Response to Reviewer 1

We thank the reviewer for thorough reading of our manuscript and providing valuable feedback. The comments and remarks have aided in improving the analysis and manuscript. We understand the main concerns are that (i) the impact of STE on tropopauses and O_3 variability as tropopause-evaluation metric is not sufficiently addressed, (ii) the cross-tropopause gradient is not suitable as stand-alone assessment tool of tropopause-sharpness and (iii) the context of the case study flights is not discussed in enough detail.

In the following, we address all comments in detail (Reviewer's comments in italic, quotations of the corresponding revised text passages in blue).

Based on suggestions from all reviewers, we have added a new Section 3.4 titled "Key differences in tropopause definition methodologies". Additionally, the quantification of tropopause-sharpness has been changed from a calculation of the cross-tropopause gradient to the curvature at the tropopause.

General Comments

(i) **Impact of STE on O_3 variability as a metric**

On utilizing ozone variability as an assessment tool for reliable discrimination of troposphere and stratosphere observations (i.e., Section 4.2): I'm not convinced this approach is a reliable assessment of tropopause definition performance. One particularly problematic condition to consider (which is not discussed at length in the present draft) is stratosphere-troposphere exchange (STE). Recent STE will undoubtedly result in air that chemically appears to be stratospheric/tropospheric but is located in the troposphere/stratosphere (respectively). Thus, using variability as an indicator of performance/appropriateness would seem to reward definitions that "remove" or otherwise minimize the inclusion of STE events. This complexity should at the very least be acknowledged and considered further in the interpretation of the meaning/significance of these results.

We agree that a more thorough discussion of the differences between tropopause definitions that are due to different calculation approaches should have been presented more clearly. Generally speaking, this paper aims to highlight the (dis-)advantages and sensitivities of the different definitions without referencing specific meteorological situations. We compare three main types of tropopauses, which are being used in literature for varieties of studies. While the chemical tropopauses are more representative of an air mass' origin and thus represent recent STE / tropopause folds with an emphasis on what the current chemical composition is representative of, other tropopause definitions put the focus on the location within large structures (aka representing the tropopause through the means of global almost-continuous surfaces).

We chose to use the variability of ozone relative to the tropopause as a measure of how well the natural variability of ozone can be disentangled from dynamic processes in the atmosphere. As discussed e.g. in Millán et al. (2023), separating these effects from each other is necessary to calculate for example long-term trends in ozone. Keeping this in mind, we cannot say that one single tropopause definition will be better for all types of studies. We merely aim to increase the knowledge base of benefits and limitations of the tropopauses studies in the context of aiming to homogenise air masses over long spatial and temporal scales.

We have added some discussion of STE in the newly added Sect. 3.4 Key differences in tropopause definition methodologies.

(ii) **Cross-tropopause gradient**
On the cross-tropopause gradient (i.e., Figure 7): using the magnitude of the ozone gradient across the tropopause as an assessment tool also seems problematic to me. One could achieve the largest gradient by simply introducing a high bias in tropopause altitude such that the ubiquitous sharp stratospheric increase in ozone contributes entirely to the diagnosed cross-tropopause gradient. I think the alternative approach of comparing (or rather, differencing) the gradients/slopes in ozone vs. altitude BELOW and ABOVE tropopause would be a more meaningful and reliable assessment of the appropriateness of each definition. That is nearly accomplished here, but this slight adjustment in approach would resolve an otherwise misleading means of assessing performance.

Taking into account comments from both reviewers, we have revised this part of the analysis to focus on the curvature of the vertical profile at the tropopause to better highlight the "sharpness" of the transition. However, we would like to clarify the use of the normalised tropopause gradient in the analysis: 1) The high bias in the gradient for tropopauses at a higher average ozone value was taken into account through the normalisation to the average tropopause value. As could be seen in the previous Appendix B1, the tropopause averages varied from 67-160 ppb. For tropopause definitions with a high bias, this normalisation factor also increases, which in turn reduces the normalised gradient. 2) Additionally, while admittedly not the focus of the analysis, we make reference to the difference between the gradients at the tropopause as well as 0.5 km below in the text. We agree that the non-normalised cross-tropopause gradient should not be used as a stand-alone assessment tool.

We address this comment through the revision of Sect. 4.3 (formerly "Vertical profiles and cross-tropopause gradient", now "Vertical profiles and curvature at the tropopause"). The gradient evaluation has been replaced by an assessment of the curvature at the tropopause, which we believe to be a stronger metric for evaluating the transition happening at the tropopause. For this, we describe the region ± 2 km around the tropopause with a fit and evaluate the curvature of this fit for each season. A larger curvature then represents a sharper transition between the characteristic tropospheric and stratospheric parts of the vertical profile of ozone.

(iii) **Case study** *On the case study analyses in Section 4.4: this analysis was very brief and I found to be minimally convincing. In particular, where are the observations contextually, especially those in Figure 8? Is the high ozone portion of the flight flagged as tropospheric in panels (e) and (f) of Figure 8 the result of an STE event? For example, could this be a tropopause fold? To convincingly demonstrate appropriateness of these definitions for a case study, more information should be given. Though tracer-tracer analysis was mentioned early, I thought its exclusion from this study was a missed opportunity.*

Thank you, this comment poses some very interesting questions. We would like to again highlight that the focus of this paper does not lie in specifically exploring features of the atmosphere, but in showing differences between tropopause definitions. In order to briefly showcase the potential impact, we present case studies of generally non-specific meteorological scenarios. The benefits and limitations of in-situ calculable versus modelled tropopause parameters is visualised through these flights and later discussed in more detail.

For more context, you may refer to Fig. 1 of this response showing curtains of potential vorticity (PV), temperature and potential temperature from ERA5 reanalysis data combined with the interpolated modelled tropopause pressures. In the left panel, the stop-over in Reykjavík commencing around 12:00 corresponds to a sharp descent and corresponding crossing of the tropopause. Due to changes in PV, temperature and potential temperature, the modelled dynamic and thermal tropopause show some unusual features in this region, leading to the differences between definitions described. We have added the following information in Lines 474–478 to provide the reader with more context on the case studies:

The first case study flight happened throughout the day on 21 August, including a stop in Reykjavík. While most of the flight takes place at pressure levels below 200 hPa and thus mostly in the stratosphere, tropopause crossings happen for all definitions during the ascents and descents. The second case study flight starts in the evening of 22 September until the morning of 23 September. The aircraft stayed well above all modelled tropopauses during the flight, except for ascent and descent.

Technical Edits

Line 9: "larger variability near the tropopause" would be better stated as "larger composition variability near the tropopause"
Thank you, the wording has been changed as suggested in Line 9.

Line 107: "substances" should be "substance"
Thank you, this has been changed in Line 111.

Line 151: "across the tropopause" would be better stated as "from troposphere to stratosphere in the extratropics"
We agree, the revised phrasing has been implemented in Lines 156–157.
Due to stratification above the tropopause, PV experiences a strong increase from troposphere to stratosphere in the extra-tropics, allowing for the definition of a *dynamic* tropopause.

Line 153: "the vast majority of" could be stated simply as "most"
Thank you for the hint, the sentence has been changed accordingly in Line 158.

Line 179: suggest revising "last decades" to "last several decades"
Thank you, we have decided to remove that part of the sentence in Line 185.

Line 304: I find the statement "a large number of" to be somewhat misleading. What is considered large? Could you express this as a fraction of all aberrations? From my interpretation of the analysis, it appears to have a minimal impact on the statistics, so that gives the impression that these observations are still few though certainly more numerous than remaining definitions.
We agree this should have been quantified better. The following sentence has been added in Lines 353–354 to expand on this point:

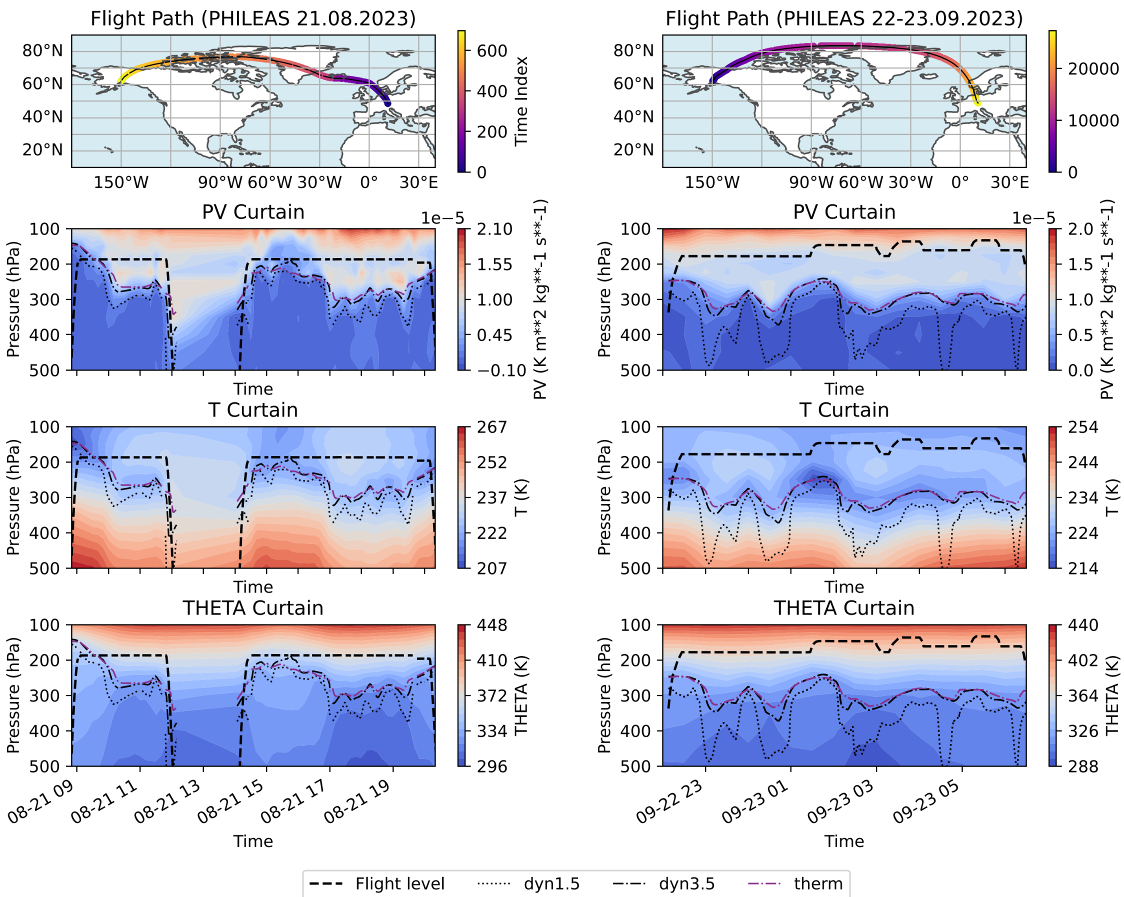

Figure 1: ERA5 reanalysis curtains interpolated onto the flight track of the case study HALO-PHILEAS flights on 21 Aug 2023 and 22-23 Sep 2023. Potential vorticity (PV), temperature (T) and potential temperature (THETA) are shown for the pressure range 500-100 hPa on a vertical resolution of ... and a temporal resolution of 30 minutes.

Note that over 4% of the points sorted into the troposphere by the thermal tropopause have ozone mixing ratios over 200 ppb, compared to 1.5% for the $\text{dynamic}_{3.5}$ tropopause and less than 0.1% for all other definitions.

Line 313: "stratosphere" should be "stratospheric"
Thank you, this has been changed in Line 366.

Lines 322-323: this sentence also appears to be an overstatement. The variability in ozone is not largest for the thermal tropopause in all seasons, though the range appears to be. In particular, the mode is largest for the N2O definition in autumn and winter.
Indeed, we have changed this sentence in Lines 376–378 to better represent the results: For all seasons, the spread of tropospheric bin variabilities of ozone is largest with the thermal tropopause with corresponding high modes in spring and summer, although in autumn and winter the chemical_{N_2O} tropopause results in a higher variability mode.

Line 338: delete floating paren after "Fig. 6b"
Thank you, we have removed this for Line 393.

References

Millán, L. F., Manney, G. L., Boenisch, H., Hegglin, M. I., Hoor, P., Kunkel, D., Leblanc, T., Petropavlovskikh, I., Walker, K., Wargan, K., and Zahn, A.: Multi-Parameter Dynamical Diagnostics for Upper Tropospheric and Lower Stratospheric Studies, Atmospheric Measurement Techniques, 16, 2957–2988, https://doi.org/10.5194/amt-16-2957-2023, 2023.