REFEREE REPORT:
A METHOD FOR ASSESSMENT OF THE GENERAL CIRCULATION MODEL QUALITY USING K-MEANS CLUSTERING ALGORITHM
U. RAUDSEPP, I. MALJUTENKO

This paper suggests using K-means as a method for clustering error estimates of the General Estuarine Transport Model (GETM), an oceanic general circulation model, to find meaningful spatial structure of model error. Error estimation is done by taking the difference between true values of daily averaged temperature and salinity found in the EMODnet Chemistry database and simultaneous values generated by the GETM for a 3 dimensional spatial region (x, y, z) around the Baltic sea and time t. Once error values for temperature $dT = dT(x, y, z, t)$ and salinity $dS(x, y, z, t)$ are found, their unique pairs (dT, dS) are plotted in the $\mathbb{R}^2$ clustering space. K-means is then performed using the euclidean metric in $\mathbb{R}^2$. Since each pair is uniquely identified by (x, y, z, t), the clustering result can be evaluated in the original (Lat, Lon) space and time.

It is obvious that the authors have a well-established understanding of the dynamics of ocean flow and the available models in this field as shown in section 2.1-2.2. Furthermore, the idea of using error structure to determine where possible biases exist geographically is meaningful. However, there are several key aspects of the method that has been introduced that lead to invalid or uninterpretable results of K-means. Because of this, I do not believe at this time that the study is complete enough for publication. I list the reasons below. Additionally, I would suggest the authors look into some statistical learning literature such as The Elements of Statistical Learning, by Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie.

1. **The position of initial centroids for K-means is uniformly distributed and does not change in the paper.** The K-means algorithm is heavily dependent on the choice of initial centroids. Hence, the algorithm may reach a local minima for the distance metric but it cannot be said that this minima is the global minima (or the "most optimal"). To overcome this dependence, the K-means algorithm is often run many times with randomly placed centroids. If the clustering outcome remains consistent, then it is seen as a relatively reliable outcome for the algorithm.
2. **There is no obvious "elbow" in the plot of the distance.** Although this step can be thought of as suggestive, an elbow in the data should correspond to the number of clusters $K = k$ such that the change in rate forms a jump in the rate function at $K = k + 1$. If the plot of minimum distance over K is connected across the histograms in figure 4, it has a smooth exponential decay which indicates that there is no obvious cluster structure for K-means.

3. **Halting the K-means process after 100 iterations does not "ensure convergence".** Line 141 states that the number of iterations is limited for the numerical implementation of K-means to 100. This is not good numerical practice as it is not guaranteed that you will reach the local minima for the set of initial centroids in this number of iterations. Instead, it is numerically common to set a threshold for example, $O(10^{-4})$ so that if the change in the distance metric is less than this for a given number of iterations then it is assumed to converge.
4. **K-means requires the clusters in the space to be able to be divided by hyperplanes (lines in the case of $\mathbb{R}^2$) and the distribution of error makes this difficult.** Contrary to the statement made on line 173, error representation can actually provide some insight on the possible structure of clusters in the $\mathbb{R}^2$ error space. It is common in statistics to assume that the error of a model (like that of temperature and salinity in this case) is the sum of many independent errors, which by the central limit theorem tend to the normal distribution with mean zero and variance σ^2. Of course there are exceptions to this in the case of dependence or outliers. Because we can reasonably assume a normal distribution for the errors of both temperature and salinity, we know if we plot these values on $\mathbb{R}^2$ that the majority of the points will be in the center with the number of observed points away from the origin decaying by the variance of their normal distribution. For an illustration, see a projection of the bivariate Gaussian distribution. This is also seen in figure 2.

   The reason I mention the normality of the error distributions for temperature and salinity is because it is not possible to appropriately divide this type of resulting cluster structure by hyperplanes unless you increase the number of clusters K to some very large amount. In fact, the approximation of a circular cluster with a center can be seen in figure 3 as the number of clusters K increases. Given this, I would suggest that the authors look at other possibilities of clustering such as kernel K-means where the divisions can be made in a functional space. This may also give some reason as to why there is no obvious elbow in the distance vs. K graph.
5. **K-means clustering does not add any information on the structure of the data. Many of the results in section 3.2 can be found without clustering.** The spatial locations provided in figure 5 and discussed in section 3.2 where over- and under- estimates of temperature and salinity occur in the model can be found by setting a threshold of over- and under- estimation, say 2 standard deviations away from the mean, for each measurement and calculating the proportion of points in the (lat, lon) space that fall above or below the set thresholds.
6. **The process described in 3.3 just describes the continuation of the K-means algorithm.** The K-means algorithm begins with a set of random centroids, assigns points to the centroids based on their proximity, recalculates the centroids based on the mean, and continues this way until a local minima of the inter-cluster distances is reached. If the algorithm is run on a uniformly selected subset of points coming from a fixed distribution (which is true in this case), the reassignment will

continue until a local minima is obtained. If more points coming from the same distribution are added and the end centroids are used, the algorithm will continue from its final centroids until it reaches the exact clustering structure that is unique to the starting centroids, this is the case with figure 7 (a) (d). If less points are added, the structure remains constant (but the same since it corresponds to the same starting centroids), this is the case with figure 7 (b) (e) and (c) (f). This does not provide information on the *stability* of the clusters (line 274-275) which depends on changing the starting centroids and performing multiple runs of the K-means algorithm.

Additionally, there are some other minor issues with the article such as grammatical issues in the switching of "the" and "a" as in the first sentence of the abstract, present tense writing should be used when describing the work done for this article, and some spelling issues (e.g. line 50). There is a lack of literature on clustering methods with Jain (2010) being the main reference. As I mentioned, I would suggest starting with the book by Friedman, Tibshirani, and Hastie for a foundational understanding of unsupervised learning algorithms.