Dear Editor Gabriele Messori,

Thank you for coordinating the second round of review for our manuscript.

Although we appreciate comments of the Reviewer 1, we are very disappointed about comments of Reviewer 3.

Reviewer 1: *"To summarise, 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. I think this, followed by the various robustness tests and the first look at CMIP6 already present would then make any reader keen to see this approach explored further. I would also remove or substantially reword some of the criticisms of PCA based clustering that are unrelated to extreme circulations for the reasons I discuss above."*

Reviewer 1 suggests to draw more attention to rare synoptic conditions and extreme weather. We agree this would improve the paper and stronger justify purposes of constructing the new classification algorithm. We also agree to shorten and rewrite our argumentation about existing PCA-based methods.

Unfortunately, comments of Reviewer 3 indicate that we were not convincing with our arguments in his/her opinion. He/she doubts about usefulness of our tests of classification methods on the synthetic data (*"I'm not clear why the authors are using synthetic Gaussian data to compare k-means and their k-medoid method"*), despite our explanation in Lines 152-157 in the manuscript: "*The first dataset, a dataset of synthetic data, is used to demonstrate the performance of the classification method explaining why modifications to the classical k-means algorithm are necessary. We generated this synthetic data set using Gaussian shaped anomalies … to illustrate how such anomalies are treated by the classification algorithm."* The anomalies in synthetic data initially have circular shapes. We demonstrate how k-means destroys such shapes producing "distorted" (due to averaging) class centers. This effect is important to keep in mind when classifications are applied on geopotential fields: classes retrieved with k-means may show unrealistic geopotential and be non-interpretable. Our k-medoid based classification overcomes this shortcoming.

Furthermore, Reviewer 3 seems to doubt our honesty as he/she says *"I find this unconvincing - presumably the most dissimilar members are shown in the figure…"* as we show dissimilar class members in Figure 5 as a result of using MSE as similarity metric. The shown fields are just the first 15 members of the class, not deliberately chosen or pre-processed in any way for showing the failure of MSE-metric more critically as it is. Figures 3 and 5 show that MSE metric does not suit as similarity measure for our data: fields dissimilar to each other are grouped into one class. The reason for this is in the formulation of MSE – it does not account for correlation of patterns that plays an important role for grouping highly structural data. In contrast to MSE, our classification that uses SSIM does.

We wonder about the following comment from Reviewer 3 *"Finally, there is no corresponding comparison of dissimilar fields in one of the larger k-medoids clusters shown in panel 4d - do these clusters suffer from similar issues?"* as there are no dissimilar fields in classifications that use SSIM because weakly correlated fields yield a negative/very low values of SSIM and are not grouped into the same class. This follows from the definition of SSIM that includes the covariance term!

Reviewer 3: *"l295 I don't follow why having a cluster with weak anomalies would then attract more fields than other clusters with stronger anomalies"*.

This so-called snowballing effect results from the averaging of multiple class elements (= k-means classification) – see explanations in Lines 286-300 of the manuscript. This leads to iteratively weakening structures in class centers i.e. the more elements are assigned to this class the more dissimilarity between each class element to the class center is tolerated by the algorithm.

We also would like to add here: we do not re-discover deficiencies of k-means clustering with MSE as distance metric. Those deficiencies are widely known and already referenced in our manuscript. See for example "Finding Groups in Data. An Introduction to Cluster Analysis by L. Kaufman and P.J. Rousseeuw or other handbooks. More examples and discussion on the problem of use MSE as similarity metric can be learned from an outstanding paper of Wang and Bovik "Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures" in IEEE Signal Processing Magazine, 26, 98-117, 10.1109/msp.2008.930649, 2009.

Regrettably we see no further way to convince Reviewer 3 with our argumentation and would like to withdraw our manuscript with an option (Option B) to resubmit a rewritten manuscript for an independent review and discussion, and possible publication in ESD at later time.

Sincerely,

Kristina Winderlich and Co-authors