Review of “Signs of climate variability in double tropopause global distribution from radio occultation data” by Alejandro de la Torre et al.

The authors present a complex statistical methodology for the analysis of double tropopause (DT) variability. The DT topic is not extensively researched, and the presented methodology has potential to provide valuable results, so the general idea of the study is a welcome one. However, especially in section 3, my impression is that not enough testing and optimization of the method has been performed, which limits the result’s performance and interpretability in later sections.

The statistical side of the work is presented in a very detailed way. However the discussion of results regarding tropopause dynamics and previous publications is nearly absent from the manuscript. Also I have concerns about the setup of the clustering analysis which serves as the base for the multivariate linear regression later. All together, I feel the present manuscript is still a long way from being fit for publication.

Although my recommendation is ‘reject’ for the manuscript in its current form, as I state above the method has good potential and I would encourage resubmission once the extensive list of issues is rigorously addressed.

#

 # Major comments

#

M1.1: There’s a lack of relation and discussion of processes responsible for DT, or DT types, with the clusters found in your analysis.

M1.2: The clustering analysis is done with very basic parameters, I am not sure they are the best choice.
→ Using standard deviation of NDT’ blends all modes and timescales of variability together, this compounds with my comment M1.1 and makes interpretation of the results very, very difficult, if one wants to go further than just showing the statistics.
→ See my “**general comment on section 3**” for suggestions on this issue.

M2: Sometimes I noticed the authors do not justify some parameter choices in their analysis, simply stating that the result is ‘reasonable’, which is not sufficient in my view.
→ In other places, the authors present some results in an exaggerately positive way, while to me they are unconvincing or even cause for concern in one instance.
→ I marked the most glaring examples of this with “**(M2)**” in the individual comments below

M3: No interpretation/discussion is provided, whatsoever, for results in section 3 in terms of STE, atmospheric dynamics or previous DT literature. For sections 4-6, this is limited to a couple of paragraphs at most.
→ On the other hand, explanations on some statistical methods are overly long and include sometimes unnecessary definitions of standard statistical measures.

Please find individual comments and suggestions for each section below:

#
Abstract
#

2nd half of the abstract has too much technical jargon on cluster analysis and linear fitting.

No main results or conclusions are highlighted. E.g.:
“the most relevant climatic indices for the distribution of NDT' are identified.” → Which ones?!

Whereas the authors state that the main focus of the manuscript is the methodological approach, main results and potential applicability should be highlighted from the beginning.

#
Introduction
#

l. 57-58: *“and detected in cloud-top inversion layers (Biondi et al., 2012)”* → shouldn’t this be considered as an artifact of the lapse-rate definition of multiple tropopauses?

l. 71-74: feels very vague, and big data are not used anywhere in your manuscript.

#
Sect. 2
#

l. 106: **Foelsche et al. (2011)** missing from reference list, same with **Angerer et al. (2017).** Please check throughout manuscript for missing items in reference list.

l. 109: why do you start at 2006 and not 2001? Should state the step-change in RO data amount from the start of COSMIC

l. 110: DT percentage or frequency has been shown in previous studies on multiple tropopauses, see e.g. references from section 4.1 in **Wilhelmsen et al. (2020).** Please cite them along, the most relevant ones.
l. 110: also, Wilhelmsen didn’t invent the lapse-rate tropopause definition, please cite the WMO definition that the algorithm uses.

l. 110: perhaps this is described in the Wilhelmsen or Angerer studies, but what happens with the RO profiles where the tropopause cannot be found? I know from experience there’s always a small percentage of those, please give some number on the discarded profiles.

l. 115: Wilhelmsen used 5x5 degrees, so this difference should be pointed out since you don’t follow their horizontal resolution.

→ Please don’t use the label N2/N1 in figure labels. “DT frequency” or $DT_{\mathbf{freq}}$ is much more reader-friendly and intuitive than N2/N1.
I even kindly suggest to use $DT_{\mathbf{freq}}$ as a substitute for NDT throughout the manuscript.

l. 121: *"complex pattern with a prevailing temporal variability that depends essentially on the latitude"* →can't really say anything from Fig. 1 in the current form, and of course, there are different variability modes at different latitude bands.

Figure 1: not publication-worthy
→ Please substitute the y-axis for Latitude, and plot DT frequency as color shading, this is the proper way of visualizing the same information.
→ Also format month number into "YYYY", perhaps label every two years.

#
Sect. 3
#

l. 139-140: *"The mean values of the NDT time series and the standard deviations of the NDT' time series are then used for the clustering."*
→ How do the authors justify these settings, why are these the optimal variables for clustering?
→ Also, the authors should describe a bit what these variables represent e.g. in terms of STE, otherwise for many readers this may seem as just a statistical exercise with the most basic parameters of the NDT distribution.

l. 165: about the cutoff distance of 0.07, perhaps some sensitivity experiments with +- 5% or 10% of that value would be reassuring, if shown in a supplement.
Also, please state in this paragraph, what is the usual range of cutoff distances in CA in general.
(M2)

Fig. 3 caption: Does each cluster have the same color in Figs. 2 and 3? Please specify this.

State somewhere in the text at the beginning of section 3 that the 'arbitrary' color scheme from Fig. 2 will be the same in all plots.

l. 175: *"a reasonable separation of objects is achieved"*.
→ This judgment is based on what property of the plot?
→ Also I disagree with the statement, Fig. 3 rather looks like a continuous scatterplot.
(M2)

Fig. 3: → Please show the corresponding probability density estimates of this scatterplot, maybe there are relative density maxima corresponding to some clusters. But from the current plot one can't say.

Figs. 2, 4, 5, and 6: I strongly suggest to number the clusters according to the distributions in Fig. 3, from left to right. In Fig. 2, please add the corresponding cluster numbers which are missing.

Fig. 6 and corresponding text:
→ Discussion missing, please at least discuss the different clusters in relation to different high DT frequency regions from e.g. **Wilhelmsen et al. (2020).**

→ In my opinion, describing the clusters as symmetrically distributed relative to the equator has little meaning: they are related to the subtropical jets on both hemispheres, so sure this will look somewhat symmetric, but it's the relation to the jets that has interpretative value.
→ Fig. 6 looks quite coarse, since it works with monthly timeseries, a refinement to 5x5 degree or better is possible from RO coverage, and would be very welcome.

l. 204-205: *"The interconnected nature of each sub-region is highlighted in the polar, sub-polar and equatorial regions by clusters 1 and 4."*
(M2)
→ I'm sorry to put it this way, but to me this sentence is quite euphemistic. What I see is the method's weakness: equatorial and polar DT's have very little in common, yet the clustering method is mixing them up.
→ Clusters 1 and 4 are next to each other in Fig. 3, and their spatial distribution in Fig. 6 has no distinct structure, it makes me doubt about the method's ability to separate some DT features – with the used settings in this manuscript. See below for further suggestions on the method.

General comment on section 3:
→ I am not sure the parameters used for the clustering analysis are the most meaningful. With the method as is, all DT types as well as all modes of variability are blended together and really difficult to separate. I have a couple of suggestions that should help with the clustering and interpretation of results:

- Separating NDT' by time-scales, e.g. into subseasonal, interannual, QBO-specific, ENSO-specific… would make any cluster regions found easier to interpret, <u>and relatable to previous works</u>. For example, a good test of the clustering method would be to compare its output regions to the El Nino – La Nina differences shown in **Wilhelmsen et al. (2020)**, their Fig. 3c.
- Combine with different parameters for clustering: e.g. DT depth (difference in height between the two tropopauses) or its standard deviation could be tested instead of std(NDT') to see whether the clustering performance improves. DT depth properties should distinguish equatorial and mid-latitude / polar DT much better than DT frequency alone.

→ I think some sensitivity tests with other parameters and/or case-studies (e.g. ENSO-specific timescale) are necessary in order to reassure the audience that this method can give some useful and meaningful output, that can be related and build upon previous DT studies. Especially the current Figure 6, in my opinion is rather far from being convincing on the method's potential, and it's done with a single setting for clustering on, again, basic distribution parameters of frequency.
→ Once clustering gives something more robust, one may expect a lot more juice coming from the multivariate regression analyses.

#
 # Sect. 4+5
#

I feel these sections could be summarized into a couple of methods subsections. They are very heavy in the present manuscript, their extended version with all the minutiae can be moved to an appendix or supplement.

Same with the first half of section 6 actually.

l. 320-345: I don't think it's necessary to explain what is the R^2, adjusted R^2 and F-statistic, these are standard things… please explain the most relevant details for model evaluation in a very summarized way.

#
 # Sect. 6
#

l. 379-385: I was thinking exactly this while reading everything after section 3: the clustering analysis feels to me like the bottleneck for the rest of the methods/results – it is imperative that the results in section 3 are robust and thoroughly tested, then sections 4-6 will benefit greatly and interpretation beyond just the statistics will be more straightforward.

l. 387-403: this is the only paragraph where the results from the multivariate linear regression are discussed, this part should be markedly extended.
→ From these results, please highlight what is added upon the referenced work.
→ From the results of this section, can you say something about which climate indices affect STE the most? Can your methodology provide a way forward to answer such question?

l. 430-435: I don't think it's necessary to explain what a q-q plot is…

Figs. 10-11:
→ Both appear a bit pixelated, please increase quality, and reduce the horizontal separation between the red boxes. Also, titles are repeated in each sub-panel, you can save a lot of space using one title above all panels.

l. 439-445: this can be considered an outlook, please create a separate section 'Summary & Outlook' or similar that summarizes main results and discusses possible applications and adaptability of your method.

Appendix C: is it really necessary that these very technical infos stay within the main manuscript? I'd suggest to move it to a separate supplement document.

Data availability statement: state here (or reference in methods section) what software is used for clustering analysis and regression model.