Earth Syst. Dynam. Discuss., referee comment RC3
https://doi.org/10.5194/esd-2022-29-RC3, 2022

Comment on esd-2022-29

Anonymous Referee #3

Referee comment on "Classification of synoptic circulation patterns with a two-stage clustering algorithm using the structural similarity index metric (SSIM)" by Kristina Winderlich et al., Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2022-29-RC3, 2022

This paper describes a novel method of clustering circulation fields, and then applies this method to assess the ability of CMIP6 models to simulate realistic circulation patterns. The paper is generally clearly written and straightforward to understand, but I feel that the authors have not sufficiently justified the use of their method over something simpler like k-means. The analysis of circulation in the CMIP6 models is also rather brief. I therefore recommend major revisions.

Major comments

The bulk of the paper describes a new two-step classification method, arguing that previously used methods are 'suboptimal'. However, I don't think that the authors have sufficiently motivated their choice of method - my suspicion is that standard k-means clustering would give similar results.

The authors argue that k-means clustering has a number of drawbacks:

i) the number of clusters has to be pre-specified.

(But the authors' similarity threshold parameter seems to play a similar role, as it is subjectively chosen and also influences the number of clusters.)

ii) k-means centroids could be misleading and unrepresentative of the fields in the cluster.

(But does this not also apply to medoids, as a single field chosen to represent a set of fields? Surely any daily field will contain its own set of small scale features that don't resemble those of other fields. The authors appear to find that the cluster centroids and medoids are pretty similar anyway.)

iii) k-means clusters could be sensitive to outliers. (But does this actually happen in the case of the geopotential height fields?)

The authors quote image processing references to justify the similarity metric used here over (say) mean square error. It would be more convincing if the authors could show actual examples of deficiencies in k-means clusters constructed from their circulation data, and/or that clusters produced using their method were superior to those produced using k-means (for example, using the criteria set out in section 3.3).

2. The analysis of the CMIP6 models is rather limited - there's a ranking of the models according to various metrics, but not much more. Why did the authors choose these particular metrics over the wide variety of other possibilities? Do the HIST statistics correspond to
biases in the mean state of the models? Can the authors suggest any reasons why some models are better than others - eg resolution?

Also, the transition statistics are likely to be very noisy with 43 different circulation types. How can we be confident that the transition results from ERA-Interim are a meaningful benchmark - is there enough reanalysis data to do this?

Again, it would be interesting to know if the results of the model evaluation analysis are signficantly different if k-means derived clusters are used instead.

Minor comments

Line 49 - "Hochman et al proved" - I think 'proved' is only an appropriate word when discussing mathematical proofs. I suggest something like 'argued' or 'demonstrated'. Also, people arguing that clusters represent genuine low-frequency weather regimes tend to find relatively few of them (four in winter seems a popular choice). Presumably the authors are not arguing that the 43 types they analyse here each represent a physical weather regime in this sense?

Line 58 - 'the moving atmosphere' - I'm not sure what this means.

Line 90 onwards - standardising the height fields means that information about the amplitude of the circulation anomalies is lost. But different amplitude anomaly patterns could produce quite different responses in eg surface air temperature and precipitation, so I'm not sure the standardisation step is beneficial.

line 111 - "The k-means clustering assigns every data element to the cluster center that is closest to it, if only by a small margin." Isn't this true of any method that assigns each field to one of a set of a classes?

line 112 - "This makes the method sensitive to noise in the data and may lead to an assignment of a data element to a structurally dissimilar cluster center." - what does "structurally dissimilar" mean here? How can we distinguish the noise from the structure in any given field? Can the authors show examples of fields that are far apart under the Euclidean distance metric but close together under the similarity metric, or vice versa?

line 116 - Doesn't using medoids also risk inflating the significance of small-scale noise in the daily field chosen as the medoid?

line 137 - "Wang and Bovik (2009) demonstrated that the MSE has serious disadvantages when applied on data with temporal and spatial dependencies" - dependencies on what? Does this mean temporal and spatial correlations?

line 194 - is the similarity between two clusters measured using their medoid fields?

line 267 - Is the algorithm stable if applied to slightly different initial subsets of the data? The number of patterns may be stable, but do the same patterns emerge from the clustering?

Figure 3 - it would make more sense to have the transition between the blues and reds in the colour bar at zero, not +0.25.

Line 245 - should there be a reference to figure 6 here?

Line 282 - "However, it is necessary to demand that a cluster medoid represents all cluster elements and their whole entity as a group." Does comparing the mediod and centroid really guarantee this?

Line 307 - is section 4 meant to be labelled 'Method', the same as section 3?

Figure 4 - Can the colour bar be included in the figure? There's room in the bottom row of panels.

Line 320 - "This correspondence gives us an evidence that, albeit not tuned to and not required to mimic semi-manual classifications, the new classification method determines not just arbitrary synoptic patterns but those described by experts in semi-manual classifications."

I'm not convinced - given that there are 43 different types, it seems quite likely that some of them could resemble Grosswetterlagen patterns by chance.

Figure 7 - the text in the figure labels could be much larger for legibility.

line 447 - again, I don't think one can infer that this is an inherent advantage of the SSIM method without making a comparison with other cluster methods.