Review of wes-2022-39

General comments

1) **On the choice of Gaussian mixture model.** I found the choice of a GMM as an alternative to PCA and other dimension reduction methods rather debatable. Firstly, I would argue that GMM is not "data-driven", a term that is usually associated with nonparametric methods such as PCA or Neural Networks. Rather, it is a model with a well specified structure who is inferred from the data, so no more data-driven than any other parametric model. More importantly, from Figure 3 and lines 264-268 it seems that the authors are proposing a GMM after a dimension reduction approach such as PCA has been applied. As such, it's not really an alternative to previous methods, but rather a second step. Was PCA performed to reduce the feature space of the GMM? Why not just using the original zonal and meridional wind as features? The choice has important implications on the spatial coherence on the clusters, see point 3).
   Furthermore, why was GMM used here instead of any other clustering algorithm? Is there anything in the PCA space suggesting that such a parametric choice was suitable? One could have adopted far simpler choices such as K-means or mean shift, which do not require the use of latent models such as GMM, which require either EM or a Bayesian framework for inference.
2) **Details and reproducibility on the GMM.** The presentation of the GMM is very short and lacks many details, which is somewhat surprising since the main contribution of this work is in the use of this model. A few examples: is the covariance matrix parametrized? This may not be simple task to do, since the covariance is defined in the feature space, not in an actual physical space, so the concept of distance is rather artificial. On the other hand, a nonparametric estimation of M $D \times D$ matrices with $D = 20$ would require a substantial number of time replicates to achieve some stable estimate, so if this was the approach, I would expect these estimates to be very unstable. Also, how was the EM algorithm performed? Since this is one of the main

points of the paper, a much more detailed and formal exposition would have been necessary.

3) **Lack of spatial coherence of GMM clusters.** Figure 3 shows the schematic algorithm, but it's rather deceptive, as the final outcome will not result in spatially coherent clusters. This is because the clustering is performed in the feature space, not in a physical space. As such, I believe the original clusters will actually be scattered all around the domain, a clearly unappealing feature which prompt the natural question: why not using just the original data for clustering?
4) **Some clarification on the problem statement**. Using the language of design of experiments, I believe what the authors are proposing is to discretize an I-optimal design for a spatial design of experiment, which is continuous by nature. In more abstract terms (I remove the time component as it is ultimately unnecessary for understanding the methodology), we denote by $\boldsymbol{X}(\boldsymbol{s})$ the true spatial process for a location $\boldsymbol{s} \in \boldsymbol{\Omega}$ for a domain $\boldsymbol{\Omega}$. Given a set of observations $\chi = (\boldsymbol{X}(\boldsymbol{s}_1), \dots, \boldsymbol{X}(\boldsymbol{s}_D))$ at locations $\boldsymbol{s}_1, \dots, \boldsymbol{s}_D$, the authors aim at finding the optimal χ that minimizes

$$\int_{\boldsymbol{\Omega}} \left(\boldsymbol{X}(\boldsymbol{s}) - \widehat{\boldsymbol{X}}(\boldsymbol{s})\right)^2 \mathrm{d}\boldsymbol{s},$$

where $\widehat{\boldsymbol{X}}(\boldsymbol{s}) = E(\boldsymbol{X}(\boldsymbol{s})|\chi)$, i.e., is the conditional expectation of the process at a generic location $\boldsymbol{s}$ given the set of observations. Such conditional expectation can be computed under any choice of model, and the authors refer to a data-driven basis decomposition such as EOF extrema and QR, or to a EOF + clustering algorithm based on a GMM. Under this reformulation it is natural to ask why was the problem discretized on a grid in the first place. While I understand the necessity of using climate simulations as the baseline data for assessing the design, such choice would lead to different choices depending on the dataset used. If, as the authors suggested in the final part of the manuscript, we would use ERA5 instead of AROME, would the different results emerge because of the different information in the wind process, or because of the reanalysis being on a different grid?