We would like to ask the Editor to make the judgement, which options the manuscript may still have.

Below, we would like to address comments of Reviewer 1. For an easier reading, we highlight comments of Reviewer 1 in blue, our comment in black and citations from our paper in *black italic*.

We regret that Reviewer 1 suggests rejecting the manuscript. But we would like to draw editor's attention to the fact that this suggestion is based on a number of misunderstandings on the part of Reviewer 1 with regards to our manuscript:

1) "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" We did neither reproduce Grosswetterlagen nor tried to produce any set of other predefined synoptic circulations similar to them.
2) "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 vague formulation of distance between data elements, correlation of some attributes of data element that may group them into different clusters. We solve this problem not by reducing the dimensionality of data, but by using the SSIM, that mimics human image-perception, instead of a classical distance measure (Lines 135-149) and by using the medoids (instead of centroids) for more stable representation of clusters (Lines 116-122).
3) ".. 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. Yes, we comment on low representativeness of cluster means (and therefor do not use them!) and discuss the choice of an alternative representation of clusters (Lines 116-122).
4) Reviewer 1 suggests as an option to extend the paper "… by exploring the socioeconomic impacts of their synoptic patterns (such as by looking at their relation to energy, agriculture, extreme event management, etc.)." Our manuscript presents a new clustering method, not its possible applications, for pre-selecting "good" models for subsequent applications such as impact-modelling etc.
5) 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).
6) 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. An objectively "optimal" number of classes for representing synoptic circulation does not exist [4]. The choice of this number strongly depend on the purpose of classification. About 1/3 of 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 Quality Index, and vice versa, rare transitions make smaller contributions.

7) 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. This novel approach allows accounting for rare synoptic situations, which may be linked to severe weather, and to avoid PCA-related deficiencies in pattern extraction discussed in detail in [1]. In our opinion, this alternative method 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 as an example.

We appreciate the suggestion of Reviewer 1 to perform the analysis on robustness of the method to the temporal variability of the data. We are willing to include results of this analysis into the next version of the manuscript.

Our answers to comments of Reviewer 1 in detail:

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 intuitive 'maximum similarity' hyperparameter.
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.

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 in

[5] 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.
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.
The paper presents the method for clustering synoptic patterns that differs from existing methods of clustering:

- It does not require reduction of data-dimensionality such as to PCA-filter because it uses new similarity metric SSIM
- It builds clusters stable to outliers as medoids (instead of centroids) represent classes
- It does not discard rare synoptic circulations, which may be important for further analysis i.e. extreme weather occurrence etc.
- It avoids PCA-related deficiencies in pattern extraction discussed in detail in [1]

We present one of possible applications for our classification (for illustrative purposes) – the evaluation of CMIP6 models as compared to the reference ERA-Interim over 1979-2015 period. We aim to provide the solid reference and documentation of this novel classification method, 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 in the Manuscript (Table 3, Lines 473-478) show

1) how close the frequencies of synoptic patterns $QI_{(HIST)}$ produces by CMIP6 models are to 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 for model developers as they tell about particular deficiencies in the flow simulation by the models.

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) or by exploring the socioeconomic impacts of their synoptic patterns (such as by looking at their relation to energy, agriculture, extreme event management, etc.). This would of course represent another major extension to the current work.

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.

Reviewer 1 suggests an extension of the paper by either investigating "the drivers of good/bad synoptic pattern representation in CMIP models" or "exploring the socioeconomic impacts of their synoptic patterns". We consider both of these suggestions superfluous for the paper that presents an evaluation method for climate models. Firstly, because 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. Secondly, 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".

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

The dimension space of the original data was reduced from the original ERA-Interim spatial resolution to the sampled 22x22 points (every 2^{o} in latitude and every 3^{o} in longitude directions over the CORDEX-EU domain, Figure 1 Page 4 of the manuscript) according to [3] for representing 500-hPa geopotential height at the synoptic-scale.
We wonder why the size of 22x22=484 grid points is "bound to add to the issues on instability" as Reviewer 1 says. We show exactly the opposite in our paper. Increasing the spatial resolution of the classified fields to 44x44=1936 grid points and reducing it to 11x11=121 grid points yields essentially the same set of synoptic patterns (Figure 9, Page

20), which indicates the stability of the method to the horizontal resolution of the input data. *"Figure 9 shows six SP-classes at the original resolution (centre plots) and their counterparts in the low- and high-resolution sets of classes. Please note: the SP-classes are built at each resolution independently and are not just re-sampled copies of the same classes. Therefore, some discrepancy must be tolerated among the classes at different resolutions as they are medoids of independently formed classes. Despite of such discrepancies the SP-classes show essentially the same synoptic situations at all spatial resolutions*" (Lines 436-440).

Clustering of high-dimensional data poses two serious problems: 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).

Yes, we comment on the low representativeness of cluster means and discuss the choice of an alternative representation of clusters in Lines 116-122 of the manuscript. 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, Page 21): *"The similarity value between medoid and centroid for each class is computed and listed for all classes in the Table 2. The "strong similarity" between medoids and centroids for all 43 classes was found indicating the very good representability of clusters by their medoids. The mean similarity over all 43 classes is 0.84*" (Lines 453-455).

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. 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.

We do discuss why PCA technique was not used for initialization of the clustering algorithm in the part "3 Method" of the manuscript *." (*In Lines 123-133*)*: *"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."*

Additionally, using the PCA-based pre-filtering eliminates rare synoptic patterns from the analysis, but we want them to be included as 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) and 3) as they may be linked to extreme weather events and would be attributed to “common” synoptic patterns otherwise.

Reviewer 1 references to the study by Dorrington et al. (2022) on wintertime Euro-Atlantic circulations split in four main patterns, which investigates 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 stable modes of circulation. In contrary, our clustering method does not have a purpose to identify few such modes. We do explicitly want to classify infrequent synoptic patterns in separate classes.

MSE has multiple serious disadvantages (it is insensitive to a contrast 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 (discussed in detail and illustrated in [6]). Another alternative distance metric, Pearson correlation coefficient, is insensitive to differences in the mean and variance [2]. 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 “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

The frequent classes/transition-elements dominate the quality indices as we use the Jensen-Shannon distance (Equations 6-9, Pages 11-12): the influence of each signal to the final score is proportional to its frequency (similar to computation of Kullback–Leibler divergence). The usage of Jensen-Shannon distance makes the Quality Index most sensitive to the frequent classes/transition-elements and least sensitive to the “noise” from infrequent elements. There is no universally optimal number of classes to represent synoptic circulation in all applications. The choice of this number 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 study [4] 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 ...” Resulting number of classes in the present paper 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.

References

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

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

[3] 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.

[4] 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.

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

[6] 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