Comments of Reviewer 2 are in blue, answers of authors are in black/**black**.

Review of the manuscript entitled: "Classification of synoptic circulation patterns with a two-stage clustering algorithm using the structural similarity index metric (SSIM)" by Kristina Winderlich, Clementine Dalelane and Andreas Walter

Summary

The authors develop a new classification method for synoptic circulation patterns with the aim to extend the evaluation routine for climate simulations. Its unique novelty is the use of the structural similarity index metric (SSIM) instead of traditional distance metrics for cluster building. This classification method combines two classical clustering algorithms used iteratively, hierarchical agglomerative clustering (HAC) and k-medoids. The authors apply the classification method to ERA-interim and NCEP1 reanalysis, and CMIP6 models. The authors wish to demonstrate that the built classes are consistent, well separated, spatially and temporally stable, and physically meaningful. Finally, the authors rank the CMIP6 models according to their ability to represent the weather types using different quality indices.

Dear authors,

The purpose of using synoptic circulation patterns to evaluate climate models is a welcomed aim, but is not the first time this is done, as it may seem from the text. Indeed, the ability of models to capture the characteristics of synoptic patterns is an important aspect of improving climate model simulations. The SSIM is generally an interesting and seems to be promising approach for the classification of weather regimes. The article is generally well written, however it should be extended to serve as a high quality research article in ESD.
We do not pretend to be the first in the field. Therefore, we often refer to the project named "Harmonisation and Applications of Weather Types Classifications for European Regions" (https://www.cost.eu/actions/733/) that finished by the time we started developing our method. This international project joined research groups of 23 European countries to produce 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).
Pointing to this and other publications, we tried to shorten the introduction. It seems being cut too short. **We suggest extending the introduction of our manuscript by a new chapter, with an overview of existing synoptic classifications and their applications to model evaluation.**

My comments and suggestions to improve the manuscript are as follows:
General comments

- Many classification algorithms attempt to categorize weather types/regimes over the Atlantic-European-Mediterranean region. If the authors suggest a new procedure, they should at least demonstrate why their classification is better than other classification procedures. Indeed, the authors try to explain their choices, but do not demonstrate how their procedure is superior in comparison to other classifications. Perhaps the authors can randomly select days and subjectively see for how many of them the classification does a decent job? Comparing to the original classification you mention in the text would then provide a semi-quantitative way of demonstrating the improvement from one classification to the other.

We suggest including following parts into the manuscript:

1) **an additional comparison of synoptic classes derived with a "standard" k-means routine for illustrating the advantage of using our k-medoids method.**
2) **application of our two-stage classification on different sets of randomly chosen data for comparing the resulting classes.**

- Forty-three classes seems a rather large number of weather types and can probably be significantly reduced by some sort of EOF analysis. If not, it should at least be explained why the authors do not use this approach as it is very common. Furthermore, I would like to see some further explanation on how do these synoptic types relate to the four canonical weather regimes.

Our choice of not using PCA-based pre-filtering of the data is explained in detail (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."*

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. Additionally, we suggest emphasizing in the text of the manuscript that the PCA-based pre-filtering technique does not suit our purpose because it eliminates rare synoptic patterns from the analysis taking only few PCAs with the largest load. This approach does not suit our purpose because we do want to include rare synoptic situations into separate classes.

We do 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.

We introduce the new method for clustering synoptic patterns as an alternative to existing methods of clustering. Our classification method accounts for rare synoptic situations, which may be linked to severe weather (who knows?!), and avoids PCA-related deficiencies (for example, the extraction of bi-polar structures as an artifact by approximating a single-polar structure) in pattern extraction discussed by Fulton and Hegerl (2021).
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).

The reviewer uses the term "canonical" synoptic regimes probably inspired by inter-related studies of Fabiano et al. (2020) and Dorrington et al. (2022), who investigated the variability of the atmospheric circulation looking at four recurrent patterns: NAO+, NAO-, Scandinavian Blocking and Atlantic Ridge. This perspective of looking at the atmospheric circulation does not suit our aim of the model evaluation, because we do not focus on synoptic regimes (quasi-stationary states). The choice of the number of synoptic classes strongly depends on the purpose of the classification. In our manuscript, this number is rather large, as we do not eliminate infrequent classes as it is done by PCA-based classifications. The European COST Action 733 "Harmonisation and Applications of Weather Types Classifications for European Regions" (https://www.cost.eu/actions/733/) also says: there is no universally optimal number of classes to represent synoptic circulation in all applications: "*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* ..." (Tveito et al, 2016). Participants of the COST733 showed a wide range of class numbers from few to over 50.

- The CMIP6 model evaluation section in its current form is rather short and does not provide very useful information for model developers. This section should probably be extended. It would be nice to have some discussion as to why you think some models are better or worse. Additional analysis is of course welcomed, but should probably be balanced with the length of the article.

We use the CMIP6 models only for the illustrative purposes. The aim of our manuscript is to present a new classification method (as the title says; CMIP6 is not even mentioned in the title of the paper). We use CMIP6 models only to demonstrate how the models could be ranked according to the Quality Index computed on the reference set of synoptic patterns. In our opinion, such Quality Index should be used to extend the traditional set of metrics in model evaluation routine. A typical evaluation routine depends strongly on users demands and often includes bias-estimation and the analysis of extremes for scalar variables such as temperature and precipitation as stated in Lines 41-47 of OM: "*...traditional techniques for model evaluation mainly focus on individual variables and/or derived indices and do not take into account, how well models simulate synoptic weather patterns and their frequencies of occurrence (Díaz-Esteban et al., 2020)*".
Our Quality Index computed on the reference set of synoptic patterns gives additional information on model quality, firstly, to the users of the models supporting their choice for a model suitable for their applications and, secondly, to the model developers showing how well their model performs in comparison to other models. The bit of information that could be provided by our analysis to the model developers would sound like "Model X under-represents synoptic situation A in summer months" or "Model Y hast a strong mismatch in representing transition matrix in spring moths".

We believe that the aim of the evaluation routine is to quantify deficiencies in a model's performance and not to investigate/detect their reasons (for over 30 CMIP6-models it is also not feasible as these models differ in their grid, numerics, processes resolved and drivers used). This task can only be performed by the developers themselves. We would be happy to provide any modelling group with much more detailed evaluation results as to which synoptic pattern is malrepresented in which season to analyse possible deficiencies.

Specific comments

Abstract

- What do you mean with physically meaningful? There may be different meanings to physical, and you should probably clarify this in the text.

The "physically meaningful" synoptic class is a synoptic pattern known to exist in the data i.e. one that represents a realistic circulation.

- Line 10: This sentence should be at the very end of the abstract.

ok

- Do you think your classification would be useful for extended-range weather forecasts? If so, mention this and in the abstract and discuss in the conclusions.

As an optional application besides the model evaluation, a linkage of synoptic classes to extreme weather could potentially be addressed. For example, such linkage was used by Nguyen-Le and Yamada (2019), who classified anomalous weather patterns associated with heavy rainfall in Thailand and implemented classification results into a Global Spectral Model (GSM) of the Japan Meteorological Agency improving the forecast skill with the lead time up to 3-days. We are ready to extend the introduction/discussion part of the manuscript by such overview. However, we doubt that using synoptic classes in a form of "precursor" for improving a weather forecast beyond 3 days lead-time would be the best-suited instrument for improving such forecasts.

Introduction

- Line 43 – 47: From the introduction, it sounds as if you are the first and only group evaluating models based on weather regimes. However, there is an increasing body of knowledge working in this direction. To name a few articles:

References

Dorrington, J., Strommen, K., and Fabiano, F.: Quantifying climate model representation of the wintertime Euro-Atlantic circulation using geopotential-jet regimes, Weather Clim. Dynam., 3, 505–533, https://doi.org/10.5194/wcd-3-505-2022, 2022.
Fabiano, F., Christensen, H.M., Strommen, K. et al. Euro-Atlantic weather Regimes in the PRIMAVERA coupled climate simulations: impact of resolution and mean state biases on model performance. Clim Dyn 54, 5031–5048 (2020). https://doi.org/10.1007/s00382-020-05271-w
Hochman A, Alpert P, Harpaz T, Saaroni H, Messori G. 2019. A new dynamical systems perspective on atmospheric predictability: eastern Mediterranean weather regimes as a case study. Science Advances 5: eaau0936. https://doi.org/10.1126/sciadv.aau0936

We are certainly not the first to evaluate models accounting to weather regimes. Just to name a few: an exemplary study on CMIP5-CMIP6 model evaluation over multiple regional domains by A. Cannon (https://doi.org/10.1088/1748-9326/ab7e4f), a study by U. Riediger using weather types obtained with a threshold-based classification method for the Central Europe (DOI: 10.1127/0941-2948/2014/0519), a full set of more than 20 classification schemes inter-compared and described in the COST733 Action, and many others.

We suggest extending the introduction with an overview of other applications of synoptic classification methods.

- Line 58: Please discuss the number of regimes some more. There are a few articles focusing on this aspect in the literature. Some use two regimes (Wallace and Gutzler, 1981), others use four (Vautard 1990), six (Falkena et al., 2020) or seven (Grams et al., 2017) regimes. This is important as you use an outstanding number of 43.

References

Falkena, S. K., de Wiljes, J., Weisheimer, A., & Shepherd, T. G. (2020). Revisiting the identification of wintertime atmospheric circula-tion regimes in the Euro-Atlantic sector. Quarterly Journal of the Royal Meteorological Society, 146, 2801–2814. https://doi.org/10.1002/qj.3818

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., & Wernli, H. (2017). Balancing Europe's wind-power output through spatial deployment informed by weather regimes. Nature Climate Change, 7, 557–562. https://doi.org/10.1038/nclimate3338

Vautard, R. (1990). Multiple weather regimes over the North Atlantic: Analysis of precursors and successors. Monthly Weather Review, 118,2056–2081. https://doi.org/10.1175/1520-0493(1990)118<2056:MWROTN>2.0.CO;2

Wallace, J. M., & Gutzler, D. S. (1981). Teleconnections in the geopotential height field during the Northern Hemisphere winter. MonthlyWeather Review, 109, 784–812. https://doi.org/10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2

There is no universally right number of synoptic classes for all applications. Each application requires a number of classes best suitable for its purposes. A large number of classes is often used by classification 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. For example, the ZAMG-classification with 43 classes (Baur 1948, Lauscher 1985), and the Grosswetterlagen-based classification by James et al. (2007) with 58 weather types (29 for winter and 29 for summer). This attempt may produce 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.

We suggest adding a discussion part about the choice of number of classes into the manuscript.

- Line 64-66: This is a very strong critic on all prior classifications and should be further explained why none fit your purpose. These classification procedures were all used extensively in the literature. If you state this, you should at least demonstrate how your classification is superior.

Our aim is to create an automated classification scheme that gives high priority to a structural differentiation among synoptic patterns, including rare patterns in separate classes (not leaving them unclassified and not attributing them into dissimilar classes).

We suggest extending the part that explains why existing classifications do not suit for our purpose.

Data and methods

- Line 80: If you use ERA-interim and not ERA5 reanalysis, you should at least say why, and mention some of the studies comparing the two data sets. I do not expect much difference for large-scale weather regimes, but this should be at least discussed.

At the time the work on the new classification started, the ERA-Interim was the data set with the largest temporal coverage (1979-2018) and commonly used as a reference data set for evaluating climate models. Certainly, the newer data set ERA5 could be used as the new reference, once available and routinely tested for "standard" evaluation applications such as bias-estimation and extreme statistics for models scalar variables.

- Line 82: Please justify why you use 12:00UTC and not daily or all 6-hourly data.
- Line 82: How did you coarse grain the data and why to 2×3 degrees?

Answer to both comments above: The spatial and temporal resolutions of the domain were chosen as recommended by the inter-comparison project COST Action 733 "Harmonization and Application of Weather Type Classifications for European Regions" (Tveito et al., 2016): every 2° in latitude and every 3° in longitude, daily at 12:00 UTC. The model output at 12:00 UTC is often chosen for evaluation for two practical reasons: 1) it often matches mid-day peak in extreme weather conditions and 2) is a typically available time for model output. The coarse-scale sampling of the geopotential field was also suggested by Muñoz et al., (2017) as sufficient due to the fact that the synoptic-scale 500-hPa geopotential height does not require high resolution to reproduce the key physical mechanisms associated with.

- You often use 'synoptic scale', but I think it is more accurate to consider these regimes as large-scale features. I would try being more accurate on this. Perhaps change throughout the text.

We use the term "synoptic scale" consciously addressing the weather-patterns that consist of positive and negative geopotential anomalies at a horizontal scale of about 1000 km seen together at a time. We treat all weather patterns equally independent on their temporal duration so that short-term patterns are deliberately included (not eliminated) into the classification. The term synoptic regime, from our point of view, describes rather a recurrent, quasi-stationary and temporally persistent state of the atmospheric circulation that can be associated, for example, with a NAO phase. **We do not aim to detect and classify such quasi-stationary regimes.**

- Line 95: Why 151 days of smoothing? Please justify this choice.

The 151-days smoothing with equal weighting to each element was done with the purpose to produce the very smooth seasonal curve of 500hPa-geopotential and its standard deviation. Such strong smoothing allows us to preserve as much as possible of the anomaly of the 500hPa-geopotential height after the normalization.

Results

- Lines 436-440: I do not completely understand how you obtained high resolution relative to coarse resolution in figure 9.

For testing the stability of the method in space, additionally to the classes on the reference data set (2°x3°), further two sets of classes were built by a resampling of the original data (the spatial resolution of the data set is approximately 80 km, T255 spectral): on the low-resolution (4°x6°) and on the high-resolution (1°x1.5°). This explanation should be added to the manuscript.

- Line 454-456: Your motivation was not to use centroids in the introduction and methods section, but then you test your medoids and say that they are very similar to the centroids. Is this not a circular argument?

No, it is not a circular argument. We use medoids, not the means for building clusters for the reasons of stability of the classification method. 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 obtained with such routine is rather small, but each cluster is likely to include elements strongly unsimilar to other elements (and we tested 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 all 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.

Then, 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 is not surprising as any new element attribution is done to the most similar medoid (similarity of a new cluster element to other elements is guaranteed in this case). 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).

- Section 4.6: Perhaps provide some illustrations of the different classes in the CMIP6 models, in addition the quality indices in the table.

The classification is only done on the reference reanalysis data. The only classes used in the evaluation are the classes (shown in Figure 4, Page 12) derived on these reanalysis data. CMIP6 model data are not used in the classification algorithm. The output of all CMIP6 model is assigned to these reference classes using the maximum similarity measure (SSIM). Lines 284-289: "*The model output was assigned to the 43 reference classes derived from ERA-Interim and the following statistics were computed: histogram of frequencies (HIST) for SP-classes (year through), histograms of frequencies for each season ($HIST_{DJF}$, $HIST_{MAM}$, $HIST_{JJA}$, $HIST_{SON}$), matrix of transitions (TRANSIT) between available classes (frequency for each SP to follow another SP), and probability of persistence (PERSIST) of each SP for 1,2, .. 25 days. For each of these seven statistics an individual quality index (QI) is computed. The overall quality index is then computed as the mean of the seven individual quality indices.*"

If we would repeat the classification on each CMIP6 model output, it would produce different sets of synoptic classes not necessarily comparable.

- Table 3: I believe that there is not much difference between the models in the 'transit' and 'persist' values because there are so many classes. In addition, for the other indices the standard deviation is rather low, which is a bit surprising for more than 30 models. They all do pretty much the same job, which is again a bit surprising.

Absolute differences between models may be small. For the evaluation of these models their relative difference to the reference is important. As more classes are used the more differences between a model and a reference is captured (contributes to the Quality Index). Vice versa, with fewer classes – less differences are captured.
At best, in an extreme case, if we use just one single class – the difference would be characterized just by a single number.

- Are the models evaluation criteria significantly different from one another? I think you should test this.

The metrics to evaluate the classification method (chapter 3.3) are not independent; they are adapted form COST Action 733.
The Model evaluation criteria are independent. Although, the persistence and transition matrix may be viewed as dependent in some extent: the diagonal elements of the transitional matrix represent the probability of transition for each synoptic pattern to itself the next day, i.e. persistence, but not the duration of such transition. The persistence matrix represents the probability of duration of 1,2, 3 … days for each synoptic pattern.

Conclusions

- This section is rather very short and should have a bit more discussion with respect to other articles evaluating models using a classification procedure. The article would also benefit from explaining what is better or similar in the new classification with respect to other methodologies used in the literature. The potential use of this methodology in climate projections or extended-range weather forecasts should probably also be discussed.

We are willing to add more discussion on the pointed topics.

Technical comments:

- Line 82-84: Please rephrase, something is missing here.
- Line 307: This should be 'Results' and not 'Method' section.
- Line 318: Change 'gives us an evidence that' to 'provides evidence that'.
- Line 357: Change 'gives an evidence that' to 'provides evidence that'.

Figures:

- Figure 4: It is very hard to see anything with so many panels.
- Figure 10: I think you mixed up between left and right in the caption. In addition, are there significant difference in the right panels?
- Table 3: It should probably be DJF for winter in the upper row and not 'JDF'.

References

Baur, F., 1948: Einfüruhrung in die Grosswetterkunde (Introduction into Large Scale Weather), Dieterich Verlag, Wiesbaden, Germany

Hochman, A., Messori, G., Quinting, J. F., Pinto, J. G., and Grams, C. M.: Do Atlantic-European Weather Regimes Physically Exist?, Geophysical Research Letters, 48, 10.1029/2021gl095574, 2021.

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

James, P., 2007: An objective classi_cation method for Hess and Brezowsky Grosswetterlagen over Europe. Theor. Appl. Climatol. 88, 17/42.

Lauscher, F., 1985: Klimatologische Synoptik Osterreichs mittels der ostalpinen Wetterlagenklassifikation (Synoptic Climatology of Austria based on the Eastern-Alpine Weather Type Classi_cation). Technical report, Arbeiten aus der Zentralanstalt f ur Meteorologie und Geodynamik, Publikation Nr. 302, Heft 64, Austria

Muñoz, Á. G., Yang, X., Vecchi, G. A., Robertson, A. W., and Cooke, W. F.: A Weather-Type-Based Cross-Time-Scale Diagnostic Framework for Coupled Circulation Models, Journal of Climate, 30, 8951-8972, 10.1175/jcli-d-17-0115.1, 2017.

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.