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., 2026

Overall comments

The study by Mirallie et al., 2026 presents a methodologically original and timely adaptation of the BAyeSian Integrated and Consolidated (BASIC) merging framework of Ball et al. (2017) to ground-based partial column ozone records from the NDACC network, deriving regional stratospheric ozone trends for the 2000–2024 period. The work addresses a genuine community need: with the anticipated decline of limb-viewing satellite instruments over the coming years, well-consolidated ground-based composites will become increasingly central to monitoring ozone recovery, and the partial column framework is a sensible and well-reasoned approach to reducing the large noise inherent to individual profile records. The paper is well-structured and clearly written throughout, with a logical flow from data description to methodology and results that makes it accessible to a broad readership. The figures are generally informative and well-chosen, and the inclusion of detailed group composition tables and trend maps in the supplement is a valuable addition that substantially increases the paper's utility as a community reference dataset. The introduction of the alternative partial column set (aPC) to isolate UTLS variability, the correlation-based regional grouping methodology, and the direct comparison with the conventional weighted mean are all meaningful contributions that go meaningfully beyond prior work in this area. I feel the paper contributes significantly to the community and, importantly, the results are largely consistent with the existing literature in the upper stratosphere while offering new regional insight in the lower stratosphere. The significant negative trends found at Lauder and in the European UTLS are inspiring, and the transparency with which the authors discuss known limitations — such as the HILO ozonesonde artifact and the North Canada data quality issues — reflects well on the overall scientific rigour of the manuscript. I recommend this paper for publication with minor revisions. Please address my specific comments below.

We thank the reviewer for the positive assessment. Regarding the significant negative UTLS trends in Central Europe: following a comment on proxy selection, we recomputed the trends using the full LOTUS proxy set (ENSO and sAOD for all groups) instead of NAO for Northern Hemisphere groups. As a result, the significant negative trend in the Central Europe UTLS layer is no longer detected by BASIC, though the WM still reports a negative trend in this layer. The signal is however still captured in the aMS layer, where BASIC reports a significant negative trend of −1.73 ± 1.18 %/decade. The significant negative trends at Lauder are unchanged, as the proxies were already ENSO and sAOD for that group.

Specific comments:

Systematic uncertainties excluded from BASIC weighting (section 2.2.2)

For FTIR retrievals, the paper states that systematic uncertainties (5–7% per partial column from spectroscopy, instrumental line shape, and temperature profile errors) are comparable to or larger than the propagated random uncertainties (~4–6%). Since BASIC only uses random uncertainties for weighting, FTIR instruments are effectively over-weighted relative to their total measurement reliability. The offset removal corrects for mean systematic offsets but not for time-varying systematic biases. This limitation should be acknowledged along with any potential implications.

The measurement uncertainties L1 are a combination of systematic and random uncertainties, as well as smoothing and retrieval uncertainties for remote sensing methods. Thus, the FTIR are weighted with respect to their full errors.

We added clarifications in section 2:

L 196 : "The L1 measurement uncertainties are the total error budget, including random and systematic errors, as well as smoothing and retrieval uncertainties for remote sensing techniques."

Ozonesonde partial-column uncertainties (lines 188–191 vs. lines 245–249)

Ozonesonde partial-column uncertainties are obtained by summing up the individual uncertainties at each pressure level and in the the Dobson Umkehr section (lines 245–249) it is stated explicitly that the root-sum-of-squares method is used. This is an inconsistency between instrument types that should be briefly acknowledged. If random errors dominate for ozonesondes, the summation overestimates uncertainty by roughly $\sqrt{N}$ relative to quadrature, systematically downweighting ozonesondes in BASIC. I believe that the study would benefit by adding even a sentence noting the choice and its direction of effect.

Because the same instrument is used for every layers during a flight, the measurements are not linearly independent. In the worst case, systematic uncertainty is correlated across all individual layers, and the total uncertainty is the sum of all individual

uncertainties. Furthermore, since only up to 4 measurements are available per month for a typical station, the statistical reduction is limited. Details for clarification were added in section 2.2.1, line 244:

"The uncertainties of the partial ozone columns are obtained by summing up the individual uncertainties of the ozone concentration measurements across pressure levels. Since the same instrument measures all layers during a given flight, the measurements at different altitudes are not linearly independent. This linear sum is the worst case scenario, assuming fully correlated systematic uncertainties in the entire vertical profile."

CAMS-based region definitions (lines 273–284)

The spatial correlation structure used to define groups is derived from CAMS EAC4, which assimilates satellite ozone data. The groups therefore reflect how a model and satellite-constrained reanalysis represents ozone coherence, rather than how the ground-based observations themselves covary — a structural dependency that is worth acknowledging given the paper's stated aim of providing an independent ground-based reference. Additionally, CAMS data is available only from 2003 onward (as stated in Line 283), meaning the group definitions extrapolate back to the 2000–2002 portion of the trend period. A brief sentence noting this as a minor limitation would be appropriate.

Following this remark, we added details acknowledging that the groups are mathematically relevant within the validity limits of the CAMS reanalysis and its assimilated data. We mentioned the backward extrapolation (for years 2000-2002) and also the limitation that the reanalysis is not perfectly independent to the timeseries since the CAMS reanalysis include ozonesondes. Details were added in section 3.1 for clarification, line 369:

"However, the trends will be computed for the period 2000-2024, so we extrapolate the correlation for the years 2000-2002. Furthermore, since the CAMS EAC4 reanalysis assimilates satellite and uses ozonesondes data for validation, the spatial merging is somehow biased by this."

Correlations between the oPC aPC in UTLS (lines 279–281)

The CAMS-based correlations are computed only for oPC and reused for aPC on the grounds that the two sets overlap in altitude range as stated by the authors. The supplement's group maps (Figures S7–S8) confirm the same groups are used across both sets. However, I think that the key difference between oPC and aPC lies precisely within the UTLS layer, which is governed by tropopause dynamics quite distinct from the lower stratosphere. I would ask the authors to add a brief discussion acknowledging that the UTLS groupings carry more uncertainty than the stratospheric ones in this regard and what their implications might be.

We agree with the reviewer that the UTLS dynamics, mainly governed by tropopause dynamics, may differ from the correlation structures in LS. Therefore, using the same groups for LS and UTLS bring more uncertainty in the last one. For simplicity, this choice was made, alongside a better comparison between LS and UTLS. We added a larger discussion regarding this limitation in section 3.1.

Line 365: "This introduces uncertainty, particularly in the UTLS layer governed by the tropopause dynamics.

To avoid having the correlations dominated by the seasonal cycle, we computed them using the CAMS anomalies time series of PC ozone monthly means. Since the CAMS data is only available from 2003 to 2024, it was the time range used to compute the correlations. However, the trends will be computed for the period 2000-2024, so we extrapolate the correlation for the years 2000-2002. Further, since the CAMS EAC4 reanalysis assimilates satellite and ozonesondes data, the spatial merging is somehow biased by this."

The BASIC prior (lines 350–354)

The prior is replaced from Ball et al. (2017)'s empirically-derived version with the McPeters and Labow (2012) ML climatology, to ensure independence from the input datasets. However, the ML climatology is itself built from Aura MLS and ozonesonde data, and several ozonesonde stations in this study (e.g., Lauder, Hohenpeissenberg, Boulder) very likely contributed to it — so the independence is not complete. More importantly, the paper acknowledges directly that the climatology "has a very large variability," making the prior essentially uninformative. I would simply ask the authors to tone down the independence claim slightly, acknowledging that the prior plays a minimal constraining role, rather than citing independence as a primary motivation for the change.

We thank the referee for this important comment. We agree that the independence of the ML climatology prior is not complete. The ML climatology is based on Aura MLS and ozonesondes data, and some ozonesondes stations used in this study may also have contributed to it. Therefore, the prior is not fully independent from all observational information used here.

After double-checking the code used to produce Fig. 6, we found that the variability of the ML climatology prior had been incorrectly represented in the figure. The corrected prior is narrower than previously shown and therefore has a more visible constraining role than suggested in the original version. Figure 6, 7, 8, 9 and 10 were updated accordingly, with small differences.

We have therefore revised the corresponding paragraph to clarify that the motivation for using the ML climatology is not to obtain a fully independent prior, but rather to avoid

deriving the prior empirically from the same regional records that are being merged, as in Ball et al. (2017). The prior remains external to the data and provides a climatological month-to-month constraint.

Line 459: "Here, the prior is derived from the ML climatology (McPeters and Labow, 2012), based on Aura MLS and ozonesonde datasets, rather than from the input records themselves. This reduces direct dependence on the merged timeseries, although the prior is not fully independent. The ML prior provides a climatological month-to-month constraint, while remaining broad enough for the posterior to be mainly driven by the observations."

MCMC implementation (line 364)

The posterior is sampled via MCMC but no information is given on number of chains, burn-in length, or convergence criteria. Given that Figure 6c (August 2012) shows an explicitly bimodal likelihood — where convergence to the correct mode is non-trivial — I would ask the authors to add at least one sentence describing what convergence diagnostics were applied.

We thank the referee for pointing out this omission. $\hat{R}$ and ESS convergence diagnostics are now extracted: The number of chains (n+1, where n is the number of instruments merged in the group), the burn-in length (5000, also called warm-up) and the number of samples were added to the manuscript, alongside the convergence diagnostics used. The results were added to the supplements S7. Thank you for your remarks.

Line 474: *"BASIC samples the final posterior distribution P(y|M,d) using a Hamiltonian Monte Carlo (HMC, in the family of Markov Chain Monte Carlo (MCMC) methods), following the code in Ball et al. (2017), with the Stan package (Stan development Team, 2024). The sampling is done on n+1 chains (where n is the number of instruments of the group), with a warmup of 5000 (burn-in length) and a sampling of 20000.*

Two convergence diagnostics were applied:

1. *The rank normalized Gelman-Rubin diagnostic ($\hat{R}$, Vehtari et al., 2021), which quantifies the convergence of the n+1 independent Markov chains by comparing the variance within each chain with the variance between the chains. The value is typically recommended to be less than 1.01.*
2. *The Effective Sample Size (ESS, Vehtari et al., 2021), which estimates the number of i.i.d. draws necessary to obtain the same standard error (since the draws in the MCMC are autocorrelated). The value is typically recommended to be at least 400. "*

Line 504*: "Because the standard HMC struggles to traverse low-probability valleys between isolated modes (Betancourt, 2017), the parallel chains remain localized within their respective initialized peaks. The diagnostics reflect this expected topological*

barrier: for the bimodal August 2012 case, we record an $\hat{R}$ of 1.044 and an ESS of 217.3 in the Lower Stratosphere (LS), and a more severe mode separation in the Upper Troposphere Lower Stratosphere (UTLS) with an $\hat{R}$ of 1.435 and an ESS of 21.8."

PCA uncertainty scaling (section 3.2.1)

The PCA-based multiplicative scaling of individual instrument uncertainties is the paper's key novel methodological contribution relative to Ball et al. (2017), however, Ball et al., (2017) used PCA to directly construct monthly uncertainties (their Eq. 5), whereas here it scales pre-existing propagated random uncertainties. These are fundamentally different operations. I would appreciate having a brief discussion on that in the study; why the scaling is multiplicative, and acknowledge that by down-weighting instruments deviating from the PCA consensus, the method could in principle suppress genuine physical signals captured only by a single high-resolution instrument.

The PCA uncertainty construction is directly inspired by Ball et al. (2017). The L3 uncertainties would naturally be used to assess the weight of each data record in the composite. However, this criterion does not appear to be sufficient for rejecting problematic datasets with low uncertainties and high time resolution. Combining PCA and measurement uncertainties has proved to be an effective solution by letting the PCA approach, combined with BASIC, perform the outlier detection and reduction of importance treatment.

To address your specific points: the combination is performed via multiplicative scaling because the L3 uncertainties generally have a much larger magnitude than the statistical scatter. Multiplicative scaling proportionally inflates the propagated uncertainties of instruments that deviate from the consensus, which allows the Bayesian framework to smoothly down-weight outliers without entirely discarding the underlying data.

Furthermore, we agree with your assessment regarding signal suppression. We have added text acknowledging the theoretical limitation that by down-weighting instruments that deviate from the PCA consensus, the method could in principle suppress genuine physical signals uniquely captured by a single high-resolution instrument.

Line 537:

"This multiplicative scaling is applied because it proportionally inflates the propagated uncertainties of instruments that deviate from the consensus. This allows the Bayesian framework (BASIC) to smoothly down-weight outliers without discarding the underlying data."

Line 543:

"We acknowledge a theoretical limitation to this methodology: by strictly down-weighting instruments that deviate from the PCA consensus, the method could, in principle, suppress genuine physical signals that are uniquely captured by a single high-resolution instrument."

Comparison with the most recent satellite literature (section 4, trend results)

The paper benchmarks its results primarily against Godin-Beekmann et al. (2022) and the WMO 2022 assessment. However, Sofieva et al. (2025) provides an updated multi-satellite LOTUS trend analysis extending to 2024, covering the same trend period as this study. Given that both works use the 2000–2024 period and the LOTUS MLR framework, a brief and direct comparison of the trend magnitudes in the upper and middle stratosphere between this paper's results and Sofieva et al. (2025) would considerably strengthen the results section and help the reader assess the consistency between the two approaches.

We do actually primarily compare our results with Sofieva et al. 2025, alongside Godin-Beekmann et al. (2022).

Line 651: "and Sofieva et al. (2025) for the period 2000-2024"

§ L. 1057: "*Sofieva et al. (2025)"* is cited 4 times.

§ L. 1183: "*Sofieva et al. (2025)"* is cited 5 times.

Line 1288: *"A trend distinction emerges in the LS compared to Sofieva et al. (2025)."*

oPC vs. aPC results presentation in section 4

The results are presented layer by layer across both PC sets, which leads to some repetition and makes it difficult for the reader to grasp what the aPC definition actually adds over oPC in terms of trend detectability. I would encourage the authors to include a brief summary paragraph (maybe even a concise tabular overview) — at the end of Section 4 that directly contrasts the key trend differences between oPC and aPC for the groups where the distinction matters most (e.g., Lauder LS vs. UTLS, European LS vs. UTLS). This would make the value of the aPC contribution considerably clearer.

We thank the reviewer for this suggestion. We have added a paragraph after the table at the end of Section 4.1 to compare sPC and aPC results for the three selected groups. The key message is that the two frameworks reconstruct a physically consistent vertical trend profile rather than systematically improving detectability. The one case where aPC generally improves detectability is Central Europe MS vs. aMS (non-significant 0.03 ± 0.91 vs. significant −1.59 ± 1.11 %/decade, BASIC), where the aMS avoids tropopause contamination identified in the residual diagnostics.

Line 672: "Table 2 allows a direct comparison between the sPC and aPC trend estimates. The two sets allow to provide a vertically coherent picture of the trend profile.

For Central Europe, the MS trend of 0.05 ± 0.92 %/decade (BASIC) is non-significant, while the aMS yields −1.73 ± 1.18 %/decade (significant at the 2σ level) suggesting that the MS layer is contaminated by dynamical noise near the tropopause, consistent with the residual diagnostics discussed in Section 3. For Lauder and Hawaii, both MS and aMS are significant, but the aMS uncertainties are larger, given the lower altitude of the aMS layer. From LS to UTLS, the UTLS systematically shows larger uncertainties and smaller trend magnitudes than the LS, as the inclusion of the upper troposphere dilutes the stratospheric signal. Taken together, the sPC and aPC layers reconstruct a physically consistent vertical trend profile: positive in the upper stratosphere, negative in the middle and lower stratosphere, and positive in the troposphere, the expected signature of stratospheric ozone recovery."