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

There is undoubted value in the use of GNSS-RO observations to monitor and understand changes in complex tropopause characteristics including conditions of multiple tropopauses. The premise of the analysis is therefore strong. The authors are to be commended for taking this on and I would encourage them to work further on it.

However, there is probably considerable work required for this to be publishable. Concerns relate to the appropriateness of the statistical approach, the lack of physical interpretation of the results in terms of fundamental processes and the overall structuring of the paper including the complete absence of a classical discussion and conclusions section.

I limit below to only major comments given the need for substantial work before this could be publishable.

Major comments

1. The consideration of solely double tropopauses is somewhat limiting. There are many interesting complex tropopause cases illuminated by RO and this should be at the very least acknowledged. Also, the sensitivity to the single definition of a double tropopause deployed is an obvious weakness. If you had chosen different objective criteria to define a double tropopause event how would your analysis have differed?

Thank you for this comment. This study focuses clearly 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.

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.

2. Overall paper structure is really far from the classical structure for a paper, that being introduction – methods – results – discussion-conclusion. Interleaving methods and results throughout makes for a very challenging read for a reader with new aspects of methods suddenly being dropped at random points in the text. Rewriting the paper in the more classical way would probably make for an easier read. In particular the lack of a discussion and conclusions means the 'so what' part is almost entirely missing. You need to close by placing your analysis in the broader context, highlight any caveats, and outline some potential future directions and open questions.

   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.

3. Figures in general need considerable work for clarity. In particular figure 1 is indecipherable to the reader as presented. This could instead, for example, have been presented as a stacked plot of timeseries by latitude bands N to S with the same vertical axes ranges extending vertically across a whole page enabling a reader to easily ascertain latitudinal variations. This could have avoided trying to find 18 colours which are challenging for most and indecipherable to colour- blind readers. Other figures have similar challenges but Figure 1 is by far the most challenging to comprehend as currently presented.

   We agree with this comment. In the revised version, Figure 1 was eliminated, as it is not essential to illustrate our results and we tried to improve the resolution of some of the remaining figures.

4. Why were the 29 indicies chosen and why do you expect these to be important in double tropopause behaviour? This married to the lack of physical interpretation is problematic. When you do compare them it currently leaves a reader with a perhaps unfortunate impression that you are proverbially throwing spaghetti at the wall in the hope that some of it sticks. I doubt this was the case but as currently written it is hard to tell on what basis you chose this set and why you think all these might, plausibly, matter. This comes to the point made in the

   opening remarks that this is very statistically heavy and you really need more physical understanding in the piece as a whole.

   Climatic indices play a crucial role in understanding the general circulation of the atmosphere by providing valuable insights into climate patterns and variability, climate change and in the link between the ocean and the atmosphere. Moreover, for improving our understanding of the interconnected nature of Earth's climate system. Overall, climatic indices are essential tools for meteorologists, climatologists, and policymakers in understanding and responding to *atmospheric dynamics. Besides, double tropopauses are produced due to specific atmospheric conditions that lead to the formation of distinct tropopause layers as a consequence of different dynamic or thermodynamic situations, i.e., stratospheric temperature inversions, vertical shear and stability, convective activity and jet streams. These comments were included in section 2.1.*

5. In the cluster analysis work from the analysis as shown it is hard for me to really tell that there truly are six distinct clusters. In Figure 3 they just look like cuts driven by the arbitrary selection of six clusters in what is very much a continuum of behaviour with no obvious centering into distinct clusters driven by likely distinct physical behaviour. This is compounded in Figure 6 where in particular cluster 4's distribution suggests this cluster is not driven in any way by the physics with cluster placement ranging across almost all latitude bands.

   The non-hierarchical K-means cluster analysis only follows the classification indicated above by the hierarchical method into 6 groups. Additional parameters of higher order than the mean NDT and the standard deviation of NDT' could also have been included. This is one of the main assumptions of our analysis, resulting in a well-defined object separation in Figure 2 of the new version. We do not expect that final physical conclusions can be obtained before the robustness of the results (time series classification and model relating NTD' to climate indices) is guaranteed.

6. Given significant seasonality in the latitudinal distribution of key aspects of circulation relevant to double tropopauses, the use of a seasonally varying criteria or criteria that track key features from e.g. reanalyses may have been considerably more elucidating. We know that double tropopauses are more common in key physical conditions as you have alluded to. Using a fixed lat-lon distribution when features may be repeatedly transient across such fixed grids on an annual and semi-annual basis probably explains much of the annual and higher harmonics structure in figures 4 and 5. Again, this is highlighting the need to really think about the physics here. The use of a fixed lat-lon grid vs a feature tracking approach e.g. following the sub-tropical and polar jets and the ITCZ throughout the year should be considered in revisions. A feature tracking approach which could be utilized by e.g. using ERA5 diagnostics for features of interest might give a clearer picture than your current fixed lat-lon approach.

   We are aware that climate studies often use latitude and longitude grids to represent global data, but alternative methods have emerged, especially with machine learning. Several approaches have been developed to model and analyze global climate without relying on traditional grid clustering. Some of these methods focus on pattern recognition, dimensionality reduction, and leveraging irregular data inputs: graph neural networks, spectral methods and harmonic analysis, gaussian process models, unsupervised learning with autoencoders or variational autoencoders and self-organizing maps, temporal and spatial attention mechanisms in transformers and principal component analysis. (We postpone a clustering of global climate data based on large-scale patterns rather than latitude and longitude grids for a next contribution).

   On the other hand, to describe the behavior of a global variable in terms of climate indices states, rather than grouping regions, we can use a few advanced methods to analyze how global variables like DTfrequency, temperature, precipitation, or wind speed are influenced by climate indices such as the ENSO, NAO, or PDO: Multivariate regression models with climate indices, state-space models, dynamic mode decomposition with climate indices, canonical correlation analysis and machine learning methods like random forests or neural networks (we begun here with the first of these methods). These comments are included in section 4.

7. The multivariate regression really needs much more physical interpretation to be of any value. At present the statistical results are presented and any physical interpretation pretty much left as an exercise for the interested reader. Statistical significance is a necessary but insufficient condition to draw robust conclusions here. It is necessary to understand physically what these results are showing us and what they mean. Why is something leading or lagging and if something is lagging does that mean that somehow double tropopauses are causing that phenomena? There is an absolute need for understanding physically what your results mean here for them to have any scientific value. I can understand how double tropopause features may lag a given phenomena, but I am unsure how to interpret a result saying they are a leading indicator. Table 1 is thus very confusing to me as a reader presently.

   As mentioned above, the results above presented must be strictly considered as a first step of a deep analysis to reveal the model that minimizes RMSE in test and training data. This presentation should be considered as the beginning of a more in-depth analysis, currently in progress, in which the robustness of the results can be verified. Prior to the development of any model, as NDT´ and the features may present the best cross-correlation for time lags k different from zero, it is worth considering k values within an interval around $k = 0$. A resulting $k \neq 0$ value may indicate the ability of NDT´ to anticipate a given feature, or vice versa. Moreover, a significant maximum CC may indicate the possible relative relevance of the respective feature in relation to the others (Section 2.2.2 and 4).

8. I am not really sure how I should interpret figure 9 as presented. In particular in clusters 1 and 3 the test RMSE is consistently lower than the training RMSE which makes no logical sense. This may highlight that the cluster definition is not appropriate (see earlier point) and that the behaviour within clusters is non- stationary in interesting ways as a result.

   RMSE is shown for the training and test samples. This is an indication of the degree of possible model overfitting ($RMSE_{test} >> RMSE_{train}$). A possible indication of overfitting appears in clusters 2, 4 and 5. In clusters 1, 3 and 6, $RMSE_{test} << RMSE_{train}$ suggests that the model performs better in the test population, as desired. We recall that if the test data are similar to the training data but without as much noise, the model may perform better in the test population, also resulting in a lower RMSE. Moreover, if the training data contain many outliers that negatively affect the model, and those outliers are not present in the test data, the RMSE also may be higher in the training data than in the test data (section 3.2.2).

9. Table 2 again you are making the reader do the lifting of the physical understanding as to why these particular modes might matter to these particular clusters. Taken together with Table 1 I have a real challenge thinking how to interpret your results here. You need to help a reader understand how to interpret these combined results.

   Table 2 lists the features selected by the AIC forward step-wise method, it is to say, which features are significant or relevant according to the best multivariate regression "with ALL" model found. R^2, adjusted R^2 and F-statistic values in each cluster are included too. Previously to the model, in Table 1 we indicate the lag corresponding to the best CC found between each of the features and NDT', with k ranging from -5 to + 5 months. The relative enhanced significance of each feature and the possibility that each feature anticipates or delays NDT´ by k_0 months is highlighted.

10. I am unclear why so much of what would nominally be considered key results is left in the supplement and not discussed at all in the main text. I may have missed it but I failed to note a reference to it and certainly a substantive analysis and discussion of these results.

    Sections 3 and 4 include further explanations of the material contained in appendices C and D. In this last, for additional information the coefficients and the features corresponding to the "with ALL" model, in clusters 1 to 6, are included. We believe that the definition of the meaning of each parameter is not necessary.

11. I am always loathed as a native English speaker to make this point as I could never even attempt to write a paper in any language other than English let alone to such as a standard, but the paper overall is a heavy read and either getting a native English proof reader or engaging a native English speaking co-author to help in the rewrite and restructure would be helpful.

    In the new version we have enlisted the help of an English proofreader.