Review of ‘Classification of synoptic circulation patterns with a two-stage clustering algorithm using the structural similarity index metric (SSIM)’

In this paper the authors introduce a new clustering method for the analysis of synoptic weather types over Western Europe, in a similar style to the traditional Grosswetterlagen approach. The main novelties of the method are the use of the SSIM instead of Euclidean distance to compute distances in the K-medoids algorithm, and a coupling of the K-medoids clustering to a hierarchical agglomerative model which replaces the ‘number of clusters’ hyperparameter with a more intuitve ‘maximum similarity’ hyperparameter.
Using ERA-Interim reanalysis data they test the robustness of the method to parameter and resolution variation, and show that it is essentially doing what they want it to do. Using these ERA-Interim patterns, they then compute a number of metrics in CMIP6 models, and use this to make a cursory assessment of model skill in representing synoptic European weather.
While I think the developed clustering method is interesting and has some potential benefits, especially the clever 2-step procedure to find cluster number, I do not think the current manuscript represents a strong research paper, and instead reads as more of a technical report. I have two main issues:

1. I do not think the analysis of CMIP6 simulations is very convincing, and in my opinion would need considerable extension to meet the stated aim of providing ‘a useful instrument to evaluate climate models, which gives an insight into the reasons for the poor model performance and the valuable feedback to model developers.’
2. Even if extended in this way, I do not believe the work fits well within the scope of ESD. To meet this scope, the work would in my view either need to engage with atmospheric dynamics (such as by investigating the drivers of good/bad synoptic pattern representation in CMIP models) or by exploring the socioeconomic impacts of their synoptic patterns (such as by looking at their relation to energy, agriculture, extreme event management, etc.). This would of course represent another major extension to the current work.

For these reasons, I unfortunately have to recommend the paper should be rejected as unsuitable for ESD.
Below, I provide more detailed comments that may be of use to the authors in developing this work further.

Detailed Comments

The choice to use a 22x22=484 dimensional space for cluster analysis is rather unusual, and bound to add to the issues of instability, and low representativeness of the cluster means that you comment on. Many approaches first reduce the phase space using EOF analysis, and are able to capture >90 percent of the variability with <40 EOFs. It might be valuable to comment on why you did not do this. Such approaches also reduce the ‘structure insensitivity’ of the standard Euclidean distance metric by the way, as they preselect large scale modes that encode the spatial structure of the flow.

The paper goes into considerable detail describing the new clustering method and demonstrating various aspects of its robustness, with the reward being a new way of validating climate model performance. However this most relevant aspect of the work is not explored in much detail, and there are some issues with parts of the analysis that is present:

- The most important element of robustness has not been explored – robustness of the method to temporal variability. If we wish to use observationally identified patterns and their statistics to evaluate the performance of uninitialised climate models, in either a historical or future context, then we must know how internal atmospheric variability alters the patterns and their statistics. While imperfect, there are many centennial reanalyses which could be used to look at synoptic patterns in different 40 year periods (as done in [1] for example). Failing this, a bootstrap approach could be used for the ERA Interim data. Without this, I find it very difficult to see how you can say that a low similarity between model and reanalysis SPs is because the model is bad, rather than due to the SPs being properties of a very particular time frame.

- The TRANSIT and PERS metrics are based on the 42x42 transition matrix of the SPs which must surely be very noisy, with less than 8 datapoints on average for every element. In my experience looking at sets of <10 clusters, much more than 50 years of data are needed to even vaguely constrain transition matrix elements, especially ones representing rare transitions. I do not think these metrics can be telling you anything real about the skill of CMIP models.

[1] “Quantifying climate model representation of the wintertime Euro-Atlantic circulation using geopotential-jet regimes”, Dorrington, Strommen and Fabiano 2022, Weather and Climate Dynamics, https://doi.org/10.5194/wcd-3-505-2022