In this paper, the authors introduce a new clustering method for the analysis of synoptic weather types over Western Europe. The paper is primarily methodological, suggesting a new approach, discussing its robustness and demonstrating its application to climate model evaluation. They make two main points:
1. That a large number (>30) of weather types is needed to properly sample the range of synoptic flows, and so also to make sure the drivers of extreme events are included.
2. That Kmeans using RMSE has deficiencies which are fixed when using SSIM and K-Medoids.

They use a two-stage clustering procedure where the SSIM is used instead of Euclidean distance to compute distances in the K-medoids algorithm, and this is coupled to a hierarchical agglomerative model which replaces the 'number of clusters' hyperparameter with a more intuitive 'maximum similarity' hyperparameter.
They compare their new method to previous approaches using synthetic data, investigate robustness to parameter changes and to temporal resampling, and then demonstrate the application to climate model evaluation, by summarising metrics based on these weather types' occurrence, persistence and transition properties into a single overall score.

For transparency, I was reviewer 1 for both rounds of revision of the previously submitted version of this manuscript.

I have some major issues with the manuscript, detailed below, and suggest major revisions.

<u>The usefulness of the suggested approach for capturing extreme events</u>

One of the major motivations given for the use of a large number of clusters in full field data was that this would help capture extreme events, whereas PCA based approaches with smaller numbers of clusters may not capture extremes. Unfortunately, neither side of this claim has been demonstrated. I also have some reasons to doubt the claim: looking at their figure 8, there is not much sign that the more common weather patterns are less extreme supporting than the rare patterns. Further, other work has shown that persistent regimes (i.e. common weather types) can drive cold and warm extremes.
As I suggested previously:
"I suggest that the authors more tightly focus the structure of the article around the importance of handling rare synoptic conditions and extremes in clustering approaches, showing an example situation where an impactful event was linked to a very rarely occurring circulation as motivation. I would then suggest a concrete demonstration that the EOF Kmeans with MSE approach more poorly handles rare circulations than the SSIM approach in ERA Interim…."

Even if it is the case that rare circulations are associated with rare extremes, when you compute the Jensen-Shannon divergence, you weight each class by frequency! So representation of rare flows has almost no impact on the resulting quality index.

<u>Usefulness for climate model evaluation</u>

The authors also emphasise the value of their method for climate model evaluation. Indeed, circulation based metrics can be very useful for such analysis. This can and has been done several different ways (although it would be easy to think otherwise reading the authors' work), with only a few regimes at one extreme as in [1], or on a gridpoint basis as in [2] at the other extreme.

However, I am seriously concerned that the method the authors suggest is not suitable for this purpose.

The author's explain that using similarity as a metric, ~37 weather patterns are needed to fully capture the diversity of European circulations. I accept this, and it is a useful perspective, and similarity is a nice way to quantify this. Exploring spatial and seasonal variations in this number of 'necessary patterns' could be an interesting dynamical study.
But, for model evaluation, the question of relevance is not how many weather patterns you need, **but how many weather types you can constrain**, given data limitations.

The authors compute error metrics for weather pattern frequency (37 elements), transition matrix (37x37 =1369 elements) and persistence probability over days 1-8 (37x8=296 elements). Simply put, using 40 years of ERA-Interim the sampling uncertainty in such fine-grained metrics are almost certainly far larger than any difference between climate models and era-interim. The fact that the inter-model variation in scores is so low reinforces this point. I believe your quality index is almost entirely noise, averaged over a few hundred variables.

I make this claim quite confidently, as I know that it is difficult to find significant differences in the frequency and persistence of models and reanalysis when only using 3-10 regimes, and 100 years of data. Of course I would be pleased to be proven wrong: if you can rigorously constrain sampling variability in model and observational statistics, and so provide upper and lower bounds on your quality index, and still get meaningful results, then the scientific contribution is strong. Otherwise, I would move away from climate model evaluation as a goal for this methodology.

Synthetic data

The synthetic data section raises some questions for me. One clear point that I found interesting is that Kmeans leads to distorted patterns (i.e. not circles as in the synthetic data). However I think the other points would be better made in ERA Interim than in the synthetic data. The synthetic data does not have multimodal structure, so there is no reason to expect any clustering algorithm to give very clear clusters: there are no clusters to identify, just 'hallucinations' of the method. In fact, you could argue that in non-structured data, a good clustering algorithm *should* give unclear structures. Also, I do not follow the claim about snowballing: the k-medoids with SSIM produces the most snowballing of all algorithms shown in figure 4.