Below we answer comments of the Reviewer 1.
Original comments of the Reviewer 1 are listed in black, our answers - in blue colour.

General Comments

1. While the manuscript is very detailed in explaining and testing the methodology that was developed to classify synoptic circulations the connection and application of the method to the main motivation for its development, "to extend the evaluation routine for climate simulations", is not given the same amount of detail and attention as it should. The manuscript as is should more clearly demonstrate how the method accomplishes this objective and how it adds value to the current evaluation of climate simulations that would warrant the effort required to implement it. One possible suggestion, given the length of the paper and detail provided to the actual methodology and its testing, could be to make the application of evaluating CMIP6 simulations with this algorithm as a separate manuscript where that specific application of the method can be discussed and demonstrated in a complete manner.
   Our proposed method provides a quality index that is subsequently fed into a comprehensive evaluation routing for climate simulations along with a number of other quality indices. It is explicitly stated in the manuscript that we only "demonstrate an exemplary application" (Line 19) to show the applicability of our method for evaluation. We indeed plan to write a follow-up paper on the "full evaluation" of the CMIP6 models.

2. While there is a good discussion in the introduction with respects to building synoptic classes in pervious work there was no mention of works that used approaches such as Machine Learning and AI which is becoming more popular within Earth system science as well as other fields. For example, Gervais et al. (2016) uses Self-Organizing Maps to classify Artic Air Masses from CESM-LE. I think it would be important to discuss how approaches like SOMs, Random Forrest, etc. have been used in the classification of synoptic patterns and how this new approach compares to them.
   We agree to this comment and will extend the introduction and the discussion with suggested approaches in the revised manuscript.

Specific Comments

LINE 20 – Why not state what the alternative reanalysis is instead of keeping it vague by just saying "alternative reanalysis"?
We chose NCEP1 as alternative reanalysis. But we believe that in our work it is not important which alternative reanalysis product exactly we take for demonstrating the „best" quality score. Any of available reanalysis products may be taken. The alternative reanalysis is only needed here to demonstrate the relative range of quality indices for CMIP6 models.

LINE 175 – Would this method also work if considering more than one atmospheric variable mapped on the same domain, or can it only work with the use of a single variable?
There is no "universally correct" recipe on how to build synoptic classes. As we mentioned in Lines 67-74, geopotential height and surface pressure are common variables used for the classification of synoptic patterns. But there is a variety of methods, which construct synoptic patterns on the basis of more than one variable (e.g. Bisolli&Dittmann 2001). The presented classification method can be extended to multiple variables by either targeting the optimization algorithm on a vector of similarity values, or defining the SSIM for vector-valued variables.

Bisolli, P. and Dittmann, E. (2001): The objective weather type classification of the German Weather Service and its possibilities of application to environmental and meteorological investigations. Meteorologische Zeitschrift, Vol. **10**, No. 4, 253-260

LINE 175 – For this work, one time step a day was used, is the reason for this due to computational/time constraints or are there other issues that may arise using this method with more regular time steps, such as all timesteps in ERA-Interim or even if moving to the hourly timesteps in ERA-5. If there are restrictions associated with the method and temporal/spatial resolution of data that can be used it would be good to mention them at some point.
Weather patterns are typically defined once per day sampled at 12 UTC to capture the mid-day peak in extreme weather conditions. Using more frequent output, for example 1-hourly, would increase the data volume but not qualitatively add more information on synoptic patterns as these patterns extend over scales of 1000 km and persist up to 20 days, they do not replace one another within few hours, so there is no necessity to use 1h, 3h or 6h data for classification.

LINE 195 – Its not clear why NCEP1 was chosen as the alternative reanalysis compared to other available reanalysis datasets. Why would the assumption "*Assuming that the alternative reanalysis captures the synoptic circulation of the reference data ERA-Interim better than any unconstrained global circulation mode*" be made? Can more be said about this decision?
The NCEP1 was chosen as an alternative reanalysis that covered the same period 1979-2018 as ERA-Interim, but use different models and data-assimilation routines. Any other reanalysis could be used instead. The assumption that an alternative reanalysis captures the synoptic patterns of ERA-Interim better that an unconstrained model is based on the construction of the reanalysis product: reanalysis data are updated weather forecasts initiated with the blend of past weather forecasts and the observations. Reanalysis data provide a complete and consistent picture of past weather and climate. Two different reanalysis data sets ERA-Interim and NCEP1, both assimilating real observations, can be seen as two "realizations" of the real weather/climate. Whereas a free-running (no data assimilation used) climate model can be seen as an approximation of the real weather/climate as it lacks usage of true observations. Therefore, it is natural to expect that any pair of reanalysis products that share the same observations used in their production match more closely than any pair of one reanalysis and one free-running model or even a pair of two free-running models.

LINE 198 – I am assuming all datasets are normalized with EQ. 1? Is this correct?
Yes. In Equation 1 the normalization around the 0-mean and by the standard deviation is used. It is necessary because the variance of the geopotential changes seasonally (larger in summer, smaller in winter). The normalization is done in order to be able to cluster summer and winter synoptic patterns without being over-sensitive to the higher summer variance in these fields.

LINE 375 – I'm not sure this is clear, is the "final cluster" what is used as the initialization clusters, or the final result of the entire method being presented in the manuscript?
The second. The two-stage algorithm stops, when no similar clusters are left to combine. [This is the final set of clusters.] The centres (medoids) of final clusters give the set of classes.

LINE 444 – When stating "well separated …from the entire data set" does this mean the clusters should be well separated from the data that is not assigned to the given cluster?
Cluster separation is a measure that quantifies the similarity of clusters as compared to homogeneous/random data or other clusters. We use 1) explained variation EV, 2) Euclidean distance ratio DRATIO and 3) similarity ratio SSIMRATIO. These measures characterise how clusters differ to other clusters (DRATIO and SSIMRATIO) and to the whole data set (EV).

LINE 451 – Are these "similarity diagrams" what is shown in Figure 10?
Yes. Similarity between classes derived with different merging threshold.

LINE 461 – If it has been established that using values such as Euclidean distance does not perform well when considering things such as synoptic patters what is the value in calculating Metric 2?
The metric EV is widely used to describe the separation and representability of classes in the wide community of classification methods for synoptic patterns. This metric is recommended within the project COST Action 733 report (Tveito et al., 2016) as we referred in Line 453, therefore we use it. Values of EV show that despite using medoids for building clusters, the final classes still explain a large portion of variance (although Euclidean distance was not targeted by the optimization!).

LINE 585 – To clarify, there are 183 "runs" but each run is done for varying data volumes from 1 to 40 years. So, is it correct to say the method is done 183 x 40 times? Or the output of each run is just saved after each year of data is added?
In total 183x40 runs: for each of 40 data volumes and for each of three merging thresholds 60+1 runs.

LINE 660 – It is difficult to see the dashed and grey lines in Figure 10.
We agree. We will update the Figure to make the lines more visible.

LINE 805 – While I understand the reasoning for showing the 5 most frequent SP-classes one of the benefits mentioned was the ability for the algorithm to preserve less frequent patterns that are more likely to be associated with extremes. I think it is important to demonstrate this ability/benefit. I would suggest maybe showing a couple of these patterns instead of just focusing on the most frequent SP-classes.
We agree. We will add a Figure with rare classes and corresponding extreme weather indicators to the next version of the manuscript. The maps for all SP-classes are also included in the response to Reviewer 1 (https://esd.copernicus.org/preprints/esd-2023-34/esd-2023-34-AC1-supplement.pdf).