Review of "Signs of climate variability in double tropopause global distributionfrom 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 potentialto 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.

Thank you for your comments. We agree with this comment. In the revised version, which was re-written following a completely different structure (Abstract, Introduction, Data and Methodology, Results, Discussion and Conclusions), starting from a DT database obtained from RO observations, we propose to explore a possible relationship between the spatio-temporal distribution of DTs and a set of monthly climate indices, with a primary focus on the methodological approach. With the main purpose to illustrate this idea, we first apply a cluster analysis to geographically associate DT occurrences. Secondly, we construct a multivariate linear regression using a progression of different models, considering train and test populations, to identify climate indices relevant for DT occurrence. Then, these preliminary results should be considered as the beginning of a more in-depth analysis, currently in progress, in which the robustness of the results is still pending to be found and established.

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.

We have included an additional discussion regarding tropopause dynamics in sections 1 and 2.1 and in 2.2.1 the proposed cluster analysis is discussed in more detail.

Although my recommendation is 'reject' for the manuscript in its current form, as I state above themethod has good potential and I would encourage resubmission once the extensive list of issues isrigorously 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.

This study focuses on the relation (and correlation) of double tropopause occurrence and climate indices. Of course, over the last two decades there were several complex tropopause studies based on RO data. Because of the properties of the RO technique tropopauses and double tropopauses can be detected precisely on a global scale. Several of these previous studies describe climatologies and even trends in tropopause parameters. Some of these studies are already listed in the references. In the revised version, we will include a broader spectrum of this previous publications related to tropopause and double tropopause investigations using RO or other datasets.

In this study we use the WMO definition of the lapse-rate tropopause (Wilhelmsen et al 2020). This definition includes also the conditions to detect double tropopauses. Of course, the WMO definition from 1957 was developed based on datasets with a coarse vertical resolution.

Due to the availability of high vertical resolution datasets (radiosondes and, e.g., RO data) some modifications (in comparison to the pure WMO definition) on the tropopause detection retrievals have been performed, e.g., Schmidt et al. (2005) and Birner (2006) (see below).

But, if you compare (double) tropopause climatologies from different authors (that usually avoid giving precise information on the tropopause detection algorithms) the climatologies are very similar, i.e. the results of our double tropopause climatology are robust.

In summary, we would argue that even if there are small differences in the tropopause algorithms the general picture of the tropopause climatologies is the same. From that we further conclude that our results based on our analysis would have no basic differences if we had chosen a (small) different criterion to define the double tropopause.

Moreover, in the revised version we include the following additional DT studies and the corresponding main focus in each of them, in addition to the already referenced:

Randel, W. J., D. J. Seidel, and L. L. Pan (2007), Observational characteristics of double tropopauses, J. Geophys. Res., 112, D07309, doi:10.1029/2006JD007904.

Temperature profiles in the extratropics often exhibit multiple tropopauses (as defined using the lapse rate definition). In this work the authors studied the observational characteristics of DT based on radiosondes, ERA40 reanalysis, and GPS radio occultation temperature profiles.

Schmidt, T., J.-P. Cammas, H. G. J. Smit, S. Heise, J. Wickert, and A. Haser (2010), Observational characteristics of the tropopause inversion layer derived from CHAMP/GRACE radio occultations and MOZAIC aircraft data, J. Geophys. Res., 115, D24304, doi:10.1029/2010JD014284.

The characteristics of the Northern Hemisphere (NH) midlatitude (40°N–60°N) tropopause inversion layer (TIL) based on two data sets. First, temperature measurements from GPS radio occultation data (CHAMP and GRACE) for the time interval 2001–2009 are used to exhibit seasonal properties of the TIL. Secondly, high-resolution temperature and trace gas profile measurements on board commercial aircrafts (Measurement of Ozone and Water Vapor by Airbus In-Service Aircraft (MOZAIC) program) from 2001–2008 for the NH midlatitude (40°N–60°N) region are used to characterize the TIL as a mixing layer around the tropopause.

Castanheira et al. (2012), Relationships between Brewer-Dobson circulation, double tropopauses, ozone and stratospheric water vapour. Atmospheric Chemistry and Physics.10.5194/acp-12-10195-2012.2012.

Statistical relationships between the variability of the area covered by DT events, the strength of the tropical upwelling, the total column ozone and of the lower stratospheric water vapour are analyzed. The analysis is based on both reanalysed data (ERA-Interim) and HIRDLS satellite data.

Liu, C., & Barnes, E. A. (2018), Synoptic formation of double tropopauses. Journal of Geophysical Research: Atmospheres, 123, 693–707. https://doi.org/10.1002/2017JD027941

As DT are ubiquitous in the midlatitude winter hemisphere and represent the vertical stacking of two stable tropopause layers separated by a less stable layer, by analyzing COSMIC GPS data, reanalysis, and eddy life cycle simulations, the authors demonstrate that they often occur during Rossby wave breaking and act to increase the stratosphere-to-troposphere exchange of mass. Moreover, the adiabatic formation of double tropopauses and two possible mechanisms by which they can occur were proposed.

Shao, J., Zhang, J., Tian, Y., Wang, W., Huang, K., & Zhang, S. (2023), Tropospheric gravity waves increase the likelihood of double tropopauses. Geophysical Research Letters, 50, e2023GL105724. https://doi.org/10.1029/2023GL105724

As the tropopause region is crucial for the stratosphere-troposphere exchange (STE) and acts as an indicator of climate change, DT events act to increase the STE process but their driving mechanisms remain an open question. The present assessment offers for the first time the linkage between tropospheric gravity waves and DT events by exploring a global data set of multi-year radiosonde measurements.

Schmidt et al. (2005), GPS radio occultation with CHAMP and SAC-C: global monitoring of thermal tropopause parameters. Atmospheric Chemistry and Physics.10.5194/acp-5-1473-2005.

Birner, T. (2006), Fine-scale structure of the extratropical tropopause region, J. Geophys. Res., 111, D04104, doi:10.1029/2005JD006301.

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, thiscompounds with my comment M1.1 and makes interpretation of the results very, very difficult, ifone wants to go further than just showing the statistics.

See my "**general comment on section 3**" for suggestions on this issue.

The choice of location and spread in terms of the mean and standard deviation constitutes a consistent start in a cluster analysis. But of course, we agree that a deeper analysis, which is not considered here, could add higher order moments to explain the symmetry of NDT'.

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 bel*ow*

We aimed to replace the word "reasonable" by consistent arguments in the new version. Throughout the new text, we have attempted to balance our preliminary conclusions and assumptions.*

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 includesometimes unnecessary definitions of standard statistical measures.

The revised version, with a different structure from the first version, includes several explanatory paragraphs added in the last two sections. We hope that the reading and comprehension of the text, together with a revision of the language, will now be more agile and clearer.

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.

The abstract has been rewritten in the new version. Main results are now more detailed.

# 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?

This is a possibility, please see above our discussion regarding definition of DT.

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

We agree, both sentences that mention big data analysis were eliminated.

# 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.

Citations included.

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

At the beginning of the COSMIC mission, all the satellites were grouped together to ensure a synchronized start for their measurements and to simplify the deployment process (around May 2006). The final arrangement of the COSMIC satellites allowed them to cover the entire globe. Between 2001 and 2006, the density of available RO is considerably lower.

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 mostrelevant ones.

l. 110: also, Wilhelmsen didn’t invent the lapse-rate tropopause definition, please cite the WMOdefinition that the algorithm uses.

In the revised version we included additional references and a discussion about the lapse-rate tropopause definition is included too (Section 2.1).

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.

We are not able to indicate the number or pencentage of them, but we agree that it is relatively considerably low.

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

We mention this reference in Section 2.1 and we explain that we prioritize the availability of a sufficient number of events in each cell

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.

Both suggestions were included in the revised 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 theproper way of visualizing the same information.

lso format month number into “YYYY”, perhaps label every two years.

According to comments made by both Reviewers, the “old” Figure 1 was eliminated in the revised version of the manuscript.

# 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.

The clustering is defined from two measures of location and spread of the time series: The mean values of the DT_{freq} time series and the standard deviations of the DT'_{freq} Additional moments of higher order could have been included too, but we feel that for a preliminary luster analysis it is advisable to consider the first two measures. A deeper analysis, which is not considered here, should progressively include higher order moments to check the symmetry of the DT_{freq} distribution. The choice of mean DT'_{freq} values instead of the mean DT_{freq} values would not provide any information.

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

We modified this paragraph: “In Fig. 1, within the interval [.065, .075] the resulting number of clusters is 6, so we chose a mean cutoff point equal to .07 in hierarchical clustering. This value indicates a commitment number of 6 clusters, retaining a significant number of individual cells in each cluster. We then proceed with the cluster analysis according to this classification in 6 groups.”

Also, please state in this paragraph, what is the usual range of cutoff distances in CA in general.

If we properly understood this comment, there is not a defined range of cutoff distances in CA

(M2)

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

No, the colors in Figure 1 are arbitrary. In the other hand, in Figures 2 to 5 there is a strict correspondence between the selected colors.

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.

The clarification regarding the choice of colors was included in Section 3 and in the legends to figures 1 *and 2.*

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, maybethere are relative density maxima corresponding to some clusters. But from the current plot one can’t say.

We agree that this phrase is inappropriate and it was reworded. As is well known k-means is an iterative data partitioning algorithm that assigns n observations to exactly one of the 6 clusters previously established in the hierarchical CA and defined by six centroids. That is, k=6 is chosen before starting the algorithm. This concept is included in the paragraph preceding Fig. 2.

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.

The numbering of each cluster in figures 2 to 5 is now indicated below Fig. 1.

Fig. 6 and corresponding text:

Discussion missing, please at least discuss the different clusters in relation to different high DTfrequency regions from e.g. **Wilhelmsen et al. (2020)**

In the revised version, the discussion section includes additional references to previous work, in particular from Wilhelmsen et al. (2020).

In my opinion, describing the clusters as symmetrically distributed relative to the equator haslittle 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.

This point was highlighted in the discussion following Fig. 5 (old Fig. 6).

> Fig. 6 looks quite coarse, since it works with monthly timeseries, a refinement to 5x5 degree orbetter is possible from RO coverage, and would be very welcome.

We agree with this comment, however after testing different possible resolutions for the cells we concluded that the optimum size for monthly time series was 10x10 degrees.

l. 204-205: *“The interconnected nature of each sub-region is highlighted in the polar, sub-polar andequatorial 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 themethod’s weakness: equatorial and polar DT’s have very little in common, yet the clusteringmethod is mixing them up.

We agree with this point. This comment was also modified and reworded, following Fig. 5.

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 – withthe 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:

- *We believe that the revised version discusses in some detail the choice of the parameters used, always considering the present work as a first stage of a more in-depth analysis.*
- Separating NDT’ by time-scales, e.g. into subseasonal, interannual, QBO-specific, ENSO-specific… would make any cluster regions found easier to interpret, and relatable to previous works. For example, a good test of the clustering method would be to compare its output regions tothe 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 andmeaningful 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 donewith 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.

As mentioned above, we agree with the potential interest of these possible new research directions, in addition to those suggested in our new section 4, concerning clustering and modeling. However, we prefer to leave them for a future contribution.

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 anappendix or supplement.

Same with the first half of section 6 actually.

The general and specific structure of the revised version is very different from the original one.

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.

This was modified in Section 4.

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 theresults 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 STEthe 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…

This was modified and reduced in Section 4.

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.

Done.

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.

The revised version, in particular section 4 was rewritten. This suggestion has been included.

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.

We believe that the appendixes are not strictly contained in the main manuscript, but we have no *problem to create a separate supplement if necessary.*

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

Done.