Review of manuscript

"A method for assessment of the general circulation model quality using K-means clustering algorithm"

by Urmas Raudsepp, Ilja Maljutenko

General comments:

Including machine learning approaches for multi-model validation has been a long-standing discussion in CMEMS. For over ten years, the four clusters validation approach has been used as a general guideline. I'm delighted to see that progress is being made with machine learning methods. Thus, there is no doubt that the subject is worthwhile of investigation. However, the study given in this manuscript did not convince me to utilize this method directly, for example, in the CMEM QUID report. My primary worries are as follows:

1. Clustering techniques are frequently used to evaluate atmospheric models, biogeochemical models, and so on. The variables in those models are multidimensional and, to an extent, "colossal." Typically, the output of an ocean circulation model is not regarded as a massive dataset. To persuade me to experiment with various clustering approaches based on machine learning, the interpretation of the clusters should be striking.

2. Prior impacts on clustering approaches, particularly hierarchical clustering methods, should be acknowledged. Without a doubt, comparing hierarchical clustering against centroid-based clustering is worthwhile.

3. The cluster interpretation should emphasize the distinct outcomes using the Taylor and target diagrams. At the moment, I see no evidence of new information being obtained (my last comment).

4. The Baltic Sea is very special. The salinity is significantly lower than that of other marginal seas, and interaction with the open ocean is extremely limited, among other factors. I have my doubts about the method applied to the Baltic Sea being universally applicable; yet, this should be discussed.

As a result, I recommend that the authors pursue two revision strategies for the paper. One possibility is to include more model data (sea level, mixed layer depth, currents, sea ice, and possibly heat fluxes and runoffs), or to use multiple models (at least two, another one can be CMEMS results). This way, I can determine the method's reliability. Another possibility is to incorporate additional clustering methods, such as agglomerative hierarchical clustering (bottom-up), divisive hierarchical clustering (top-down), or 'soft' K-means clustering (distribution-based) vs. rule-based methods (geographic areas, etc.). Clustering evaluation enables the acquisition of beneficial best-practices for clustering analysis. I believe that the work in these two areas does not require much time, and hence I recommend a major revision.

P-Page, L-Line

Introduction:

P3, L40-L41: The rationale for using clustering methods is unclear. The shortcoming is that those papers did not include enough information in data? What is 4 dimensional information embedded? For instance, vertically, even if the vertically resolution in the observation is 1 cm, but you still bin to the resolution of 5m, don't you? You did not include more information than traditional methods. I feel that the problem of standard statistical metrics (Taylor and target diagram) is their inability to express clustered error statistics, such as error in climatology, seasonal, or diurnal signals. By the way, what are your criteria for defining 'the huge dataset'?

P3, L49: It appears as though this 'K-means clustering algorithm' has fallen from the sky. This section should contain an introduction to conventional clustering algorithms. There is something missing at the start of L50.

P3, L60-64: This section should be in the 'discussion or perspective'. Why in the 'introduction'? Perhaps some previous efforts have already made used of it in an operational mode? Then they should be cited.

P3, L68-70: This article discusses the results for the entire Baltic Sea. Other validation studies of GETM in the Baltic Sea, not just in the Gulf of Finland, should be cited.

Materials and Methods:

P4 Why this subsection 2.1 is in 'Methods'? It should be in the introduction part, and review of the Baltic Sea dynamics should be included, with a reference to the discussion in the subsequent section on 'adopting this method with caution' in other seas.

P6, L120, What is meant by a 'preliminary' check? That is, by examining Fig. 1a?

P6, L127, 'This complicates data collection.' What does it mean? Perhaps you mean 'gathering of data during winter is very complicated'?

P6, L140, 'The squared Euclidean distance' is also coming from sky. Is that different clustering measures should be introduced and the reason to not choose non-Euclidean measures should be clearly stated.

Results:

P12, Figure5d, the dramatic change of clusters in recent years, e.g. big increase of K1, is it because of the smoothing you applied? BTW, add the meaning of pK in caption.

P16, Section 3.4: Interpretation of the clusters. My concern 3 reflects the issue raised in this section. Almost all of the problems in this section can be well-defined using traditional methods and have generally recommended solutions, e.g. poorly

simulated thermocline (increasing vertical resolution), Baltic inflow problem (increasing bottom inflow), Danish strait problem (too close to open boundary), river temperature problem (no easy solution), SST problem (bulk formula). Nothing novel! I would anticipate more new information if authors include more data than T and S. While one may argue that this is not critical, if not the primary need of GMD, it gives me, as a modeler, the feeling that this method is unnecessary.