We thank the reviewer for his/her comments. Below are our responses in blue.

The main changes were:

- We added the main conclusion to the abstract
- We modified the symbol size for the ozonesondeds and lidar in figure 1
- We added a rationale for the period and datasets used.
- We added a rationale for the usage of "zonal" means.
- We added a brief discussion of the WMO multiple tropopauses and tropopause breaks
- We added several references as suggested by the reviewers.

Reviewer 4

Review of « Exploring ozone variability in the upper troposphere and lower stratosphere using dynamical coordinates » by Millan et al., submitted to ACP

This manuscript aims to assess the usefulness of different transport-relevant coordinate systems (altitude, pressure, potential temperature, equivalent latitude, distance to the subtropical jet and distance to the tropopause) for dividing the measurements into bins affected by different atmospheric regimes. Then, the overall objective is to combine measurements from different platforms with different sampling characteristics for assessing the ozone trends and attributing them to changing atmospheric dynamics. This study is definitely an important milestone of the SPARC OCTAV-UTLS activity and follows a previous analysis by Millan et al, AMT 2023 (with almost the same co-authors), which was dedicated to the presentation of such dynamical diagnostics to describe the meteorological context for multi-decadal observations in the UTLS by ozonesondes, lidars, aircraft, and satellite.

The manuscript is well organized and well written. The figures are good and support the analysis. I recommend the publication after addressing some comments and suggestions, in order to improve its impact and make it useful for other data sets.

General comments :

In order to clarify some aspects of the methodology, of the results and therefore increase the impact (the usefulness) of these dynamical coordinates, I propose the following suggestions.

1. The major comment concerns the improvement in providing a further clarification of the objectives and of the results. This manuscript should better address the complementarity and/or the difference with the previous Millan et al., published in AMT in 2023, in the introduction.

After careful consideration, we have decided not to mention Millan et al., 2023 in the introduction. As correctly pointed out by the reviewer in the first paragraph of his/her review, this paper primarily focuses on the computation of dynamical diagnostics. In other words, Millan et al., 2023 is a technically oriented paper that describes the algorithms used to identify jets, tropopauses, and EqL at the measurement locations. Therefore, we believe it is more appropriate to mention it in section 2.2, where we describe the methods used to characterize the jets and tropopauses. The introduction primarily focuses on the importance of UTLS ozone and the introduction of coordinate systems. We feel that discussing the intricacies of computing dynamical diagnostics is not suitable for this section.

> At the end, the reader misses a clear opinion on the advantages of using these coordinates and a further understanding of the ozone variability in the UTLS, or at least a further discussion on the gain in consistency within the different data sets.

We believe the bullets in the summary section clearly summarize the impact on the relative standard deviation observed when using different coordinate systems. To further emphasize our main conclusion, we have added the following to the abstract (as suggested by multiple reviewers): “Overall, the use of equivalent latitude-potential temperature leads to the most substantial reduction in binned variability across the UTLS. This coordinate pairing uses PV on isentropic surfaces thus following the transport of tracers in adiabatic frictionless flow.”

> It is quite frustrating to read that interesting results will be published in two future studies without giving more information in this one.

We understand the reviewer’s frustration, however one paper cannot encompass all the material that OCTAV-UTLS will explore. The current manuscript already includes 11 figures in the main manuscript and 9 figures in the appendix, which we believe are essential to sustain our conclusions. Furthermore, many of these figures have up to 33 panels. We believe a longer paper that encompass other topics would lack a clear focus and would not do justice to the individual topics included.

> The last sentence of this manuscript “Another study will analyze how differences in sampling patterns and resolution (both vertical and horizontal) can affect the representation of the datasets as well as the trend quantification” is giving the negative impression that this manuscript is not going far enough to be really useful. It reveals that differences in sampling patterns and resolution are not addressed here. Therefore, the conclusions are somewhat weakened.

In the manuscript, we argue that despite variations in sampling and resolution across datasets, the conclusions drawn remain consistent across all datasets. The physical reason for this is that assuming a tracer conservation on the time scale of jet- and tropopause dynamics in the UTLS, the tracer relationships on isentropes are conserved for adiabatic motions (namely, “This coordinate pairing uses PV on isentropic surfaces thus following the transport of tracers in adiabatic frictionless flow.”). Notably, this is independent from sampling density and location as long as adiabatic conditions are not violated.

Thus, these conclusions are robust since the consistency across the datasets (despite their differences) reinforces the validity and reliability of the findings.

The sentences that mention this in the manuscript are:

We examine the effects of different coordinate systems on the representation of geophysical variability in UTLS ozone through production of climatologies from the datasets outlined in Section 2.1. Because the variability in these climatologies is also influenced by sampling and measurement characteristics, the use of multiple datasets allows exploration of the commonalities among differences in climatologies as a function of coordinate system for each instrument. Any common changes between coordinate systems are assumed to result from a change in the representation of the effects of geophysical variability.

To further emphasize this point we added in the last paragraph of the manuscript:

In this study we identified coordinate systems that most help to reduce binned variability over broad regions in an effort to facilitate more robust UTLS composition trend analyses. **The use of multiple datasets with different sampling and resolutions enables us to identify commonalities among them, ensuring conclusions that are independent of the specific measurement techniques. We are aware that several questions regarding the binned variability are still open and some of them will be addressed in upcoming studies.** For example, a future OCTAV-UTLS study will ...

2. Regarding the data sets used in this analysis, it is important to further argue about the selection of these data sets. Why are there not the same as in Millan et al., 2023 ? Why is IAGOS-CORE not used here in addition to IAGOS-CARIBIC ? Why is the number of ozonesondes so limited ? Why not use the ones from the SHADOZ network with the advantage of sampling the tropical regions ?

Although, we would ideally like to include all datasets available, the reality is that time, computing time, and even funding precludes us using all available ozone datasets. In fact, we considered using IAGOS-CORE in the beginning of the OCTAV-UTLS formulation, but due to the lack of resources we were not able to pursue it at that time. However, we are constantly increasing the datasets included in OCTAV. To address this, we have added at the end of the dataset section:

The datasets used in this study are not intended to be comprehensive; numerous other ozone records are available. For example, limb scattering satellite sounders, such as the

Optical Spectrograph and Infrared Imager System (OSIRIS; Llewellyn et al., 2004) or the Ozone Mapping and Profiler Suite (OMPS; Seftor et al., 2014), the long term airborne measurements from IAGOS-CORE (Petzold et al., 2015), and the ozonesondes included in the Southern Hemisphere ADditional OZonesondes (SHADOZ; Witte et al., 2017; Thompson et al., 2017). However, the records included are representative of the currently available measurement techniques in terms of resolution and geophysical sampling of the UTLS.

> My suggestion would be to consider adding, as a result of this analysis, a list of recommendations for using such dynamical coordinates with other datasets. That would be valuable for the scientific community focused on providing ozone data sets and would increase the impact of this study.

To further emphasize the coordinate system that overall reduces the binned variability, we have added in the abstract: "Overall, the use of equivalent latitude-potential temperature leads to the most substantial reduction in binned variability across the UTLS. This coordinate pairing uses PV on isentropic surfaces thus following the transport of tracers in adiabatic frictionless flow."

We have also added in the summary section: "These conclusions were drawn using a variety of ozone measurements (i.e., ozonesondes, lidars, and satellite and in-situ aircraft measurements) with a plethora of vertical and horizontal resolutions, as well as sampling characteristics. Therefore, we anticipate that these results are applicable to other datasets not included in this study, such as OMPS, OSIRIS, IAGOS-CORE, and additional ozonesondes and lidar data available elsewhere."

3. Regarding the sampling patterns, the manuscript would be improved by adding a discussion on the impact (or not) of the differences in measurements locations. This comment is indeed linked to the one on the selection of the used data sets. MLS and ACE data sets are clearly "global" data sets but all the others are not and cannot be considered "symmetric zonally" like the satellite data sets. The sondes and lidars used in this study are only or mostly located in the "Western Hemisphere", on contrary of the CARIBIC aircraft data sets which spans a wider range of longitudes. What is the impact of such differences in discussing consistency in terms of zonal averages?

As mentioned in our response to question 1, we argue that despite variations in sampling and resolution across datasets, the conclusions drawn remain consistent across all datasets and hence they are robust through all available datasets.

Also a brief discussion or a simple pedagogic explanation on the use of zonal averages for presenting these transport-relevant coordinates would be a valuable addition to the manuscript. A few comments in the manuscript mention some characteristics varying with longitudes (e.g., double tropopause, strength and sharpness of the subtropical jet). These differences in the representativeness of the different data sets should be addressed or the differences (if any) between the two hemispheres (western vs eastern) should be discussed using the data sets providing the full range of longitudes. For example, Cohen et al., 2028, showed that the IAGOS-CORE data sets have different levels of ozone between the North America – Atlantic and the Eurasian sectors in the UTLS, when the tropopause is defined as the 2 PVU. Providing zonal averages have clear advantages, but when it comes to reducing and analysing the ozone variabilities, this longitudinal dependence has to be clearly discussed.

We have added in the Coordinate mapping section:

For this initial study, we use averages over all longitudes with different horizontal coordinates, similar to zonal means when using latitude. However, many dynamical and chemical processes exhibit significant longitudinal variations. Consequently, as mentioned in the introduction, coordinates that are most helpful to study geophysical and transport properties may vary depending on the region or phenomenon of interest.

4. Regarding the sampling period: What is the rationale to cover the 2005-2018 period while some of the data sets (i.e. aircraft) cover only a few years, and not the same for all (according to Table 1) ? A further discussion on this choice and on the impact (or not) of merging data sets from 2008 with those from 2015-2016 and those apparently equally distributed over the long 2005-2018 period would be valuable to add.

To clarify our rationale, we added at the beginning of section 2.2.2:

We chose this period due to the current availability of dynamical diagnostics (discussed in section 2.2.1), which require significant computing time to generate. This period allows for ample overlap among all measurement techniques used here, i.e., ozonesondes, lidars, aircraft in-situ campaigns, and limb sounders. While the aircraft in-situ measurements from PGS, TACTS/ESMVAL, and START08 do not cover the entire period, we include them to enhance the coverage of this measurement technique. However, it's worth noting that the bulk of the variability is driven, in the aircraft results, by the overwhelming quantity of CARIBIC-2 measurements.

5. Regarding the differences in the vertical resolution: What is the impact (or not) of different vertical resolutions, among the data sets themselves (i.e. from 3 km for MLS to 100 m for the ozonesondes and probably less for aircraft) and with the vertical spacing of theMERRA-2 products (1.2 km) ? I recommend that table 1 includes the information on the vertical resolution and for aircraft, the detailed “Range” instead of flight levels which is not very informative as research aircraft may fly higher or lower than passenger aircraft.

Again, we are precisely exploiting such differences to draw robust conclusions. As requested by the reviewer we tried to add the vertical resolution information to table 1 but were unsuccessful, there is not enough horizontal space to add another column. We have added the typical “range” of the flights as a note in table 1.

Specific comments:

1. The abstract should better highlight the main findings by adding a few sentences from the Summary section.

We added at the end of the abstract: Overall, the use of equivalent latitude-potential temperature leads to the most substantial reduction in binned variability across the UTLS. This coordinate pairing uses PV on isentropic surfaces thus following the transport of tracers in adiabatic frictionless flow.

2. Line 120 : A more general and recent reference to IAGOS should be added here, or at least the web site, e.g. htpp://www.iagos.org; Petzold, et al., 2015; Thouret et al., 2022

We have added the recommended references

3. Line 155 : Can you further explain this gridding ? Is such a 100 m gridding to reduce computing power applied to other data sets ? It is quite surprising as the ozone data set is probably not the “heaviest”. In general, it would be nice to have the same types of details in all sections describing the different data sets.

When computing the dynamical diagnostics, the algorithm interpolates the reanalysis fields to the measurement locations. Many years ago, in one of the first OCTAV-UTLS meetings, it was decided that we were going to reduce the fine resolution measurements to avoid needless interpolation of the reanalysis fields (which typically have about 1-km vertical spacing in the UTLS) to very fine grids of, say, 5 meters (as the ozonesondes) where the information they supply would be redundant. We have modified the manuscript to:

In this study, ozonesondes were gridded to 100 m to reduce computing power when calculating the dynamical diagnostics (see Section 2.2). **It is important to note that this gridding resolution has no impact on the study's results, as the reanalysis fields only contain information at about 1-km vertical spacing and measurements will be averaged together in approximately 1-km bins.**

4. Line 158 : year is missing in the reference Smit and Thompson, as well as in the references list, line 607. It has been published in 2021.

Corrected

5. Lines 164-166 : The question is then "why not using more ozone sondes stations in this analysis ?"

We have added at the end of the dataset section:

The datasets used in this study are not intended to be comprehensive; numerous other ozone records are available. For example, limb scattering satellite sounders, such as the Optical Spectrograph and Infrared Imager System (OSIRIS; Llewellyn et al., 2004) or the Ozone Mapping and Profiler Suite (OMPS; Seftor et al., 2014), the long term airborne measurements from IAGOS-CORE (Petzold et al., 2015), and the ozonesondes included in the Southern Hemisphere ADditional OZonesondes (SHADOZ)(Witte et al., 2017; Thompson et al., 2017). However, the records included are representative of the currently available measurement techniques in terms of resolution and geophysical sampling of the UTLS.

Lines 204-206 : Further details are necessary here regarding the vertical resolutions, from the native ones to the gridded ones.

The native vertical resolutions are explained in the paragraphs corresponding to each dataset, below copied and pasted from the manuscript below:

Aura MLS: The MLS ozone vertical resolution in the UTLS is around ~ 3 km.

ACE-FTS: These measurements achieve an effective vertical resolution of around 1 km in the UTLS region due to vertical oversampling (Hegglin et al., 2008).

Airborne in-situ instruments: However, CARIBIC-2 aircraft operate at cruising altitudes of 10–13 km, near the climatological location of the extratropical tropopause. The high temporal and horizontal sampling of CARIBIC-2 provides a very detailed view of the tropopause and a very long time series (starting in 1997). In contrast, the other aircraft missions studied here, START08, PGS, and TACTS/ESMVal, have more limited regional and temporal coverage, but provide more extensive vertical coverage of the UTLS, making them ideal for process-oriented studies.

Lidars: Most lidars achieve high vertical resolution, on the order of less than 1 km.

Ozonesondes: In this study, ozonesondes were gridded to 100 m to reduce computing power when calculating the dynamical diagnostics (see Section 2.2).

6. Line 220 : “than” should be replaced by “that”.

Corrected

7. Line 223-224 : Would it be because the range of sampled longitudes with these data sets is restricted to the “western hemisphere” (se also General Comment #3) ?

This is just a consequence of the sparse sampling of the ozonesondes and lidars used, it has nothing to do with the actual ozonesonde or lidar technique. The paragraph explaining this has been left unchanged.

References:

Cohen, Y., Petetin, H., Thouret, V., Marécal, V., Josse, B., Clark, H., Sauvage, B., Fontaine, A., Athier, G., Blot, R., Boulanger, D., Cousin, J.-M., and Nédélec, P.: Climatology and long-term evolution of ozone and carbon monoxide in the upper troposphere–lower stratosphere (UTLS) at northern midlatitudes, as seen by IAGOS from 1995 to 2013, Atmos. Chem. Phys., 18, 5415–5453, https://doi.org/10.5194/acp-18-5415-2018, 2018.

Petzold, A., Thouret, V., Gerbig, C., Zahn, A., Brenninkmeijer, C.A.M., Gallagher, M., Hermann, M., Pontaud, M., Ziereis, H., Boulanger, D., Marshall, J., Nédélec, P., Smit, H.G.J., Friess, U., Flaud, J.-M., Wahner, A., Cammas, J.-P., Volz-Thomas, A. & Team, I. (2015). Global-scale atmosphere monitoring by in-service aircraft – current achievements and future prospects of the European Research Infrastructure IAGOS. (Vol. 67, pp. 28452). https://doi.org/10.3402/tellusb.v67.28452

Thouret, V., Clark, H., Petzold, A., Nédélec, P. & Zahn, A. (2022). IAGOS: Monitoring atmospheric composition for air quality and climate by passenger aircraft. (pp. 1-14). https://doi.org/10.1007/978-981-15-2527-8_57-1