Comments of Reviewer 3 are in blue, answers of authors are in black/**black**.

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.

Using the medoids instead means for representing clusters was based on our simple experience: we constructed the two-stage clustering method using the k-means procedure (as the second step) and run it. In the application on 40 years of Reanalysis cluster means are computed on multiple (often >600) synoptic maps (that are geopotential anomalies). This lead to building of "smooth" maps to a degree that these means do not represent realistic geopotential anomalies anymore, but "blur" pictures of some unidentifiable circulations. We realized then that the danger of using cluster means as cluster centers is that the "blur" centers attract multiple unsimilar elements into one cluster making it even more "blur". This effect is often called "snowballing". The final set of clusters obtained with such routine is the rather small, but each cluster is likely to include elements strongly unsimilar to other elements (and we tested this indeed).

Therefore, with the aim to avoid such "blur" cluster centers, we discuss the choice of an alternative representation of clusters (Lines 116-122). A medoid of a cluster can be seen as "the representative element" of this cluster i.e. element most similar to other elements in the cluster. Once the cluster is changed (merged with another one by the hierarchical step for example) the medoids are recomputed. Every new attribution of an element is done to a cluster to whose medoid the element is the most similar. This ensures only attribution of similar elements to clusters.

**As answering to comments of Reviewer 2, we suggest including an additional comparison of synoptic classes derived with a "standard" k-means routine for illustration the advantage of using our k-medoids method.**

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.)

The threshold on similarity of a pair of images is based on human visual perception and can be estimated intuitively well. We estimated this threshold being th≈0.45 for similar pairs of geopotential anomaly images in the study domain asking 20 persons (our colleagues, 10 male and 10 female). As an example we show similar, weakly similar and dissimilar patterns in Figure 3, Page 9.

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.)

It is crucial to differentiate when medoids are used: in the classification algorithm (in the hierarchical clustering part and in the k-medoid part). Once the cluster is changed (merged with another one by the hierarchical step for example) the medoids are recomputed. Every new attribution of an element is done to a cluster with the most similar medoid. This ensures, at each iteration of the algorithm, only attribution of similar elements to clusters. After the clusters are finally built using medoids, we show that also the cluster means are strongly similar (Figure 10, and Lines 453-455). This means: we used medoids (single elements for representing clusters) in the classification algorithm assuring that we avoid the "snowballing" and produced the final classes those cluster means resemble respective medoids. Therefore, using medoids is an efficient strategy for clustering and producing homogeneous clusters (clusters that only have elements that are similar to their centers).

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

Yes, it does happen. Furthermore, it happens often. We did start using k-means in our algorithm and it produced fewer classes. But: multiple classes contained pairs of elements dissimilar to each other! This is not surprising, as the classical k-means optimizes the within-class variance, that is the distance between the members and the mean, and does not (!) require elements of the class being similar to each other.

In a classical k-means algorithm, each cluster is represented by its mean. Such cluster mean computed on multiple (often >600) synoptic maps (that are geopotential anomalies) is often "smoothed" to a degree that it does not represent any realistic geopotential anomaly anymore, but a "blur" picture of some unidentifiable flow. The danger of using cluster means as cluster centers is that the "blur" centers attract multiple dissimilar elements into one cluster making it even more "blur". This effect is known as "snowballing". The final set of clusters obtained with such routine is rather small, but each cluster is likely to include elements dissimilar to other elements (and we have seen this indeed!). This is the low representativeness of cluster means we comment on in the manuscript (Lines 116-122). With the aim to avoid such "blur" cluster centers we discuss the choice of an alternative representation of clusters (Lines 116-122). A medoid of a cluster can be seen as "the representative element" of this cluster i.e. element most similar to other elements in the cluster. Once the cluster is changed (merged with another one by the hierarchical step for example) the medoids are recomputed. Every new attribution of an element is done to a cluster to whose medoid the element is the most similar. This ensures only attribution of similar elements to clusters and is called stability of the method.

**We suggest including an additional comparison of synoptic classes derived with a "standard" k-means routine for illustration the advantage of using our k-medoids method.**

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

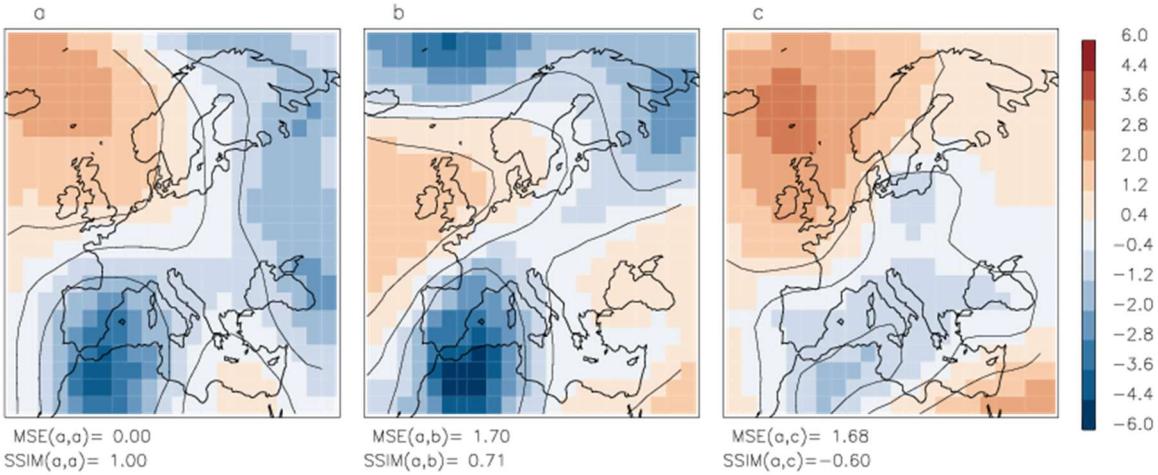
data, and/or that clusters produced using their method were superior to those produced using k-means (for example, using the criteria set out in section 3.3).

Choice of the SSIM metric. The choice of the SSIM instead of MSE was done at the very beginning of this work as the first classes were built that included pairs of elements visually dissimilar but with a rather small MSE. Therefore, we searched for a better measure to the structural similarity and turned to the field of image processing.

Example 1. The pairs of images (a,b) and (a,c) have nearly the same small MSE ( for comparison, the max MSE=11.6 for data of all daily data in 13 selected yeas) but are strongly different in terms of SSIM:

*MSE(a,b)*=1.70 and *SSIM(a,b)*=0.71 → means images *a* and *b* are similar

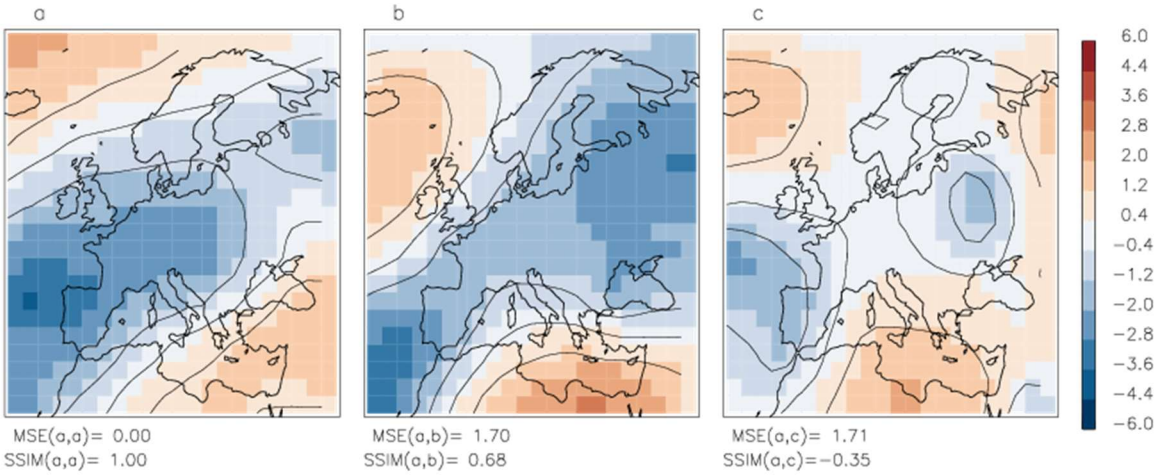
*MSE(a,b)*=1.68 and *SSIM(a,b)*=-0.60 → means images *a* and *b* are dissimilar (SSIM <0!)




Example 2. The pairs of images (a,b) and (a,c):

*MSE(a,b)*=1.70 and *SSIM(a,b)*=0.68 → means images *a* and *b* are similar

*MSE(a,b)*=1.71 and *SSIM(a,b)*=-0.35 → means images *a* and *b* are dissimilar (SSIM <0!)




Choice of the cluster centers. **We suggest including an additional comparison of synoptic classes derived with a "standard" k-means routine for illustration the advantage of using our k-medoids method.**

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?

Evaluation methods range from single-variable to multi-variables biases, from climatic mean assessment to climate change (trends, periodicity, interseasonal/interannual/interdecadal variability), extreme values, abnormal values and quantitative evaluations of uncertainty for computed model variables. There are not many metrics in model evaluation besides those that use above-mentioned variables, such as, mean bias, extreme value statistics, and frequency of occurrence of a particular signal. We tried to choose a set of independent and informative metrics. In the present manuscript, we only introduce an additional Quality Index to be used in an evaluation routine (that may include additional metrics; this depends on requirements of the user that does the evaluation).

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?

The histograms represent only the frequency of occurrence of synoptic patterns. From HIST we can see which patterns are under- and overrepresented. We expect models with shared/similar dynamical core doing similar errors.

However, we consciously restrain of judging possible reasons for better or worse model performance, as we believe that the aim of the evaluation routine is to quantify deficiencies. An investigation of reasons for a bad/good model performance would require good knowledge of all analyzed models (there over 30) regards their grid, numerics, processes resolved and drivers used. Last, but not least, knowledge of a model's heritage (history of development) would be an important source for judging its performance as, for example, models sharing a similar dynamical core with the reanalysis model may have advantages as compared to models with different dynamical cores. These issues are not trivial to summarize and discuss in the scope of our humble manuscript about one clustering method.

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?

"Noisiness" of statistics: The transitional matrix has 43x43 elements. In our case, about ½ of all these elements contain the main load (probability of transition > 0) and contribute to the Quality Index most as we use the Jensen-Shannon distance (Equations 6-9, Pages 11-12). The choice of the Jensen-Shannon distance weights the contribution of each matrix element by its frequency (similar to computation of Kullback–Leibler divergence): frequent transitions govern contributions to the distance measure, and vice versa, rare transitions make smaller contributions (it is quite robust against the "noise" from infrequent elements).

Benchmark: The Quality Index of NCEP1 (TRANS) is 0.91 – the higher value than any CMIP6 model gives and additionally significantly overshooting the models with respect to the inter-model range of Quality Indices. This gives the confidence that the ERA-interim is meaningful benchmark.

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

We suggest demonstrating the clusters built by k-means. As discussed above, these classes contain dissimilar elements due to "snowballing" effects owing the use of the "blur" means for representing clusters. We chose representing of clusters by medoids in order to build clusters that are more homogeneous (contain elements similar to clusters medoid).

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?

We do not aim to classify particularly low-frequency weather regimes. Our classification method was build with a purpose to include frequent as well as rare synoptic situations independently on their temporal occurrence and persistence. For searching low-frequency weather regimes other techniques such as PCA analysis (this approach "cuts off" PCAs with the largest load excluding other circulations as "noise") might be more suitable, not our approach.

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

atmosphere in motion

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.

The normalization of the fields is necessary and essential as data for all seasons are classified. In the present manuscript, the geopotential fields are not linked to any weather phenomena such as temperature and/or precipitation.

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?

It is true for other methods, which do not employ probabilistic/fuzzy assignment to a class, as well. The focus of the sentence is on the phrase "if only by a small margin". K-means assigns all elements to the given number of classes. If this number is small, each element must be assigned to one class anyway regardless if there is any similarity to that class. As the number of classes in k-means should be defined prior to the classification, the danger is high to produce classes with dissimilar (to each other) elements. Our method builds classes from only similar elements grouping them in the first step (merging in hierarchical step) and assigns each element to only similar class in the second step (k-medoids). The requirement of similarity of elements for building classes persists throughout the algorithm.

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?

A pair of images is structurally dissimilar when it shows patterns perceived by an observer (or characterized with any structural similarity measure) as dissimilar.

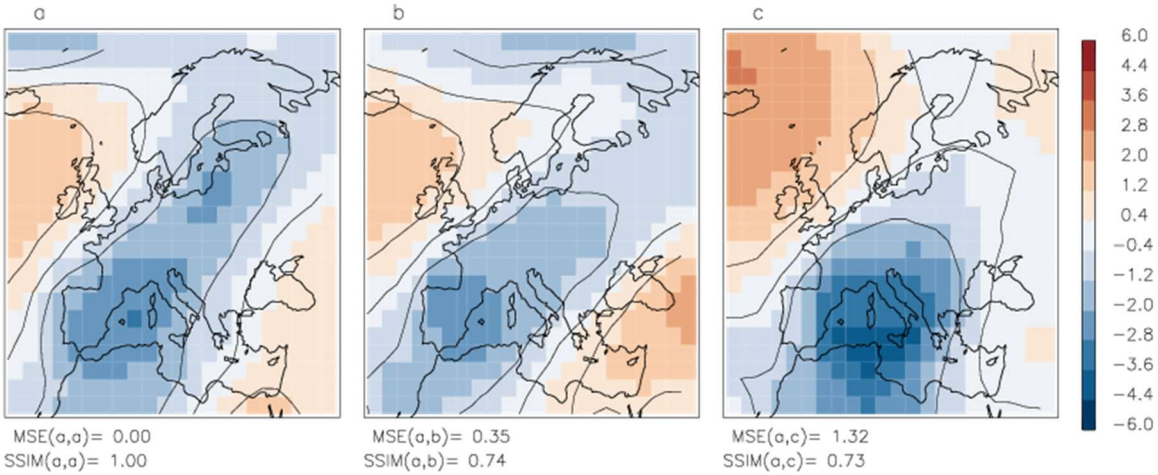
How can we distinguish the noise from the structure in any given field?

Using the SSIM index as it is constructed from three parts (structure/covariance, luminance/mean, contrast/variance) in order to detect the similarity between two images based on their shapes (covariance), match in intensity (mean) and structure/noise (variance).

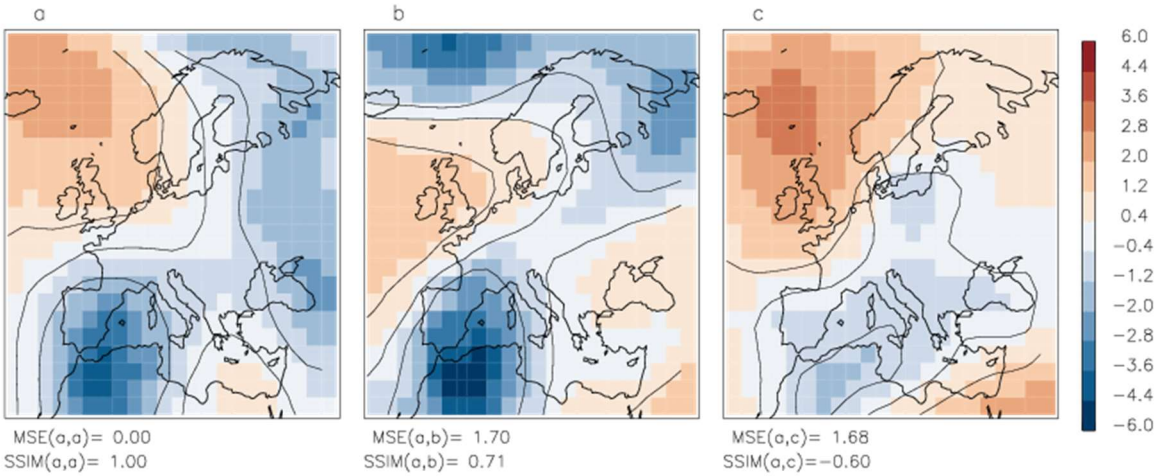
The noise/variance of an image, which can be understood as “roughness” or “texture” of the image. is included into a part of SSIM.

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?

An example of pairs of images *(a,b)* and *(a,c)* with different MSE to the reference *a* and similar SSIM: both images *b* and *c* have the same structural similarity to the reference *a*, but the MSE is different.




An example of pairs of images with similar MSE and different SSIM: both images *b* and *c* have similar MSE to the reference image *a*. But image *b* is structurally similar to *a*, *c* – dissimilar.




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

No, it does not as the “small scale noise” is only a part of the structural similarity measure (variance as contrast).

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?

MSE remains the standard criterion for the assessment of signal quality and fidelity. It is widely used in optimization routines. However, MSE is not always a good measure to evaluate signals fidelity because it is insensitive such distortions of the original signal/image as: "*a contrast stretch, mean luminance shift, contamination by additive white Gaussian noise, impulsive noise distortion, JPEG compression, blur, spatial scaling, spatial shift, and rotation*" (See Figure 2 and Figure 7 in Wang and Bovik (2009), DOI: 10.1109/MSP.2008.930649). Using MSE is justified when evaluated patterns are 1) independent temporally and spatially, 2) the error of the signal is independent of the mean signal, 3) the sign of the error plays no role in the evaluation of the signal, 4) all errors are equally important in the evaluation of the original signal. None of these assumptions holds when we process geopotential height data, temperature, precipitation, surface pressure etc.

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

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?
Similar patterns, not exactly the same, with similar frequencies.
**We suggest including a part into the manuscript that shows derived classes on randomly chosen sets of data.**

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.
Yes, we suggest adding more contour lines.

Line 245 - should there be a reference to figure 6 here?
Yes, it was "lost"

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?
Representing a cluster by a medoid guarantees that the medoid has a minimum similarity to each of the cluster elements, furthermore, it is the element with the largest total similarity to all of cluster elements. In other words, medoid is the representative element of the class. If the medoid and the centroid are similar, it guarantees that there are no or negligibly few "extravagantly" dissimilar members of that class. Otherwise, the mean (centroid) would have lost its similarity to the medoid distorted by the averaging of dissimilar members.

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

Figure 4 - Can the colour bar be included in the figure? There's room in the bottom row of panels.
*The caption of the Figure 4 says "The legend for colour shading is the same as in Fig. 3."* It can be repeated, yes.

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.
Of course the only seemingly alike looking images of a derived synoptic class and a Grosswetterlage could be arbitrary. But we compared our classes to those of James et al. (2007), which have also a similar frequency of occurrence i.e. we detect similar patterns with similar frequency of occurrence. These frequencies can be added to the text.

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

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.
**We suggest including an additional comparison of synoptic classes derived with a "standard" k-means routine, that uses MSE distance, for illustration the advantage of using our k-medoids with SSIM method.**