Dear Referee,
Dear Editor,

We would like to thank the referee for the detailed review of the manuscript and for the constructive and helpful comments and suggestions. We have taken these remarks into account and revised the manuscript accordingly, with marked changes.

Below, the referee's comments are given in black, our responses in blue, and the corresponding changes in the manuscript in italics. We hope that we have satisfactorily addressed the suggestions and remarks.

Best regards,

Louis Mirallie, on behalf of all the coauthors.

Review of Mirallie et al., *"Ozone stratospheric trends from regional Bayesian composite of ground-based partial columns"*

The submitted manuscript evaluates 2000-2024 trends in ozone at several defined layers of the atmosphere, from the troposphere to the upper stratosphere, using data from long-running ground-based measurement stations. Merged composites are formed of records from different stations using a Bayesian method, and then trends calculated for different heights and regions using multi linear regression closely following the LOTUS model. Finally the derived trends are compared with those derived in other recent work and differences discussed.

Overall this is a very thorough and clearly explained analysis, and in my opinion very suitable to publish in APC once some minor comments below have been addressed.

General comments

My main suggestion is that the authors should spend a little bit more time talking about the motivation for what they're doing. The Bayesian merger of multiple instrument types and locations together is a much more abstract construction than, for example, a single station record, or a zonal mean, or a trend estimated based on one type of instrument only. In many places the data sets show significant divergences from one another in various ways, (eg different magnitude of annual cycles or extremes, varying bias between different instruments) so the question I am interested in is what exactly is represented by this composite? (That is, not how you constructed it, but how the reader should think about it and what it actually represents?)

We thank the referee for this important suggestion. The composite should be interpreted as the regional common ozone signal, that is to say the ozone variability shared across all coherent group members, rather than a simple spatial or instrumental mean. The regions represent the volume of atmosphere that evolves coherently (in

anomalies, at monthly scale) and therefore represent the same ozone signal, although with lower spatial resolution than each individual site.

The differences between instruments' records, diverging from the regional consensus, can be caused by instrumental drift, gaps or biases or by physical differences. The later is supposed to be minimal, by the representativeness study. These differences are reflected within the posterior uncertainty of BASIC's merging, via the PCA-uncertainty and Box-Tiao distribution. In the end, the uncertainty envelope represents the range of variability supported by the ensemble of observations.

The composite cannot perfectly cancel instrumental drifts, but it statistically reduces their influence. But the alignment of the timeseries can result in an overweighted drifting timeseries for some months: if we align it with the beginning, it will contribute more than the end, drifting away. If we align it with the end, similar reasoning inverted. In such cases, the result is an increased uncertainty of the final composite.

Practically, the reader should consider the composite as the best estimate of what a stable instrument with spatial redundancy and the temporal sampling of all contributing records would have measured in that region. It is not a zonal mean, but a regional ozone timeseries with appropriate uncertainties. We added this description in the introduction of section 3.2.

Line 122: "The main motivation of this study is to improve from single-station or single-instrument trend analyses, which are often affected by local data gaps and instrumental drifts. By using a Bayesian merging methodology over coherent regions, we aim to extract the underlying regional common ozone signal. This approach takes advantage of the spatial redundancy within the NDACC network to go beyond a simple spatial average, resulting in a more stable composite for long-term trend estimation."

Line 421: *"The composite aims to be the best estimate of what would have been measured by a single stable instrument, with the spatial redundancy and combined temporal resolution of all the instruments in the group. It is a regional ozone timeseries with appropriate uncertainties. "*

Secondly, it seems like an assumption to me that short-term correlations between sites would necessarily imply similar causes and values of trends, most particularly in the case shown in Figure 5 when the northern hemisphere mid-latitudes and southern hemisphere subtropics are well correlated (but not the higher latitude southern midlatitudes). It seems unlikely to me that the trends in these regions would have influences sufficiently in common to make forming a composite physically meaningful. This point should also be justified.

The grouping relies on the assumption that sites highly correlated in anomalies are driven by the same large-scale forcings. This hypothesis is supported by the fact that the

same forcings dominating interannual ozone variability — QBO, solar cycle, ENSO, and sAOD — are hemispheric to global-scale phenomena that act coherently across wide regions. Sites responding coherently to these forcings are therefore expected to share the same long-term trend drivers, making their composite physically meaningful. This is particularly apparent in the upper stratosphere, where the dominant signal is the solar cycle and BDC transport, both hemispheric to global in scale, which explains why NH midlatitude and SH subtropical sites can be well correlated in this layer. However, other forcings such as QBO and ENSO respond asymmetrically between hemispheres, meaning that high correlation alone is not sufficient to justify cross-hemispheric merging. This is why Maïdo was manually excluded despite meeting the correlation threshold. This is now justified Line 392:
"The grouping relies on the assumption that sites highly correlated in monthly anomalies share the same dominant large-scale forcing mechanisms — notably QBO, solar cycle, ENSO, and sAOD — which are also the proxies used in the trend analysis. This assumption is supported by the fact that these forcings are hemispheric to global-scale phenomena that dominate interannual ozone variability beyond the annual cycle. Sites responding coherently to these forcings are therefore expected to share the same long-term trend drivers. This is particularly apparent in the upper stratosphere, where reduced variability and the dominance of the solar cycle and BDC transport explain correlations across wide latitude bands. However, some forcings such as QBO and ENSO respond asymmetrically between hemispheres, meaning that high interannual correlation alone is not sufficient to justify merging across hemispheres. Interannual correlation is therefore used as a necessary but not sufficient condition for grouping, and physical judgment is applied where the statistical criterion alone would lead to physically questionable composites."

Furthermore, our grouping procedure is more robust than the correlation maps suggest: we require every pair of members within a group to correlate above the 0.75 threshold, not just each member with a reference site. This ensures that even the most geographically distant members within a group remain sufficiently correlated with each other. In practice, the final groups do not cross the equator, with the exception of the equatorial groups in LS and MS which remain within a narrow latitude band around it.

We acknowledge that the assumption has limits. In the Mid-Latitude UpS group, Maïdo (La Réunion, 21°S) met the correlation threshold with NH midlatitude sites but was manually excluded because the two hemispheres respond differently to some of the same large-scale forcings. This statement is now added explicitly to Section 3.1:

"Finally, as visible in Figure 5, the mid-latitudes group in UpS extends from Zugspitze to La Réunion island, including Hawaii. Despite meeting the 0.75 correlation threshold, Maïdo (La Réunion, 21°S) was manually excluded from the final Mid-Latitudes group (see Table S2) because the Northern and Southern hemispheres respond differently to

the same large-scale forcings which could lead to different long-term trend drivers despite short-term correlation."

I very much appreciated the clear explanations of the methodology and comparison to other methods, for example in shown in figure 6 and 10. While some reviewers would possibly object to the inclusion of this material, I think it is very helpful to readers at the present time while these techniques are still fairly new in the community. It was very good that you included more conventional weighted-mean results for comparison in many places.

We thank the referee for this positive feedback.

The decision to only use data from HEGIFTOM ozonesondes is slightly surprising to me. Looking at the project map (https://hegiftom.meteo.be/datasets/ozonesondes) the geographic coverage of your study could be very usefully extended if you included the other ozonesonde sites (eg east Asia, southern hemisphere). While I can see that there is a motivation to take advantage of the presumably greater degree of homogenization that has been carried out in HEGIFTOM, in your analysis there are points where there appear to be unresolved inhomogeneities in the different ozonesonde records anyway, such as lines 451-456 and 579-582, so it seems HEGIFTOM can only provide a partial solution.

We thank the referee for this comment. The choice of HEGIFTOM ozonesondes is first motivated by the homogenization required for long-term trend analysis. As described in Section 2.2.1, ozonesonde records are subject to discontinuities and biases from changes in instrument type, sensing solution, and preparation/processing. HEGIFTOM provides homogenized records with uncertainty estimates at the individual measurement level, suitable for the uncertainty propagation to the monthly mean.

Beyond the homogenization requirement, the BASIC merging method requires co-located records from multiple instrument types. Several instruments sites, like the Antarctic stations or non-HEGIFTOM ozonesondes in East Asia, are either too far or not enough correlated with any other NDACC instrument type (for instance, our current East Asia group relies exclusively on highly correlated FTIRs), making it impossible to form a regional group regardless of ozonesondes data quality. These sites therefore do not appear in our analysis not by choice, but because no meaningful multi-instrument composite can be constructed there. Finally, regarding the remaining inhomogeneities in HEGIFTOM mentioned by the reviewer, this is precisely why we use the BASIC algorithm, whose Gaussian-mixture model is designed to dynamically down-weight such unresolved artifacts. Line 155:

" We strictly limit our ozonesonde selection to this HEGIFTOM dataset because it uniquely provides the per-measurement uncertainties required for our error propagation to the monthly means. Furthermore, isolated sites or instruments that do not correlate

sufficiently with other NDACC records (such as specific ozonesondes in Antarctica or East Asia) are naturally excluded by our regional grouping methodology."

Specific comments

Lines 50-65 The abstract seems too strident in its statements. Is it really true that no ground-based station anywhere in the world shows a statistically significant ozone trend in the lower stratosphere?

The referee is right: we have softened this statement in the abstract to clarify that large uncertainties and variability are among the factors that limit trend detectability in the lower stratosphere, without implying that no significant trend exists anywhere.

Line 50: "Large uncertainties and variability in individual ground-based instrument records limit the detection of statistically significant ozone trends, particularly in the lower stratosphere. Available merging studies are typically performed by latitude bands on satellite-based data records. This study derives correlation-based regional composites of ground-based timeseries towards reducing trend uncertainties."

Lines 64-66 This statement doesn't seem supported by figures 21 and 22. The negative trends are only in very specific regions and most of them are not significant. The one very clear result seems to be the negative trend in the tropical middle stratosphere based mainly on ozonesondes. The generally negative trends seem hard to reconcile with Figure 3 (b) of Sofieva et al. 2025 but I see that in your conclusions you say you will be looking at total ozone in further work.

We thank the referee for both remarks. We have revised the abstract accordingly. The statement on lower stratospheric trends has been softened to reflect that large uncertainties limit detectability. The description of the trend results has been revised to more accurately reflect Figures 21 and 22: robust positive trends in the upper stratosphere, negative trends confined to equatorial and Southern Hemisphere regions in the lower and middle stratosphere, and non-significant trends in the Northern Hemisphere.

Lines 78-79 If co-located instruments can show significantly different trends, doesn't this call into question the validity of combining them in any way at all? Shouldn't these differences be accounted for first?

Discrepancies between trends from individual instruments can occur due to instrumental or data record issues. Combining the data records by applying a similar weight to all of them ignores potential artifacts, thus is a problem. A valid way of combining them is to apply a lower weight to problematic data record (which is appearing different from the others). The differences between the records are then considered during the merging process. We have added this consideration to the text.

Line 94: Differences in collocated data records should be taken into account when combining them using an appropriate merging technique.

Line 80 "large discrepancies" in what?

“large discrepancies in the trend values”. We have corrected this in the text.

Lines 89-91 If it is "critical" to use the different co-ordinate systems of Millan et al. as you say, then shouldn't you also be using co-ordinates relative to the tropopause and the jets?

Yes, the referee is right; this could have been our choice. However, we opted for a simpler approach, selecting a pure UTLS layer with an alternative pressure range. Furthermore, we should be more precise and say that the critical point is not using co-ordinates relative to the tropopause but lowering the trend uncertainties by reducing the variability of the data record, which is the purpose of Millan et al. (2025). Our aim here is also to decrease trend uncertainties by reducing data record uncertainty, despite keeping traditional pressure coordinates. We have now specified this in the text.

Line 103: This allows the data in each domain to be affected by the same dynamical processes, reducing their variability. This reduction is critical for trends detection particularly in UTLS where opposite ozone changes cancel each other in standard pressure coordinates.

Lines 96-97 The way the sentence is written implies these specific satellite instruments had a particularly limited lifetime compared to other instruments not listed – is that what you meant?

We meant that the lifetime of satellites may be limited when compared to the lifetime of ground-based instruments. We have modified the sentence accordingly.

Line 110 “Satellites provide excellent geographical coverage, although their lifetime may be more limited than that of ground-based instruments.”

Lines 97-99 You say the resolution of some of them is "sufficient" – do you mean that for the others it is not sufficient?

Yes, exactly. MWR cannot resolve the lower-stratosphere for instance. We have now specified this in the text.

Line 112: … the resolution of some of them in the lower stratosphere is sufficient: less than 1km for Lidars (Leblanc et al.,2016), a few hundred meters for ozonesondes (Smit et al., 2020), and insufficient for others: MWR cannot resolve the lower-stratosphere below 20 km (Maillard Barras et al., 2020).

Lines 103-107 I would like you to be clearer here about what exactly you're hoping to achieve. Is the motivation to determine representative trends over an area rather than

single points? Is it to remove the influence of very short scale spatial variability? Is it to hope instrumental drifts cancel out in a composite?

Our purpose is to reduce the uncertainty of the trend estimates by using all the information given by individual data records. By merging data records of correlated regions, we get more information as a single data record in this same region. This leads to the determination of a trend for a region, the region coming directly from the CAMS correlation study. This can give the impression of global coverage by ground-based instruments but should be seen as a by-product of the combination of two methods: the CAMS correlation study and the BASIC merging.

There is no canceling out of drifts: the BASIC method considers a drifting data record with less weight as a non-drifting data record in the composite.

We hope to have made this clearer in the text now. Line 118:

There is therefore a sustained need for ground-based global coverage of ozone measurements, which can be achieved only by increasing the number of ground-based stations. A global coverage impression can be reached by grouping the data records into regions and by considering these regions as entities with similar ozone variability. Furthermore, by combining all existing information into a regional composite, the uncertainty of the resulting time series, and thus of its trend estimate, can likely be reduced (e.g. Ball et al., 2017; Arosio et al., 2019; Sofieva et al., 2021; Keppens et al, 2025).

Lines 124-128 You should state that for ozonesondes, though, you don't use NDACC, but HEGIFTOM, which only partly overlap.

Yes the referee is right. We say “are mostly part of the NDACC” on line 124 though.

Our ozonesondes data records do not come from NDACC, but from HEGIFTOM. However, the HEGIFTOM ozonesondes data records used in this study are on their way to being submitted to the NDACC by the individual PIs (HDF formatting made available by R. Van Malderen), while other sites belong to networks like SHADOZ or WOUDC. All other NDACC ozone soundings are either non-homogenised or short-term data records, and have not been used in this study.

We have now modified the text to reflect the delayed submission of the HEGIFTOM data record to NDACC as of the time of this publication.

Line 150: “The data records used in this study are mostly part of the Network for the Detection of Atmospheric Composition Change (NDACC), except that not all the homogenized ozonesondes site data used here belong to NDACC. The NDACC-affiliated ozonesondes data records are on the way to being submitted to the NDACC Data Host Facility by the instrument PIs after having been homogenized within the Harmonization and Evaluation of Ground-based Instruments for Free-Tropospheric Ozone

Measurements (HEGIFTOM) working group of the Second Tropospheric Ozone Assessment Report (TOAR II).”

Lines 132-133 This statement is too strong. The homogenization attempts to account for and correct the jumps and drifts but in reality is unlikely to completely accomplish this. (If it did it would make your life easier!) Your assumption that the differences are now due to random errors and spatial offsets is a reasonable methodological approach but you can't possibly be sure that this is so.

We agree and have softened this statement. The text now acknowledges that homogenization aims to correct artifacts and drifts where possible, but that residual inhomogeneities may remain. The assumption that remaining differences reflect random errors and spatial offsets remains as a methodological working assumption.

Line 150: “The data records used in this study are mostly part of the Network for the Detection of Atmospheric Composition Change (NDACC), except that not all the homogenized ozonesondes site data used here belong to NDACC. The NDACC-affiliated ozonesondes data records are on the way to being submitted to the NDACC Data Host Facility by the instrument PIs after having been homogenized within the Harmonization and Evaluation of Ground-based Instruments for Free-Tropospheric Ozone Measurements (HEGIFTOM) working group of the Second Tropospheric Ozone Assessment Report (TOAR II). We strictly limit our ozonesondes selection to this HEGIFTOM dataset because it uniquely provides the per-measurement uncertainties required for our error propagation to the monthly means. Furthermore, isolated sites or instruments that do not correlate sufficiently with other NDACC records (such as specific ozonesondes in Antarctica or East Asia) are naturally excluded by our regional grouping methodology. “

Line 164: “The homogenization of the time series aims to identify and correct artifacts, jumps and drifts where possible, although some inhomogeneities may remain.”

Lines 148-149 I don't see that forming monthly means "resolves" this problem. For ozonesondes there is the problem of very limited sampling rates, but for Lidars and FTIR it's more the need for clear skies which introduces a bias. I think some more discussion would be beneficial here.

We agree that monthly means does not fully resolve the sampling disparities. We have replaced 'resolve' with 'reduce' and added a discussion acknowledging the two distinct limitations raised: the limited ozonesondes launch frequency and the clear-sky sampling bias affecting Umkehr, Lidar and FTIR measurements. We note that these residual biases contribute to the inter-instrument disagreement captured in the BASIC posterior uncertainty.

Line 178: "The diurnal cycle is not expected to contribute to the disparities caused by different temporal resolutions, given its limited amplitude, maximum in the upper stratosphere, reported to be generally below 4% (Sauvageat et al., 2023). To reduce these disparities, we aggregate all data into monthly means (L3). This approach does not fully resolve all sampling biases: for ozonesondes, the limited launch frequency (2–12 per month) leaves monthly means potentially unrepresentative of episodic events, while for Dobson Umkehr, Lidar and FTIR, the requirement for clear-sky conditions introduces a meteorological sampling bias that monthly averaging cannot eliminate. These residual biases are partially absorbed into BASIC's posterior uncertainty through the inter-instrument disagreement, but are not explicitly corrected."

Line 150 Personally I find the diagram to be very clear and helpful.

We thank the referee for this positive feedback.

Line 160 I can't make much sense of this sentence unfortunately. Do you mean that the proxies used in ozone trend analyses are usually representative of large areas and low frequency variability?

We agree the sentence was unclear. We have reworded it to state explicitly that the regression proxies represent large-scale, low-frequency variability, and that small-scale or high-frequency signals are not captured by the model, leading to larger trend uncertainties. Vertical integration into partial columns reduces this effect by averaging over scales not represented by the proxies.

Line 203: "The proxies used in ozone trend analyses typically represent large-scale, low-frequency atmospheric variability. Small-scale or high-frequency variability is therefore not captured by the regression model, leading to larger trend uncertainties."

Line 165 A very minor point, but if it's "standard" and "alternative" why are they called "o" and "a" and not "s" and "a"?

This is a good suggestion. The naming has been changed to sPC and aPC, as it is clearer. Thank you.

Line 170 You don't specifically explain that you're trying to allow for the decreasing height of the tropopause with increasing latitude by "stepping" the boundary. This should be added to line 168. Even so, you would only get this right in a climatological sense, not for every single day of the record, does that matter?

We agree and have added an explicit statement explaining that the stepped boundary accounts for the climatological decrease in tropopause height with increasing latitude. We acknowledge that this correction is climatological and does not account for daily tropopause variability. This residual effect likely contributes to the larger uncertainties observed in the UTLS layer.

Line 222: "To account for the decrease in the tropopause height with increasing latitude, we define the layers depending on the latitude. For the standard partial columns, this is done using a separation at 30°, and with finer resolution for the alternative partial columns, to ensure that the boundaries remain consistent with the climatological tropopause height.

The daily tropopause variability is not included, which contributes to the noise in LS and UTLS layers. This could be addressed in future work by including a dynamical proxy for the tropopause height. "

Line 207-209 Does it matter that two of the locations are not using the latest version? Wouldn't this also be an inhomogeneity?

The use of older retrieval versions at Altzomoni and Izana may introduce small inhomogeneities relative to other FTIR records. However, both sites contribute only to specific regional groups where they are merged with other instrument types. Any version-related bias should be partially absorbed into the BASIC posterior uncertainty. Updating these records to the latest version is beyond the scope of this study but is noted as a potential improvement for future work.

Line 232 – The need for "quasi cloud free conditions" must introduce a significant bias in the lower stratosphere at least and the UTLS?

The requirement of cloud-free conditions reduces the number of measurements per month and may introduce a sampling bias. However, we do not expect this to introduce a significant bias at the monthly scale. No partial column is preferentially sampled over the others. Only the number of measurements is reduced, which is properly propagated in the uncertainties and accounted for in BASIC's posterior uncertainty.

Line 285: "The differential absorption technique is "self-calibrating", thus minimizing potential sources of long-term drifts, which are typically linked to systematic misalignment of the laser beams within the telescope field of view, leading to beam-telescope partial overlap at the bottom of the profiles (below 10 -16 km depending on the systems), and inconsistent signal-to-noise ratio at the top of the profiles (above 45 km). These issues are normally well-controlled, resulting in long-term drifts typically not exceeding 2-3% over several decades, as highlighted by Hubert et al. (2016)."

Line 269 "should resume by March 2026" – hopefully by now it has restarted?

No data later that 2016 can be found on the NDACC website, consulted on May 10th, 2026. Even so, it could not be included in this work.

Lines 180-270 One thing missing from all the instrument discussions though is a statement about the possibility of long-term drifts. There are a large number of instrument specialists among the co-authors who could comment for the different instrument types.

We thank the reviewer for pointing out this omission. Long-term drifts can affect all instrument types and represent a major source of uncertainty in trend estimation (Hubert et al., 2016).

Extensive efforts within the NDACC and associated networks are dedicated to minimizing these artifacts. For instance, Lidar differential absorption is naturally "self-calibrating", minimizing drifts to <2-3% per decade (Hubert et al., 2016). For ozonesondes, homogenization corrects known instrumental and processing changes (Stauffer et al., 2022), though small residual record-specific behaviors may remain (which explains the Legionowo behavior mentioned by the reviewer). FTIR records are highly stable due to constant spectroscopic parameters, though partial columns can be sensitive to Instrumental Line Shape (ILS) or temperature profile degradation (Björklund et al., 2024; Jonas et al., 2026). Umkehr and MWR records also follow strict regular calibration protocols.

However, as undetected residual drifts can still occur, this provides one of the strongest motivations for our methodology. If an individual instrument (like the residual behavior at Legionowo) drifts away from the regional consensus, the BASIC algorithm will statistically down-weight it over time. The drift is therefore translated into a larger posterior uncertainty envelope rather than a biased regional trend. We have added a dedicated paragraph summarizing these points at the end of Section 2:

"When combining long-term ground-based measurements, instrumental long-term drifts represent a major source of uncertainty in trend estimation (Hubert et al., 2016). Within the NDACC and TOAR-II frameworks, extensive protocols are established to minimize these artifacts across techniques.

For ozonesondes, homogenization substantially corrects known instrumental and processing changes. Comparisons with satellite data demonstrate a stability within ±2% for total columns and ±5% for stratospheric profiles (Stauffer et al., 2022), though small residual station-specific biases may remain. For Lidars, the differential absorption technique is intrinsically "self-calibrating". Misalignment or signal-to-noise issues are well-controlled, keeping long-term drifts typically below 2-3% per decade (Hubert et al., 2016). Regarding FTIR, total columns are highly stable due to constant spectroscopic parameters. While partial columns are more sensitive to instrumental line shape (ILS) and temperature profile uncertainties (García et al., 2012), regular cell measurements ensure stability, with recent studies showing partial column drifts mostly contained within 1.5% to 3.5% per decade (Björklund et al., 2024; Jonas et al., 2024, 2026). Finally, for Umkehr records, inherent data normalization and regular calibrations against regional standards minimize the development of long-lasting drifts (Petropavlovskikh et al., 2022; Maillard Barras et al., 2022).

Although these techniques deploy strict quality controls, undetected residual drifts may still occur in individual timeseries. This strongly motivates our use of the BASIC merging methodology (Section 3.2). If an instrument develops a drift, its relative weight within the composite dynamically decreases. BASIC inherently accounts for such anomalies by translating the reduced agreement into an appropriately increased posterior uncertainty envelope rather than a biased regional trend."

Lines 288-315 This is all R and not R-squared, right? I am surprised you don't see any negative correlations?

The given Pearson coefficients are R and not R-squared. Weakly negative correlations were computed but are not visible on the maps since we only display r > 0.4 (see scale in Figs 4 and 5). Weak negative correlations between distant sites in opposite hemispheres can arise from asymmetric dynamical responses to the same large-scale forcings, but their amplitude remains small.

Figure 5 – As earlier, I am intrigued by the fact that there is a high correlation between Zugspitze and the southern hemisphere subtropics. Could you please comment on this? To me this suggest a weakness of the correlation method because I don't think the short-term variability and long-term variability would be caused by the same processes across such distinct latitude bands.

In the upper stratosphere, the variability decreases notably (see Fig. 7, timeseries in UpS). The correlation between Zugspitze and the Southern Hemisphere can be explained by common drivers notably the Solar cycle (long-term), transport with the BDC and QBO (medium term). These forcings are supposed to contribute the most to the signal captured by the Pearson coefficient. Furthermore, our grouping is more robust that what the maps show: We require every member to be two by two correlated with more than our threshold, implying that the elements the furthest apart are still sufficiently correlated. The final groups are not crossing the equator, apart from the equator groups in LS and MS, that stay in a narrow latitude band around it.

Finally, the same large-scale forcings are also the drivers of long-term ozone variability in the upper stratosphere. This is precisely why they are represented as proxies in the LOTUS regression model. The assumption that correlated sites share common variability and common trend drivers is therefore physically consistent in this layer.

Line 392: "The grouping relies on the assumption that sites highly correlated in monthly anomalies share the same dominant large-scale forcing mechanisms — notably QBO, solar cycle, ENSO, and sAOD — which are also the proxies used in the trend analysis. This assumption is supported by the fact that these forcings are hemispheric to global-scale phenomena that dominate interannual ozone variability beyond the annual cycle. Sites responding coherently to these forcings are therefore expected to share the same long-term trend drivers. This is particularly apparent in the upper stratosphere, where

reduced variability and the dominance of the solar cycle and BDC transport explain correlations across wide latitude bands. However, some forcings such as QBO and ENSO respond asymmetrically between hemispheres, meaning that high interannual correlation alone is not sufficient to justify merging across hemispheres. Interannual correlation is therefore used as a necessary but not sufficient condition for grouping, and physical judgment is applied where the statistical criterion alone would lead to physically questionable composites.“

Lines 320-326 If I understand correctly, the method assumes that there is a constant offset caused either by distance on the ground or two different instrument types, but do you have any reason to think it is a constant value over time?

Following what was done in Ball et al. (2017), we align the datasets to one common mean. This will indeed correct only the constant part of an offset. However, we have no reason to think it is a constant value over time. However, this assumption is implied within the homogenization assumption: records that have been homogenized are assumed to be mostly free of time-varying drifts, so that remaining differences reflect spatial representativeness offsets rather than instrumental evolution. The BASIC methodology, with PCA uncertainties, reduces the impact of the residual inhomogeneities by construction.

Lines 342-345 This is an interesting point, because if the outlier is completely real but caused by spatial variability, is it right to exclude it from the mean of the area? Later I notice you use the term "regional consensus".

If the outlier is real and caused by spatial variability, then it means that the representativeness assumption behind the regional grouping fails for that particular month or site. This is indeed a clear limitation of this study. In that case, the value should not be interpreted as a measurement error, but as a local geophysical signal that is not representative of the regional composite.

BASIC does not strictly exclude such values, but down-weights them when they are inconsistent with the regional consensus. The composite is interpreted as a robust estimate of the regional mean, in which values inconsistent with the regional consensus are down-weighted to keep representative local variability. This is what we mean by “regional consensus”: the method favors the signal supported by the mutually correlated records, while increasing the posterior uncertainty when the agreement between instruments is poor.

Line 346 Is this method actually in Rodgers?

The reviewer is correct that BASIC is not described in Rodgers (2000). The citation refers to the general Bayesian inference framework, of which BASIC is an application. We have clarified this by rephrasing line 395:

BASIC (Ball et al., 2017) uses Bayesian inference (as defined in Rodgers, 2000) to determine…

Lines 360-381 This is an extremely helpful discussion and figure, and I commend you for including it.

Thank you, although this remains an illustration, and not a description of the way the MCMC method (Hamiltonian Monte Carlo) works.

Lines 418-427 You should give a specific source for these datasets for reproducibility

All links to the datasets are listed in the supplements, alongside the last date of access.

Lines 421-427 I found these third and fourth bullet points difficult to follow. You don't motivate why you changed from ENSO in LOTUS to the NAO. Could you have used both? Then for Lauder you do use ENSO even though its latitude is comparable to some of the NH areas. Then you say that you used it "with" GLOSSAC sAOD but don't explain what you did here at all. The implication is you don't have an aerosol proxy for the northern hemisphere? There have been significant eruptions in the 2000-2024 time period so this seems odd.

We thank the referee for this comment. We agree that the previous description of the third and fourth proxies was unclear. Following this comment, we reconsidered our proxy selection for the full LOTUS model, using ENSO and AOD (Glossac). This provides a consistent treatment of Northern and Southern Hemispheres. We chose not to include both ENSO and NAO simultaneously in the final model for the same reason of staying close to the LOTUS model. The previous sensitivity discussion comparing NAO and ENSO in Hawaii has therefore been removed. We also clarified the role of GloSSAC sAOD in Sect. 3.3. The sAOD proxy is now included for all regions in order to account for volcanic aerosol perturbations.

In addition to this proxy revision, the BASIC HMC sampling size was increased in the revised analysis (following a comment from referee 1). Therefore, all BASIC composites, trend estimates, uncertainties, adjusted R^2 values, figures and tables were recomputed. The main qualitative conclusions remain unchanged, but the revised analysis generally leads to larger trend uncertainties and therefore to some changes in statistical significance.

For the three selected regions shown in Table 2, BASIC trend uncertainties remain smaller than those of the WM on average, with an average reduction of 9.4 %. The reduction is 15.2 % for Central Europe sPC, 18.6 % for Central Europe aPC, 5.3 % for Hawaii sPC, 8.6 % for Hawaii aPC, 6.3 % for Lauder sPC, and 2.5 % for Lauder aPC.

The main regional changes are as follows. In Central Europe, the LS BASIC trend changes from a weak negative value to a weak positive value and remains non-significant, while the WM LS trend remains negative but loses significance due to the

uncertainty increase. MS and UpS are only slightly modified. For the alternative partial columns, the BASIC UTLS trend is now close to zero and no longer significant, while the aMS trend remains significant for BASIC. The WM loses significance in both UTLS and aMS. This loss of significance in the lower layers, largely driven by the introduction of the sAOD proxy, suggests that part of the previously detected trends was likely aliased by aerosol-driven transient anomalies. Standardizing the LOTUS proxies across all regions therefore yields a more conservative, yet physically robust, trend estimate. The adjusted R^2 values remain generally better for BASIC than for the WM, except in UpS and aUpS, as it was in the original version.

At Hawaii/Mauna Loa, the impact is more visible. For the standard partial columns, the general vertical shape is preserved, but the BASIC trends in the troposphere and UpS are no longer significant, and the WM MS trend also loses significance. This is excepted as ENSO is more relevant than NAO in the tropics. For the alternative partial columns, trend magnitudes are generally smaller and uncertainties larger. The BASIC UTLS trend is now close to zero, whereas the WM UTLS trend becomes negative. The adjusted R^2 values improve for all partial columns in both sets.

At Lauder, the changes are smaller, as expected because the proxy setup was essentially not changed. The changes are mostly due to the prior uncertainty adaptation. The standard partial-column results and adjusted R^2 values are mostly the same. For the alternative partial columns, aUpS remains slightly negative and non-significant, aMS becomes more negative and remains highly significant, and UTLS and aTROPO remain similar to the previous version.

The global maps have also been updated. The revised UpS maps still generally support positive trends over most regions, but not all positive trends are significant; for example, the broad mid-latitude group is now nearly non-significant, and North Canada shows a significant negative trend that is discussed as unreliable. In MS, negative trends are more evenly distributed and are significant mainly over the tropics and Lauder, while most other regions remain non-significant. In LS, the Equator group becomes significantly negative, positive Northern-Hemisphere trends are non-significant and generally remain close to zero, and Lauder remains significantly negative. In UTLS, the previously significant negative Equator trend becomes non-significant; Northern-Hemisphere trends are mostly positive but non-significant, and Lauder remains negative but non-significant.

These revisions were added in Sect. 3.3, in the updated regional trend discussion, in Table 2, in the adjusted R^2 figures, in the global maps and in the Conclusion.

Lines 446-450 (and figure 7) Some of the stations clearly have much larger annual cycles than others and are also display extremes of much greater amplitude. Therefore,

I wonder whether it is really valid to merge them into a 'consensus' timeseries and expect the trends to be physically meaningful?

This comment is addressed in our response to the earlier comment on what the composite represents (Lines 103–107). Records that differ from the group consensus, in annual-cycle amplitude or in extremes, receive inflated uncertainties through the PCA-based uncertainty propagation, and the posterior uncertainty reflects the degree of inter-instrument disagreement. Additionally, working with partial-column instead of more resolved vertical profiles helps reduce differences in annual-cycle amplitude between instruments with different vertical resolutions. The composite represents the common regional ozone signal shared across the group, not a simple average of potentially incompatible records. The extremes in Figure 7 are treated as values outside the regional consensus; whether they arise from measurement outliers or local variability, they are reflected in the inflated uncertainties and down-weighted in BASIC rather than removed by the offset correction.

Lines 451-459 The fact the Legionowo "(and other)" ozonesondes appear first higher than the other instruments and later, lower, seems concerning to me, when the main goal is looking for trends. Does this call into question the validity of combining data from different instrument types? Could you comment on this please.

The Legionowo ozonesondes record, appearing systematically higher than other instruments in 2005–2006 and lower in 2017–2018, is indeed concerning from a trend perspective of the individual instrument, as it could bias towards an artificial negative trend. This is precisely the type of instrumental inconsistency that illustrates the advantages of BASIC over a simple weighted mean. As described in the discussion of Figure 8, BASIC remains centered on the majority of the instruments, with inflated uncertainties reflecting the lack of consensus. The WM, by contrast, is pulled towards Legionowo without any corresponding increase in uncertainty. This asymmetric influence on the WM would directly bias the estimated trend, whereas BASIC's down-weighting of the inconsistent record limits this effect. This does not call into question the validity of combining different instrument types in general, but rather illustrates why an outlier-robust merging method is necessary when doing so.

Line 476-477 You say BASIC can "retrieve the common geophysical signal" – this implies that there exists a "common geophysical signal" and the rest of the variability is noise? Can you expand on this please?

Yes, with a clarification. The 'common geophysical signal' refers to the ozone variability shared across all instruments in the group, driven by large-scale forcings. The remaining variability is not necessarily noise in the physical sense, but it is not representative of the regional signal. By construction of the representativeness study, variability that is not shared by the group is interpreted as instrumental, sampling-related, or below the

regional scale targeted by the composite, and is treated equivalently in BASIC: it inflates the posterior uncertainty while limiting its influence on the composite. We have replaced 'common geophysical signal' with 'the regional ozone variability shared across all instruments in the group' to avoid implying that all non-shared variability is noise.

Line 630: "This agreement confirms that BASIC is robust enough to retrieve the regional ozone variability shared across all instruments in the group, even in the most heterogeneous layer of the atmosphere, with the remaining variability left as unexplained noise."

Lines 515-519 It's very helpful that you've plotted the BASIC results with the weighted mean and the individual stations.

Thank you for your positive feedback.

Lines 545-546 The one sigma is in the darker blue, not lighter, it looks to me.

Yes, thank you. It is fixed.

Lines 545-556 It seems a very big assumption to me that is meaningful to compare trends in such widely diverse locations, even if the monthly correlation is reasonably high.

We agree that trends from individual instruments located in diverse locations should not be interpreted as directly comparable estimates of the same local trend. This is precisely why we do not average the individual trends, but estimate the regional trend from the merged timeseries.

The individual trends shown in Figs. 11–18 are included only as a diagnostic comparison. They illustrate the spread of record-level trend estimates and provide context for the uncertainty reduction obtained with the composite approach. As stated in the manuscript, the trend of a merged composite is not equivalent to the mean of the individual trends, because the merging and trend-estimation operations do not commute.

We have clarified the figure discussion to emphasize that individual trends are shown for context only. The physically interpreted trend is the one estimated from the regional composite, built from mutually correlated records selected through the representativeness analysis.

Lines 579-582 This is another example where ozonesonde seem to be differing from other instrument records. Should you exclude stations suffering from the "drop-off"?

As in Van Malderen et al. (2025) and HEGIFTOM, our study excludes stations suffering from the 'drop-off'. Moreover, the BASIC methodology is designed to account for steps effects. In theory, the ozonesonde should be excluded gradually from the merging, unless the other instruments show similar trends.

Specific comments

Lines 100-101 "data desert in Salawitch" doesn't make sense, please re-word

Done.

With the foreseen reduction of the number of limb-viewing satellite instruments in the coming years (Salawitch et al., 2025), ...

Lines 147-148 Please reword this sentence – "is not expected to contribute" to what?

Corrected:

The diurnal cycle is not expected to contribute to the disparities caused by the different temporal resolutions, given its limited amplitude...

Line 170 In the second table the pressure values don't match in the first column between aMS and UTLS.

Fixed.

Line 289 "Following Weatherhead" would be better wording than "According to Weatherhead".

Done.

Line 322 A better wording would be "following Ball et al ." rather than "following Ball's approach"

Done.

Line 499 "who" -> "which"

Done.

Line 563 For me, the expression "In a nutshell" is too informal for a scientific paper in a journal, I suggest something like "in summary"

Done.