Response to Reviewers – "Evaluating satellite limb measurements of nitrous oxide" by Laura Saunders et al.

We thank the reviewers for their time. The reviewers' comments are given below **in bold** with our responses in red.

Referee #1: Xin Zhou

1. Organization of the "intercomparison" and "validation" sections

I do not think it is helpful to separate the "validation" from the "intercomparison" in the current way. When comparing N_2O profiles from different satellite instruments, readers would naturally want to know both how the satellite datasets differ from one another and how well each of them agrees with independent balloon measurements.

In the current structure, the satellite intercomparison is presented first, with SMR used as the reference for evaluating differences among the satellite datasets. However, at this stage, the reader does not yet know how well SMR itself agrees with the independent balloon observations, or whether there is sufficient justification for using it as the reference. This makes the interpretation of the satellite intercomparison somewhat difficult. For example, a difference between another satellite dataset and SMR does not by itself indicate which dataset is closer to the independent observations.

I therefore suggest reconsidering the organization of these two sections. At a minimum, it would seem more logical to present the validation against the independent balloon measurements before the satellite intercomparison. Alternatively, the authors could integrate the validation and intercomparison more closely, so that the differences among satellite datasets can be interpreted alongside their respective agreement with independent observations.

We thank the referee for this comment, but we disagree with changing the order of discussion. The goals of the project are first to test how consistent the satellite datasets are to see if a merged product is even possible (intercomparison), and second to check for systematic biases through comparisons with independent reference data (validation). Furthermore, the terms "intercomparison" and "validation" are those defined by the ESA LOLIPOP project.

For the intercomparisons, we are not assuming that SMR is the best instrument to compare to; this section highlights just one example of the comparisons that were done. We expect that the further clarification on this point below, which addresses a comment by Anonymous Referee #2 about P8 L202 below, will be helpful.

Technical comments

1. Line 251: "MLS v5 has a clear low bias from about 20–27 km."

The wording "low bias" is potentially confusing. It could be interpreted as meaning that MLS v5 has a small bias, rather than that MLS v5 is biased low relative to the reference. I suggest using wording such as "MLS v5 shows a clear negative bias" or "MLS v5 is biased low from approximately 20 to 27 km." This would also be more consistent with the

subsequent discussion, where MLS v5 is identified as an outlier over much of this altitude range.

The manuscript has been updated so that all instances of "low bias" and "high bias" are replaced with "negative bias" and "positive bias", respectively.

2. Line 254: **"unphysical features"**

Please be more specific about what the "unphysical features" are. What features of the retrieved profiles are considered unphysical, and on what basis are they identified as such?

This has been clarified in the text. The sentence previously read: "Some instruments, in particular HIRDLS, MLS, and SMR also have unphysical features at lower altitudes, usually below about 18 km in the upper troposphere-lower stratosphere (UTLS)."

It now reads as: "Some instruments, in particular HIRDLS, MLS, and SMR also have unphysical features, such as sharp changes in VMR at lower altitudes where air is expected to be well-mixed, usually below about 18 km in the upper troposphere-lower stratosphere (UTLS)."

3. Figure 12:

The biases above approximately 25 km generally have opposite signs for MLS v5 and MLS v6. What causes this substantial change between the two MLS versions? Is it related to changes in the retrieval algorithm, spectroscopy, calibration, or other aspects of the processing? Given that MLS provides an important long-term N_2O record, I think this difference deserves some discussion.

We assume this comment refers to what was previously Figure 13 and is now Figure 14, in which case this is worth explaining in more detail. The following has been added to the end of Section 3.3, after line 352 (in the preprint):

"The v5 to v6 MLS N_2O changes result from different approaches to correcting a calibration parameter for the 190-GHz radiometer that was producing a long-term drift in the v4 H_2O product. In v5, a correction was introduced that improved the H_2O drift, but the N_2O product was produced in a separate retrieval stage. In v6, a different approach allowed the joint retrieval of H_2O and N_2O in the same retrieval stage."

Anonymous Referee #2

Photolysis of N2O does not produce reactive nitrogen (P1 L17). Most of N2O is indeed photodissociated but only the slow oxidation with O1D leads to production of NO and consequent reaction NO + O3 produces NO2 + O2. (e.g. Seinfeld and Pandis, 2012; Prather et al., 2023, or Krysztofiak et al., 2023 that authors already cite).

The wording has been changed slightly to avoid this implication. The first part of this sentence previously read: "It is photolyzed in the stratosphere to form reactive nitrogen, and with the reduction in chlorine-containing ozone-depleting substances such as chlorofluorocarbons…".

It now reads as: "It is photolyzed in the stratosphere, eventually forming reactive nitrogen, and with the reduction in chlorine-containing ozone-depleting substances such as chlorofluorocarbons…".

What are the consequences of the imminent data desert described by Salawitch et al., 2025 on N2O? I believe a short discussion on this would strengthen the motivation of the manuscript.

This study is focused on existing climate data records and evaluating their reliability. The eventual data desert will make studies such as ours, as well as those listed in the introduction, more difficult in the future. The imminent data desert is not our motivation for assessing existing data; however, our study is one of many demonstrating that having multiple datasets available is useful. The bolded note to this effect has been added to the end of the summary (line 534 in the preprint):

"With some adjustments to account for biases, they could be merged into a continuous time series, in the framework of the ESA Climate Change Initiative, for future trend analysis and examination of changes in stratospheric circulation. **This work can be undertaken more easily and further into the future if global N_2O datasets continue to be made available through current and new satellite missions."**

Colocated or coincident? Please, chose one term and use it throughout the manuscript.

All instances of the word "colocated" have been changed to "coincident".

I recommend to split Sect. 3.1 to " Sect. 3.1.1 Colocation (or coincident) criteria" and "Sect. 3.1.2 Types of figures". Similar suggestion for Sect. 4.1.

Section 3.1 has been split into "3.1.1: Finding coincidences" and "3.1.2: Types of intercomparisons". Section 4.1 has been split into "4.1.1: Finding coincidences" and "4.1.2: Types of validation".

Why do not authors use sPV (or similar metrics) for the intercomparison of satellite profiles described in Sect. 3 and only for the validation with balloon data (Sect.4)? Would it not help to improve the intercomparison to ensure that "tested" and "comparison" dataset represent the same air mass? It should improve the results in Fig.1, 60-90N in MAM and 60-90S in SON.

We have investigated this. It turns out that filtering by scaled potential vorticity does not improve the comparisons between satellite instruments. Unlike with balloon comparisons, there are many

profiles available from each instrument, so it is much less likely that coincident profiles will be sampling different air masses. In addition, a 5-degree latitude coincidence criterion for the balloon comparisons was needed to find enough coincidences, and this is more likely to result in "coincident" profiles on opposite sides of the polar vortex edge than using a 300 km distance-based criterion, as was done for the satellite intercomparisons.

To demonstrate this, the medians of the relative differences between SMR and ACE-FTS are highlighted. In one comparison, the profiles are not filtered by scaled potential vorticity (Figure R1) and in the other, they are (Figure R2). There are a few barely perceptible differences between the figures, shown in Figure R3, but the overall biases are not reduced with vortex filtering. These differences are larger at high altitudes because of the small VMRs there. Since ACE-FTS is the only "sparse sampler", and most of its measurements are at higher latitudes, this is a reliable indication that sPV filtering is not required for any other pairs of instruments.

A clarifying sentence has been added to the Validation methodology section (line 367 in the preprint), which previously stated "… Colocated measurements must be in the same region with respect to the polar vortex. The vortex edge…".

It now reads as "… Coincident measurements must be in the same region with respect to the polar vortex. This is necessary because there are fewer coincidences than for satellite intercomparisons, and coincident pairs that sample different air masses occur more frequently. The vortex edge…".

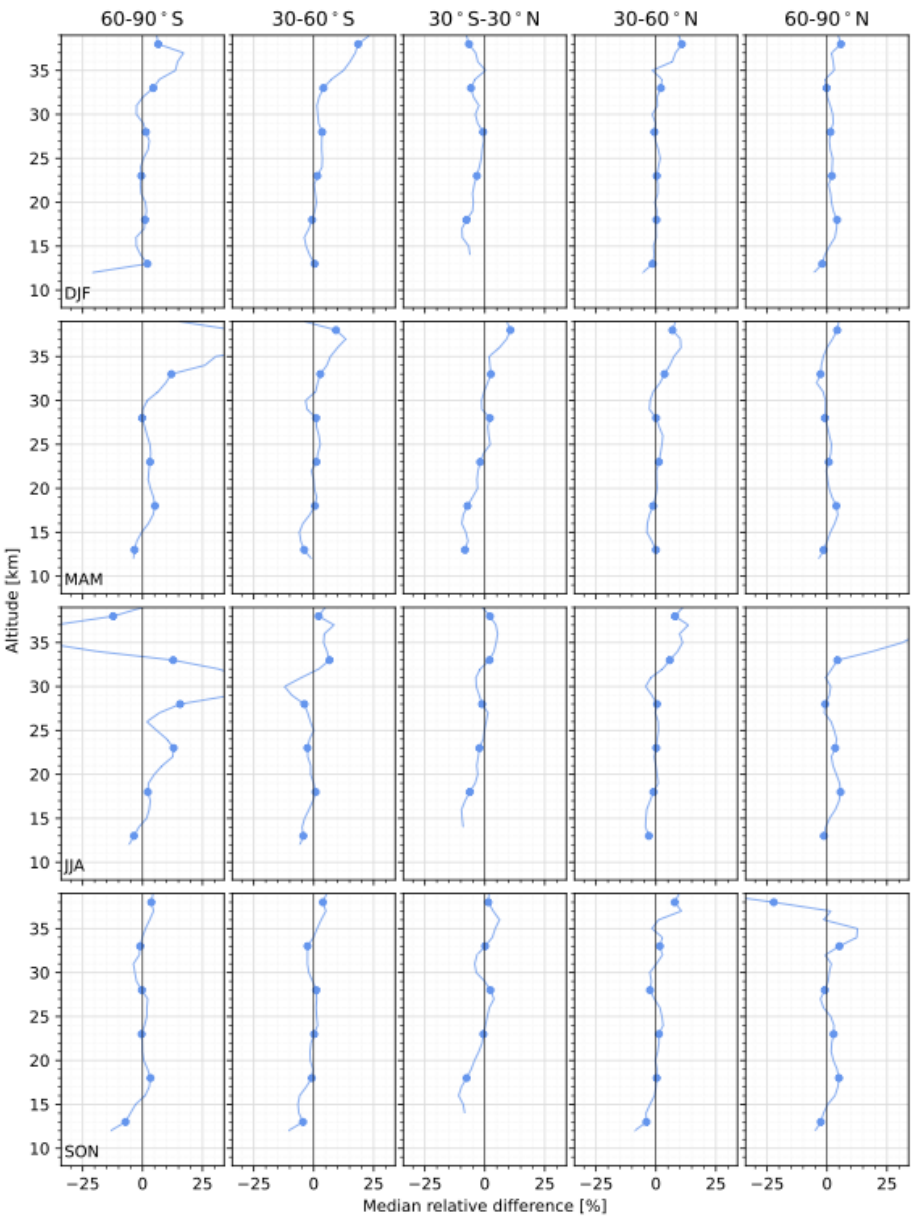


Figure R1: Median relative difference between coincident ACE-FTS and SMR measurements of N_2O, without considering scaled potential vorticity in the coincidence criteria.

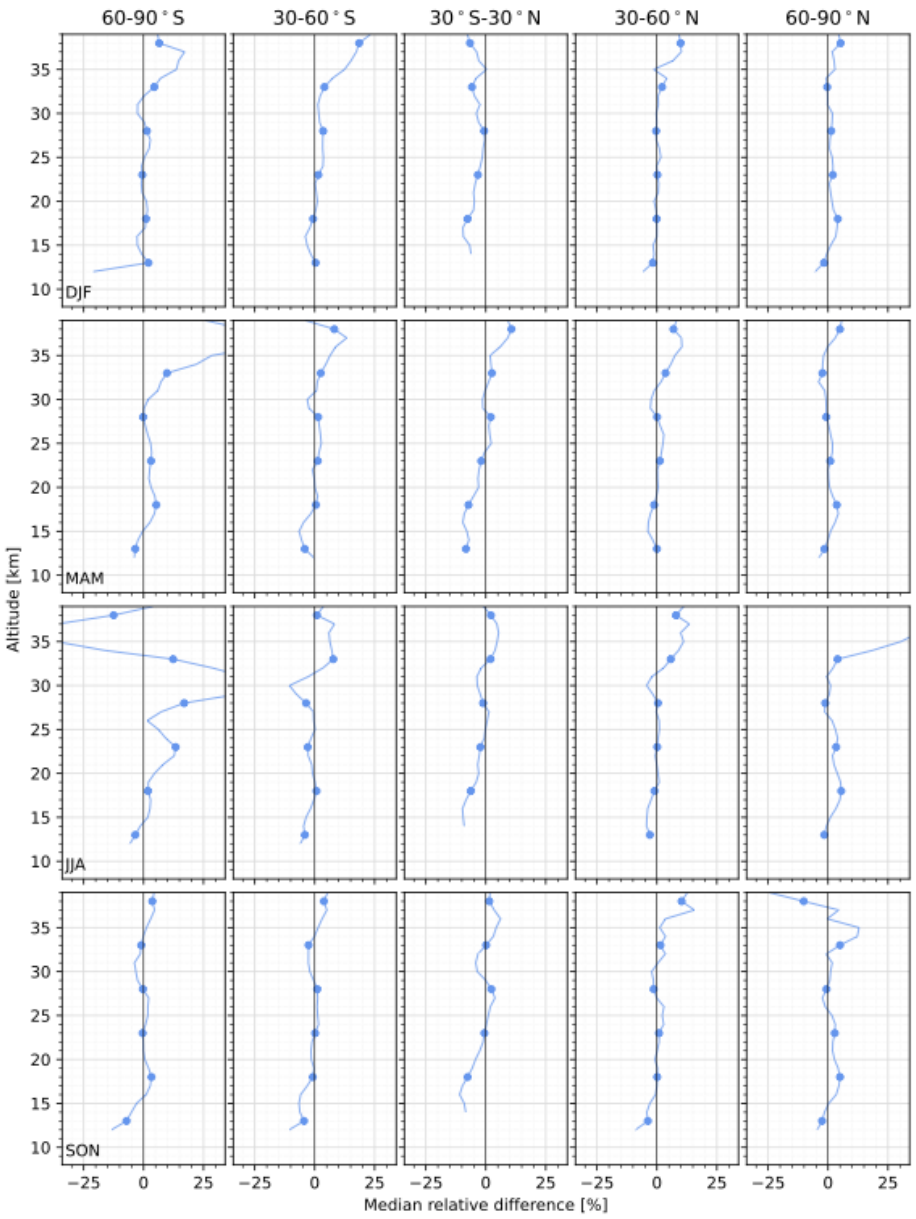


Figure R2: Median relative difference between coincident ACE-FTS and SMR measurements of N_2O, considering scaled potential vorticity in the coincidence criteria.

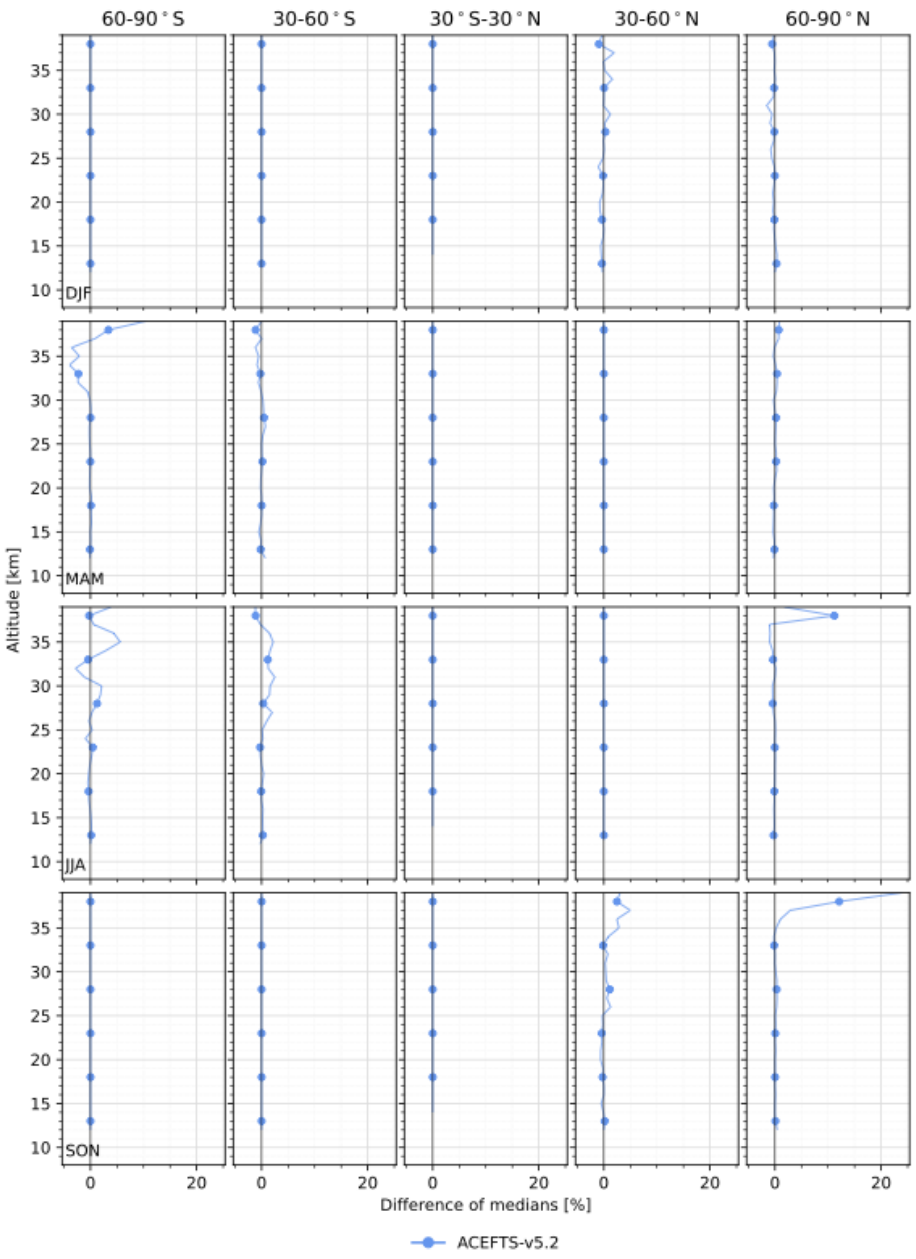


Figure R3: Absolute difference between the median relative differences shown in Figure R2 and Figure R1 (Figure R2–Figure R1).

I was wondering if the authors could add the number of colocated profiles to the text or Figures or maybe any existing/new tables? Especially, taking into account that MIPAS data are treated separately based whether they are full- or reduced resolution, the number of colocations with tested instrument will be very much different. Same comment for the validation section (Sect.4): it would be great to know the number of colocations that the presented results are based on.

This would be useful. A new figure (now Figure 1 in the text, Figure R4 here) has been added to Section 3.1.1 ("Intercomparison methodology: Finding coincidences"), described by the following text, inserted after line 212 in the preprint:

"As an example, Figure 1 shows the number of coincident measurements between SMR and each other instrument at different latitudes and times of year. Especially with ACE-FTS, the coincidences are mostly concentrated poleward of about 60º. There are also fewer coincidences in the northern hemispheric summer due to the Odin-SMR sampling pattern. The equivalent figures for the other instruments are shown in the Supplementary Information (Section S2)."

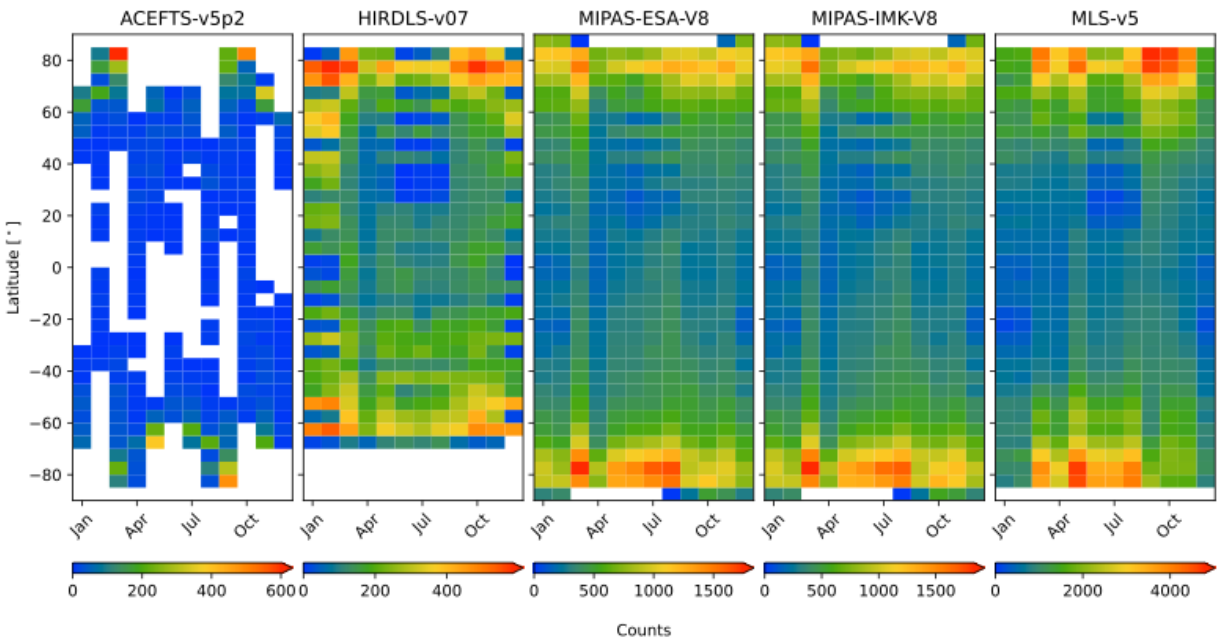


Figure R4: Distribution of coincident measurements between Odin-SMR and the other datasets. Bins with zero coincidences are white; the colourbar is not inclusive on the left side. Note the different scale for the different datasets.

Please, increase markers of the profile plots (e.g. Fig. 1, 12, 14). Also, brown and orange are very similar, maybe increasing both line width and markers will help. Otherwise please, consider another color for either of the datasets.

The marker size has been increased as much as possible in Figures 2, 13, 15, 16, and 17 (previously Figures 1, 12, 14, 15, and 16). If they are any larger, they overlap too much. This is also true for increasing the line width. The brown cannot be changed because, while it would be

easier for those with normal vision, it would make Odin-SMR and MIPAS-ESA impossible to distinguish for those with red-green colour blindness. These six colours were chosen to balance ease of separation for readers with both types of vision. Some colours are similar, but they are distinguishable, especially with the markers.

P1 L2 "… have become available from the suite of limb-viewing satellite instruments." And what about occultation instruments? Especially since authors include one of them to analysis?

"Limb-viewing" refers to the measurement geometry, and not to whether the instrument is measuring absorption (via occultation) or emission. All of the satellite instruments included in this study are limb-viewing.

P1 L10 the two MIPAS datasets → this is confusing by the time the reader gets to Sect. 2.1.3. Maybe the following would be better: "both MIPAS scientific and operational datasets"

This change has been made.

P2 L42 I suggest : "Together, these recent (i.e., since 2002) satelite-based datasets comprise a continues 23-year global record of N2O". Otherwise the sentence gives an impression that each of the datasets is 23-years long.

This sentence has been changed to: "These recent (i.e., since 2002) satellite-based datasets together comprise a continuous 23-year global record of N2O."

P2 L44 I recommend to use "valuable" (as in the abstract) instead of "excellent".

This change has been made.

Table 2. Why is there mean latitude for Table 1 and not for Table 2? Why is there Date/time for Table 1 and not for Table 2? The meaning of e.g. 01, n3, n3a etc. in the Sequence is not clear.

The GLORIA-B profiles used in this study are mean profiles over multiple scans, whereas the MIPAS-B and MkIV profiles are from individual scans. This is explained in Section 2.2.1: "For GLORIA-B, the mean of the profiles taken over the course of each flight was used for comparison, so the mean latitude of these profiles is provided in addition to the nominal latitude of the launch site."

To make sure this is clear, the first sentence in the caption for Table 1 has been modified from: "Details of measurements from GLORIA-B." to: "Details of mean profiles from GLORIA-B."

The sequence numbering identifies the measured MIPAS-B profiles using an internal identification code. The specific meaning of this code is not relevant to the comparisons or conclusions presented in this paper. The sentence discussing it (line 183 in the preprint) has been updated with the bolded text: "The MIPAS-B sequences (i.e., profiles) used for this study are listed in Table 2, which is similar to Table 1 but contains a column for the "sequence number", **which identifies the specific profile used**."

P8 L200 "It is not reasonable to search for coincidences between six datasets simultaneously" → please finish the thought by providing the reason why, e.g. ".. because it would lead to only a few collocated measurements" or so.

This sentence has been changed to: "It is not reasonable to search for coincidences between six datasets simultaneously as it would unnecessarily limit the comparison dataset. It is more practical to search for coincident measurements between pairs of instruments."

P8 L 202 Regarding "tested" and "comparison" dataset → based on the style of authors' writing, it is important for the reader to understand what is tested and what is comparison dataset. However, the next sentence/paragraph on P9 L204 starts a new discussion. The expected explanation comes only on P9 L219, which makes it hard to follow. Please, consider merging this text together. Would it make sense to mention that SMR is "tested"in Fig. 1? And others are "comparison" datasets. Currently Fig. 1 does not mention "tested" and "compared" terminology, I think if authors want to keep it, they should use it throughout the manuscript.

Some additional explanation of the terms has been incorporated into this paragraph, and the explanation for using SMR has been moved here (line 201 in the preprint). Previously, it only read: "it is more practical to search for colocated measurements between pairs of instruments. The process for finding colocated measurements between a "tested" and a "comparison" dataset is described below".

Now, it reads: "It is more practical to search for coincident measurements between pairs of instruments. As such, each dataset must be evaluated separately, using the other datasets as a reference. Here, the dataset being explicitly evaluated is referred to as the "tested" dataset, while the others are "comparison" datasets. Since no instrument is known to be a more reliable reference than the others, all of them are treated as the tested instrument for each type of comparison. This leads to a large number of results, so only results using Odin-SMR as the tested instrument are highlighted while the others are in the Supplementary Information (see Section S2). All of these intercomparisons are summarized in Section 3.2. Odin-SMR was chosen as the tested instrument to highlight because it has the most temporal overlap with the other instruments and a high measurement density, so more coincidences can be found. The process for finding coincident measurements between a tested and a comparison dataset is described below."

In addition, the caption for Figure 2 (previously Figure 1) has been updated from: "… For SMR, all profiles were included in the median. For other instruments, only profiles coincident with an SMR profile were included. …" to "… For the tested instrument (Odin-SMR), all profiles were included in the median. For comparison instruments, only profiles coincident with an Odin-SMR profile were included. …"

Figure 3 (previously Figure 2) previously had the caption: "Median relative difference between SMR and coincident measurements from each other instrument…" and now reads: "Median relative difference between the tested instrument (Odin-SMR) and coincident measurements from each comparison instrument…"

Figure 4 (previously Figure 3) previously had the caption: "PDFs of SMR (coloured) N_2O measurements that are coincident with each other dataset (grey)." and now reads: "PDFs of

tested (Odin-SMR, coloured) N_2O measurements that are coincident with each comparison dataset (grey)."

P9 L204 "All analyzed" instead of "the"

This was changed from "The datasets were first harmonized…" to "The datasets to be analyzed were first harmonized…"

P9 L206 "" ...Supplement Information, see S1" or something like this (same for P10 L240). Please, add references to your Supplements.

Parenthetical references to specific sections in the Supplementary Information have been added.

P9 L216 "..per latitude band" → I recommend that authors introduce the latitude bands here (as they did in Sect. 4.1).

This parenthetical has been added: "(60-90ºS, 30-60ºS, 30ºS-30ºN, 30-60ºN, 60-90ºN)".

P9 L221 "are made up" (and L222) → calculated from/ are based on also I recommend "For other instruments" → "For each other instrument"

This previously read: "The SMR medians and MADs are made up of all profiles for the given time of year and latitude band. For other instruments, that is the comparison instruments, the medians are made up of profiles that are coincident with SMR."

It now reads as: "The Odin-SMR (the tested instrument) medians and MADs are calculated from all profiles for the given time of year and latitude band. For each other instrument (the comparison instruments), the medians are calculated from profiles that are coincident with Odin-SMR."

Fig. 1. I think adding MIPAS full-resolution and reduced-resolution on the legend would make it easier to follow the text in the manuscript. Same for Fig.2? On P4 L117-119 the authors stated that MIPAS data are often treated separately, but unfortunately it was not clear that it is exactly what authors do too.

The legends for Figures 2 and 3 (previously Figures 1 and 2) have been updated to include both MIPAS time periods.

P10-11 While explaining Fig.2, please refer to specific panels in the text.

All references to the polar vortex on these pages have been clarified with "60-90ºN in DJF" or "60-90ºS in JJA", as are references to the Arctic ("60-90ºN") and the Antarctic ("60-90ºS").

Fig.2. Please, unify throughout manuscript: M-ESA or MIPAS-ESA-V8 (same with IMK; same with SMR → figures legends have it labeled as Odin-SMR, see e.g. Fig. 1, caption says SMR but legend shows Odin-SMR).

All references to "SMR" in the text have been changed to "Odin-SMR". "MIPAS-ESA" and "MIPAS-IMK" do not fit in Figure 4 (previously Figure 3, the PDFs), and the following note has been added to the caption: "M-ESA refers to MIPAS-ESA and M-IMK refers to MIPAS-IMK."

P11 L299 Please, add a sentence and say that Figs.4-11 will discuss median absolute and relative biases of each instrument compared to every other instrument.

The following sentence has been added to the beginning of the paragraph (line 299): “Figures 5 through 12 show median differences between each tested instrument and its set of comparison instruments.” These were previously Figures 4 through 11.

P14 L393-394 “The left panels show…” it is unclear which figures the authors are referring to.

The first reference to the figures has been clarified with the bolded text: “Instead, as in Wetzel et al. 2022 where MIPAS-ESA was compared with MIPAS-B, the mean comparisons for flights within the Arctic (60-90ºN), the midlatitudes (30-60ºN), and the tropics (south of 30ºN) are presented, and the figures **(Figures 15 through 17)** are separated by balloon instrument. For these figures…” Note that these were previously Figures 14 through 16.

P18 L518 “..difficulty in measuring in the upper troposphere” → upper stratosphere? Or maybe middle stratosphere?

As explained earlier in the sentence, difficulty in measuring the upper troposphere refers to the lower ends of the profiles (in altitude), which are in the upper troposphere.

Anonymous Referee #3

Narration would benefit from clear explanation of the choice of satellite instruments. I suppose satellite instruments were reduced to post-2000 limb-viewers only, but then why TES and ILAS-II are not included. I am not saying they must be included, and I suspect they are not good candidates for upcoming merging, but a rationale of the choice should be explained.

To exclude ILAS-II more explicitly, the following sentence (line 36 in the preprint) has been modified with the bolded text: "Since the beginning of the "golden age" of satellite observations in the early 2000s, multiple simultaneous global satellite datasets **with mission lifetimes longer than three years** have become available for N_2O limb measurements." TES does not have limb measurements of N_2O.

Difference between first and second type of analysis (inter-comparison and validation) is not clearly explained: what is the "test dataset"? was the analysis performed by taking each of 5 satellite datasets as "test dataset" and only SMR-results were chosen to be shown? If not, the whole inter-comparison part looks rather like validation against SMR.

See response to comment from Anonymous Referee #2 about P8 L202.

l.42 : these records

"Record" is singular in the previous sentence, so this was not changed. The previous sentence was changed to "These recent (i.e., since 2002) satellite-based datasets **together** comprise a continuous 23-year global record of N2O." This emphasizes that this is one record made up of multiple satellite datasets.

l.49: I would add https://acp.copernicus.org/articles/17/11177/2017/acp-17-11177-2017.pdf somewhere at the line 49 and explicitly mention the use of N2O there for positioning the subtropical transport barriers.

This reference has been added. This sentence previously read: "As a long-lived tracer, N_2O is also useful for age-of-air studies (e.g., Linz et al., 2017; Garny et al., 2024; Dubé et al., 2025, and references therein), which provide observation-based information about changes to the Brewer-Dobson Circulation (BDC)."

It now reads as: "As a long-lived tracer, N_2O is also useful for age-of-air studies, which provide observation-based information about changes to the Brewer-Dobson Circulation (BDC, e.g., Linz et al., 2017; Garny et al., 2024; Dubé et al., 2025, and references therein) and subtropical mixing barriers (Stiller et al., 2017)."

l.55: better say SPARC CCMVal and give abbreviation

SPARC needs to be defined because the abbreviation is used later, but CCMVal does not. No change has been made here.

• l.67: are those all available nitrous oxide datasets?

They are all available limb-viewing datasets for that period that lasted at least three years. The previous modification (see response to the first comment from Anonymous Referee #3 above) addresses this.

• l.73: … balloon measurements, all 5 satellite instruments being compared to a same reference, so that …

This sentence has been updated with the bolded text: "This need can be fulfilled through validation, which involves comparing **all six satellite datasets** with **the same** independent reference data such as balloon measurements so that each satellite instrument's biases can be understood as well as possible."

• l.74: … balloon measurements taken for reference are described …

This change has been made.

• l.76: … and conclusionS …

This change has been made.

• l.84: in ACE-FTS description, v5, v5.2 and v5.3 are used and "current versions" is used for both v5 to v5.3, please rewrite accurately.

"Version 5" refers to all v5 sub-versions, but a small change has been made to clarify: "N_2O retrievals were not affected and so the comparisons are still valid for the current version" was changed to "N_2O retrievals were not affected and so the comparisons are still valid for version 5.3".

• l.112: it is 1000 profiles per day, not 1300 (Fischer 2008, section 8.2)

This change has been made. In addition, the description of the reduced resolution period on line 116 of the preprint has also been updated: "The time saved was exploited both to implement a finer vertical sampling of the atmospheric limb and to acquire additional limb scans, about 20% more, within each orbit." was changed to "The time saved was exploited both to implement a finer vertical sampling of the atmospheric limb and to acquire additional limb scans within each orbit, up to 1300 per day."

• l.119: … and in nominal mode typically measured

This change has been made. The sentence now reads: "MIPAS had daily global coverage from 90ºS-90ºN and, in nominal mode, typically measured profiles…".

• l.120: 6 – 68km (not 70, Fischer 2008, section 3.2)

This change has been made.

• l.125: .. averaging kernelS row ..

"Averaging kernel diagonal" is correct (e.g., https://doi.org/10.5194/amt-16-5357-2023 and https://doi.org/10.5194/amt-19-629-2026).

• l.126: … uncertainties constitute the total error budget ...

It is not clear what is meant here as this is the current text.

• l.127: All MIPAS-IMK data – levels or whole profiles?

This was changed to: "All MIPAS-IMK data **points** with…"

• L.129-130: The useful vertical range is from 6 to 60 km: it is said for MIPAS ESA, should earlier be also given for MIPAS IMK.

The bolded sentence was added: "The vertical resolution for MIPAS-IMK, about 3-4 km, is provided in the data files for each individual measurement and is based on the full-width-half-maximum (FWHM) of the averaging kernel row. **The useful vertical range is up to 60 km.**"

• l.130: It is taken – reformulate.

This was changed to "The vertical resolution is calculated as…"

• l.138-139: it not uniformed, it is log-uniformed => uniform log-pressure grid.

"Uniform" does not refer to the grid spacing, and it has been changed to "common" to avoid this confusion.

• l.147: uniform log-pressure grid.

"Uniform" does not refer to the grid spacing, and it has been changed to "common" to avoid this confusion.

• l.147-148: corresponds to 1-sigma precision of the data – really? Please give the reference.

This is stated on page 5 of the MLS v5 quality document: https://mls.jpl.nasa.gov/data/v5-0_data_quality_document.pdf (Livesey et al., 2022). A second citation for this document has been added on line 149: "The median uncertainties are around […] in the polar vortex **(Livesey et al., 2022)**."

• l.150: less than or equal to, greater than or equal to.

This change has been made.

• l.159: The profiles extend from about 12-70 km – reformulate.

This has been changed to "from about 12 to 70 km".

• l.161: which corresponds to the error due to thermal noise and all interfering smoothing errors

– give the reference (it is probably Urban et al 2005, but better to be clear here).

– what does it mean?

This sentence previously read: "The reported uncertainties are the total retrieval error, which corresponds to the error due to thermal noise and all interfering smoothing errors."

It now reads as: "The reported uncertainties include both instrument and retrieval model-related errors, which are explained in detail in Urban et al. (2005)."

• l.163: which ranges from about 3-6 km – reformulate.

This has been changed to "which ranges from about 3 to 6 km".

• l.204: should be explained how the dataset given on pressure grid was transformed to altitude grid: I presume you used temperature profiles from MLS, but it must be clearly stated.

The MLS altitudes are one of the JPL DMPs, so the bolded sentence has been added: "…or about 22 km in the polar vortex. **The altitudes corresponding to each data point are taken from the Jet Propulsion Laboratory (JPL) Derived Meteorological Products (DMPs, Manney et al., 2007, 2011; Millán et al., 2023), which are based on the MERRA-2 (Modern Era Retrospective analysis for Research and Applications; Gelaro et al., 2017) reanalysis**. The data are flagged…"

The JPL DMPs were previously introduced on line 368 in Section 4.1 (now Section 4.1.1). The definitions and citations have been removed from the relevant sentence, which now reads as: "The vortex edge was identified using scaled potential vorticity (sPV) values provided as part of the JPL DMPs."

• l.219: In the comparisons shown below, SMR is used as the tested dataset – this statement is misleading. It reads as you announce an inter-comparison of 5 datasets, the reader expects pairwise inter-comparisons, instead you compare 4 datasets to the fifth one (SMR) alone. Reformulate.

The response to Anonymous Referee #2 about P8, L202 addresses this.

• l.224-232: the usefulness of both type of comparisons must be explained (which one is useful when drift will be calculated, which one when the merging will be done using one instrument as reference dataset etc).

This paragraph only refers to one type of comparison. The opening sentence has been changed from "The second type of figure, the other profile comparisons, are the median relative differences…" to "In the second type of profile comparisons, the median relative differences are shown. A positive value indicates…" to clarify this.

• l.251: variabilities agree - please explain.

In this sentence, "the profiles and variabilities agree quite well" has been changed to "the VMRs and MADs agree quite well".

• l.287-288: these instruments are the outliers – please explain.

This sentence was changed from "There is greater disagreement with HIRDLS and MLS, but the profile comparisons showed that these instruments are the outliers at this altitude" to "There is greater disagreement with HIRDLS and MLS, which were also shown by the profile comparisons to agree less well with Odin-SMR at this altitude".

• l.333: This is due to the deviation in the N2O profile – deviation from what?

This feature was referred to as a "deviation" earlier in this section, but some additional clarification has been added. The feature is now referred to as an "aberration". This section previously read: "MLS v5 has a clear negative bias from about 20-27 km, due to a deviation in the profile, in most regions except for the polar vortex (60-90ºN in DJF; 60-90ºS in JJA and SON)."

It now reads as: “MLS v5 has a clear negative bias from about 20 to 27 km, due to an aberration in the profile, in most regions except for the polar vortex (60-90ºN in DJF; 60-90ºS in JJA and SON). This aberration is characterized by a too-strong decrease with altitude starting near 20 km and a nearly constant VMR between about 23 and 27 km.”

In other sentences where this feature in the MLS v5 profile is mentioned, “deviation” has been changed to “aberration”. This separates this feature from other instances where the word “deviation” is used.

• l.343: Overall, SMR appears to perform very well compared to the other instruments; this is another reason that it was chosen as the tested instrument to highlight earlier in this section.

– This statement is misleading: where the same study (i.e. “test instrument” assigned to the four other instruments) being performed within the project?

The instruments were each treated as the tested instrument, but including all of these results is repetitive and too much for the main text. It was decided that the comparisons for one instrument should be highlighted and that the rest should be shown in the Supplementary Information. The use of the phrase “to highlight” indicates that these are not the only results that exist.

• l.365: where the abbreviation “hr” for “hour” introduced?

All instances of “hr” or “hrs” have been changed to “h”. This is the standard Copernicus formatting for hours. “Regarding the abbreviation of such units, SI base units and SI-accepted units must be abbreviated in conjunction with numbers” and according to their linked brochure, hours are an SI-accepted unit (source: https://publications.copernicus.org/for_authors/manuscript_preparation.html).

• l.399-400: the 1 April 2003 Kiruna flight was separated from the other Kiruna flights because no polar vortex air was sampled – please explain how it was determined.

The following sentence has been added after this statement: “The type of air being sampled was determined with sPV, again using MERRA-2 values (see Section 4.1.1).”

• l.442: all the way from 9-33 km, ranging from 2-25%. – is it from 9 to 33, from 2 to 25? Reformulate

These, and all other instances of “from [number]-[number]”, have been modified to “from [number] to [number]”. In addition, other numerical ranges were changed to use an en dash instead of a hyphen.

• l.450-451: It is the only satellite dataset that does not fully capture the decrease in N2O caused by descent within the polar vortex – I find the statement too strong for what is shown on the graphs.

To soften the statement, “that is seen by the other instruments” was added to the end of the sentence.

• l.461-462: no polar vortex air was sampled during the 2003 flight – same as l.399-400

The previous fix for L399-400 addresses this.

Comments to figures:

• fig.10: why this figure is shown for version 5 only, if the obvious candidate for merging is the version 6?

This is explained in the Aura-MLS section (Section 2.1.4): "Since v6 was released after this analysis was completed, all results presented here are based on the v5 datset [*sic*], with a brief update including v6 presented in Section 3.3."

Additional reasoning has been added to the sentence, and the typo has been fixed: "Since v6 was released after the LOLIPOP study was designed and this analysis was completed, all results presented here are based on the v5 dataset, with a brief update including v6 presented in Section 3.3." Figure 11 (previously Figure 10) shows that MLS v6 is likely a candidate for merging. This is also stated in the conclusions.

• caption of Fig. 17: "Summary of N2O comparisons in the tropics" => "Summary of ballon-borne vs satellite comparisons of N2O in the tropics"

This change has been made.

• caption to Fig.2 and subsequent captions: "each other instrument" => "each of the other instruments".

In the caption Figure 3 (previously Figure 2), "each other instrument" has instead been changed to "each of the comparison instruments". For Figure 4 (previously Figure 3), "each other dataset" has been changed to "each of the comparison datasets". For Figures 5 through 12, the captions that previously read "Median absolute N_2O bias (in ppbv) and relative bias between [dataset] and each other instrument at various altitudes during each season" have been changed to "Median absolute N_2O bias (in ppbv) and relative bias between [dataset], the tested dataset, and each of the comparison datasets at various altitudes during each season." In addition, this further addresses the comment from Anonymous Referee #2 about P8 L202.