand the main working idea of your method. Finally it is your main classification method and would deserve a little more description (although it is similar to Fioletov et al.)

Condition 2 is to say that if any time bin has one outlier, or one outlier and one suspect, this time bin will be labelled as having minor issues. To make things a bit clearer, we included a table here:

This categorization is summarized in Table 1.

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*

- l. 399-401: the parameters given in the parentheses seem more important to me than to be given in form of such a sidenote. Either move to some methodology section or a table.

Done. We removed the parentheses.

This happens mostly because the coincident high-quality satellite overpass data is fewer than the required number to calculate the statistics. ~~(e.g., for~~ 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~~)~~.

- Fig. 6/7: why is Davos not mentioned as a continuation of the Arosa time series? The operators and PI's of this site are presenting this time series as one with a generic Arosa-Davos station and there are ongoing homogenisation efforts for an improuve merge of Arosa and Davos. Stübi et al. 2017 and 2021 analysed the re-location and concluded the effect is < 0.5%.

Done.

This result extends our knowledge of the historical datasets, for example, the European Dobson references at the Arosa site show high-quality results since the 1940s. It is worth mentioning that the Dobson instruments at the Arosa site have now been relocated to the Davos site. Stübi et al. (2017, 2021) analyzed the relocation and concluded the effect is < 0.5%. Ongoing homogenization efforts to produce a generic Arosa-Davos Dobson record are in progress.

- l. 501-502: reformulate the part related to FTIR in the sense that it becomes clear that you have primarily used the reprocessed time series in the present work. Are the newer data (after 2023) processed by the individual institutes again?

Note that the newer data are also processed by individual teams. The reprocessing is named IRWG2023 to be distinguishable from the previous submissions. I.e., on the NDACC FTIR records, there are two versions of the records. Instead of using version 1 or 2, the infrared community decided to just give the reprocessing dataset extension in file names (_irwg2023). However, a lot of harmonization is made in the retrieval settings, such as spectral windows, spectroscopy, a priori profiles, etc. (see Björklund et al., 2024; Vigouroux et al., 2015). We have modified this part to make it a bit clearer to the reader.

FTIR, UVVIS, and PGN do not have a hierarchical calibration chain like Dobson, Brewer, and Filter networks. Pandora results shown in this work are centrally processed, while the FTIR and UVVIS results are processed by individual instrument teams (but following community standards). For example, the FTIR reprocessing data (IRWG2023) has gone through a harmonization work as described in Sect. 2.1 and Vigouroux et al. (2015) (FTIR records from Izana in Spain and Altzomoni in Mexico have not been reprocessed yet).

-l. 523-524: "..., which their PI might lack..." sounds strange and not really causal to me. Please re-formulate

Done. We have reformulated this part to make it clearer. The long-term operations are important for the stratospheric ozone community. Short periods of observation (<10 years) are welcome and useful for satellite validation and verification, but have much less impact than a dataset that can be directly used to verify regional trends (>10 years). Given that PGN is a very young network, we do not want a reader to misinterpret its numbers as capacity for long-term ozone monitoring. In other words, given its application, we are not sure if many PGN sites will survive for decades yet. Currently, 17% of all PGN instruments are owned by universities. The majority of owners are atmospheric, environmental or space agencies (they hold in total 193 instruments out of 230).

Another minor concern is that ~~many~~ some PGN sites are focused on air quality or satellite research, which their PI might lack the resources or even motivation for long-term operations for stratospheric ozone recovery monitoring (for instance, in the ozone monitoring community, it is very common that the instruments operated by university researchers typically retire with their PI, in contrast to the instruments operated by government agencies or institutions). Currently, ~~many~~ 17% of Pandora sites are operated by universities.

-l. 549: "The sites host ..." please reformulate this sentence/the tense of it.

Done.

The sites hosting world and regional reference instruments show~~ing~~ decent 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.

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.

Brönnimann, S., Staehelin, J., Farmer, S. F. G., Cain, J. C., Svendby, T., and Svenøe, T.: Total ozone observations prior to the IGY. I: A history, Q. J. Roy. Meteor. Soc., 129, 2797–2817, https://doi.org/10.1256/qj.02.118, 2003.

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.

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.

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

Stübi, R., Schill, H., Maillard Barras, E., Klausen, J., and Haefele, A.: Quality assessment of Dobson spectrophotometers for ozone column measurements before and after automation at Arosa and Davos, Atmospheric Measurement Techniques, 14, 4203–4217, https://doi.org/10.5194/amt-14-4203-2021, 2021.

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.

Voglmeier, K., Velazco, V. A., Egli, L., Gröbner, J., Redondas, A., and Steinbrecht, W.: The transition to new ozone absorption cross sections for Dobson and Brewer total ozone measurements, Atmospheric Measurement Techniques, 17, 2277–2294, https://doi.org/10.5194/amt-17-2277-2024, 2024.

Zuber, R., Köhler, U., Egli, L., Ribnitzky, M., Steinbrecht, W., and Gröbner, J.: Total ozone column intercomparison of Brewers, Dobsons, and BTS-Solar at Hohenpeißenberg and Davos in 2019/2020, Atmos. Meas. Tech., 14, 4915–4928, https://doi.org/10.5194/amt-14-4915-2021, 2021.