Comments of Reviewer 1 are in blue, answers of authors are in black/**black**.

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.

This comment is misleading: we did neither reproduce Grosswetterlagen nor tried to produce any set of other predefined synoptic circulations. Our two-stage clustering method derived a set of synoptic circulation patterns automatically. Some of these synoptic patterns resemble already known Grosswetterlagen. This resemblance gives us an evidence that the method is able to find known synoptic patterns, not just some arbitrary circulations (Lines 314-323).

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.

Yes, we introduce a new method for classification of synoptic patterns without prior reduction of dimensionality (PCA-based, for example) and with a new similarity metric instead of classical distance-metrics. The “similarity” parameter is intuitive as it is based on a human-perceived similarity of image pairs.

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 ERAInterim 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.

We wonder why Reviewer 1 names the Quality Index, which we introduce, “a cursory assessment of model skill”? The Quality Index (Formulae 9, Line 305) itself was introduced by Sanderson, et al. (2015) and can be computed on any similarity/distance measure. For Quality Index in this study we use the Jensen-Shannon distance measure computed on the frequency of synoptic patterns, their persistence and their transition matrix. Jensen-Shannon distance is computed on contributions of each “mismatch” between the model data and the reference weighted by its frequency (similar to Kullback–Leibler divergence) so that it is most sensitive to most frequent mismatches and least sensitive to rare ones.
If this analysis is “cursory”, we would appreciate if Reviewer 1 could make a suggestion on the analysis technique of models skill that is convincing.

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.

We introduce the new method for clustering synoptic patterns as an alternative to existing methods of clustering, which are performed on PCA-filtered data space. We developed the method to suit our purposes: evaluation of climate models including rare synoptic situations. Our approach allows accounting for rare synoptic situations, which may be linked to severe

weather (who knows?!), and to avoid PCA-related deficiencies in pattern extraction discussed by Fulton and Hegerl (2021). Huth (2021) also demonstrated that still often used unrorated PCAs result in patterns that are rather artifacts of the analysis than true modes of variability. But the main reason why we restrain from using existing methods based on PCA-analysis is that they exclude rare synoptic situations deliberately taking only few PCAs with the largest load. This approach does not suit our purpose (as we want to account for rare synoptic situations too).
In our opinion, this our method is an alternative to existing methods and it bears its own scientific value, because as the very least it corroborates previous results, but it even improves upon those previous results in both statistical (number of classes is defined automatically) and climatological aspects (all data synoptic situations are classified). We demonstrate the application of the method for evaluating of CMIP6 models (for illustrative purposes) as compared to the reference ERA-Interim over 1979-2015 period. Our purpose is not to investigate each of the CMIP6 models individually for its performance but to provide a measure that ranks their relative performance (according to the chosen reference). The manuscript is written to document this, illustrating its application, for the broad scientific community.

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.’

The analysis of CMIP6 model simulations illustrates one of the possible applications of the method for the CORDEX-EU domain. The quality indices that we provide (Table 3, Lines 473-478) show

1) how close the frequencies of synoptic patterns *QI(HIST)* produced by CMIP6 models are to the frequencies of these synoptic patterns in the Reanalysis
2) in which season of the year ($QI(HIST_{JFD})$, $QI(HIST_{MAM})$, $QI(HIST_{JJA})$, $QI(HIST_{SON})$) these frequencies are best reproduced
3) which CMIP6 model reproduces the persistence of synoptic patterns (*QI(PERSIST)*) best
4) how well the transition matrix is captured by CMIP6 models (*QI(TRANSIT)*)

We believe that findings as these, for example:

“a model X does not reproduce the correct frequency of SPs in summer”
“the transition matrix of SPs in model Y differs strongly from Reanalysis”
“a model Z fails to reproduce SP1 in winter” etc.

are valuable as they tell about particular deficiencies in the flow simulation by the models and could be addressed by 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)

Our manuscript presents a new clustering method for pre-selecting “good” models for subsequent applications such as impact-modelling etc. We do not aim to present various evaluations of as many model as possible. The computation of Quality Indicies based on the chosen reference for synoptic classes for CMIP6 models is done for demonstrative purposes. Such computations can be done for any model and reference, depending on the evaluation purpose of the user. We believe the manuscript indeed fits into the scope of the ESD journal because it contributes to the scope of the journal focused on investigations in the subject area 1."Dynamics of the Earth system" by a new concept for model evaluation in order to contribute to the model development and pre-selection for its further use such as future climate projections, impact-modelling and downscaling to smaller regions. We believe that the aim of the evaluation routine is to find deficiencies in a model's performance and not to detect its reasons (for over 30 CMIP6-models it is also not feasible as these models differ in their grid, numerics, processes resolved and drivers used). Knowledge of model's deficiencies, which we quantify, would help model developers in their future work.

> or by exploring the socioeconomic impacts of their synoptic patterns (such as by looking at their relation to energy, agriculture, extreme event management, etc.).

The evaluation of the performance of the climate model should ideally be done before impact-models are applied (and before the socioeconomic impacts are addressed with further impact-models). The main reason why we want to quantitatively access models performance independently on their subsequent application is to pre-select "the good ones".

This would of course represent another major extension to the current work.

We think that both of these suggestions are superfluous for a paper that presents a classification algorithm for synoptic situations with the aim of climate models evaluation.

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,

We are aware of problems in the clustering of high-dimensional data, such as:

1) distance measure becomes less exact as the dimensionality grows and

2) data elements may share several correlated attributes that may group them in clusters differently.

We solve both these problems by using the new similarity metric SSIM, that mimics human image-perception, instead of a classical distance measure (Lines 135-149) and by using the medoids (instead of centroids) for representing classes (Lines 116-122).

and bound to add to the issues of instability, and low representativeness of the cluster means that you comment on.

We use medoids, not the means for building clusters for exactly the reasons of stability. In a classical k-means clustering algorithms each cluster is represented by its mean. In our application such cluster mean is computed on multiple (often >600) synoptic maps (that are geopotential anomalies). This leads to a "smoothing" of such maps to a degree that the mean 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 unsimilar elements into one cluster making it even more "blur". This effect is known as "snowballing". The final set of clusters is the rather small, but each cluster is likely to include elements strongly unsimilar to each other. 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). In order to avoid the low representativeness of the cluster means we use medoids (not the means!) to represent clusters and show that the means and medoids of final classes 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).

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.

We do discuss in detail exactly why we do not use traditional PCA-based techniques to initialize clusters (Lines 123-133) in the part “3 Method” of the manuscript: *“Decision 2: use a two-stage algorithm. There are multiple ways of defining the number of classes for a k-medoids algorithm (similarly to k-means) ranging from a random guess to the analysis of the data based on principal component analysis PCA, also known as empirical orthogonal functions, Huth (2000). Lee and Sheridan (2012) suggested the initialization of the clustering algorithm by selected PCAs. The reason for this statement was the common (naïve) assumption that the first few modes returned by PCA were physically interpretable and should match the underlying signal in the data. However, Fulton and Hegerl (2021) tested this signal-extraction method and demonstrated that it has serious deficiencies when extracting multiple additive synthetic modes: false dipoles instead of monopoles, which may lead to serious misinterpretation of extracted modes. Fulton and Hegerl (2021) also found that PCA tends to mix independent spatial regions into single modes. Therefore, we back off using the PCA-based initialization of the clustering algorithm and employ another classic clustering algorithm, hierarchical agglomerative clustering (HAC), for initializing the k-medoids.”*

The PCA-based pre-filtering technique does not suit our purpose because it eliminates rare synoptic patterns from the analysis. But we deliberately want the rare synoptic patterns to be included in the analysis for three reasons:

1) they represent variance of model dynamics,
2) their frequency of occurrence may change (rare synoptic patterns becoming more frequent, for example, in the future)
3) rare synoptic situations may be linked to extreme weather events that would be falsely attributed to frequent synoptic patterns otherwise.

Reviewer 1 references to the study by Dorrington et al. (2022) on wintertime Euro-Atlantic circulations split in four main patterns combined with diagnostics on how well the tri-modal jet structure is represented by CMIP models. For this study, the small number of classes is important or even essential as it represents few a-priori known circulation modes. In contrary, our clustering method does not aim to identify known modes, it does not aim to detect few of them. We do explicitly want to classify frequent and rare synoptic patterns in separate classes.

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.
MSE has multiple serious disadvantages: it is insensitive to a contrast/amplitude stretch, shift of means, contamination by Gaussian noise, etc. as compared to structural similarity metrics when applied on data with temporal and spatial dependencies and on data where the error is sensitive to the original signal (as discussed in detail and illustrated by Wang and Bovik, 2016). Another alternative distance metric, Pearson correlation coefficient, is insensitive to differences in the mean and variance (Mo, et al., 2014). As we work with geopotential data that often reveal dependencies in time and space, as well as shifts in the mean and differing variances, we restrain from using above mentioned “traditional” distance metrics and employ the structural similarity index measure (SSIM) widely used in digital video processing software.

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.
We find this comment very valuable and are willing to include results of suggested analysis on robustness of the method on the temporal variability of the data into the manuscript.

• 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

Here we would like to remind that our study is not the first one used a seemingly large number of classes. A five-year (2005-2010) project named “Harmonisation and Applications of Weather Types Classifications for European Regions” (https://www.cost.eu/actions/733/) with participating research groups of 23 European countries produced an extensive catalogue of atmospheric circulation type classifications (cost733cat includes 17 automated classification methods and five subjective classifications, https://doi.org/10.1016/j.pce.2009.12.010) based on different methodological concepts, algorithms and parameter options. This Action systematically evaluated an extensive number of classifications within a coordinated inter-disciplinary environment and presented the results in the final project repot by Tveito et al, (2016) downloadable here (https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/deliver/index/docId/3768/file/COST733_final_scientific_report_2016.pdf). One of the statements of this action is: there is no universally optimal number of classes to represent synoptic circulation in all applications. The choice of this number strongly depends on the purpose of the classification and is often governed by the wish to reduce the numerical space of the subsequent analysis preserving the variance of the data in some degree. As the above mentioned final project report (Tveito et al, 2016) tests "three standard numbers of types: A small one with 9, an intermediate one with 18 and a large number of 27. Even though these numbers might appear arbitrary, they represent the majority of the original classifications ...". Participants of the COST733 showed a wide range of class numbers which, from few to over 40. The large number of classes is often used by methods rooted in synoptic meteorology, that give high priority to a high structural differentiation among synoptic patterns, at the same time trying to maximize the homogeneity inside classes. This attempt results in some classes, which have a small number of members or could be even empty for a different time span. On the other hand, methods that use a low number of classes may handle the pattern diversity in a sub-optimal way i.e. falsely attributing a pattern to a dissimilar class. None of these methods could be universally best suitable for all applications.

The number of classes in the present manuscript depends on the threshold of similarity between each pair of synoptic patterns: the higher the required similarity within each class, the larger number of classes will be built and vice versa. If we would aim at building fewer classes (<10) as Reviewer 1 suggests, we would have either to loosen the requirement on the in-class similarity or to eliminate classes with fewer elements. However, we estimated similarity threshold experimentally based on the human perception and loosening this threshold would mean consciously grouping patterns, which are viewed by an observer as dissimilar, into one class. The elimination of infrequent synoptic classes we avoid on purpose: we do want to retain the rare classes as they may become more frequent in historical of future climate projections and may be linked to extreme weather. These are to important reasons for the "large" number of classes we aim to use for evaluating climate models and ranking them according to their performance relative to a given reference (reanalysis in our case).

Now back to the transition matrix. In our case, about ½ of all elements in the transition matrix contain the main load 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 (is least sensitive to the "noise" from infrequent elements).

References

Huth, R. & Beranová, R. (2021). How to recognize a true mode of atmospheric circulation variability. Earth and Space Science, 8, e2020EA001275. https://doi.org/10.1029/2020EA001275

Fulton, D.J., Hegerl, G.C. (2021) Testing Methods of Pattern Extraction for Climate Data Using Synthetic Modes. Journal of Climate 34, 7645-7660.

Mo, R., Ye, C., Whitfield, P.H. (2014) Application Potential of Four Nontraditional Similarity Metrics in Hydrometeorology. Journal of Hydrometeorology 15, 1862-1880.

Muñoz, Á.G., Yang, X., Vecchi, G.A., Robertson, A.W., Cooke, W.F. (2017) A Weather-Type-Based Cross-Time-Scale Diagnostic Framework for Coupled Circulation Models. Journal of Climate 30, 8951-8972.

Tveito, O.E., Huth, R., Philipp, A., Post, P., Pasqui, M., Esteban, P., Beck, C., Demuzere, M., Prudhomme, C., (2016) COST 580 Action 733 Harmonization and Application of Weather Type Classifications for European Regions.

Sanderson, B.M., Knutti, R., Caldwell, P. (2015) A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. Journal of Climate 28, 5171-5194.

Wang, Z., Bovik, A.C. (2009) Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. IEEE Signal Processing Magazine 26, 98-117