Dear reviewer

We very much appreciate your comments on the unsupervised machine learning aspects of the model validation method. We are familiar with the book “The Elements of Statistical Learning. Data Mining, Inference, and Prediction” by Trevor Hastie, Robert Tibshirani, Jerome Friedman.

The model errors do not have obvious clusters unless there are obvious errors in the model. Our clustering strategy was to get meaningful spatio-temporal distribution of the model errors based on the distribution of data in the clusters. In addition, common statistics that have been used so far in the validation of the models can be calculated for each cluster. We wanted to keep the clustering procedure relatively simple and easy to implement for the wide audience who are not experts in the field of unsupervised machine learning. The other aspect of using the K-means, instead of the other algorithms, e.g. kernel K-means, was the relatively low computational time. Usually, in ocean model validation we deal with a huge dataset.

We can drop the idea of using uniform distribution of initial clusters, i.e. to use random clusters and run the clustering until the clustering outcome remains consistent, and use $O(10^{-4})$ criterion for the convergence. We can change the corresponding text in the manuscript. Our tests showed, while doing research and preparing the manuscript, that changes in the results are minor.

Considering your concerns:

1) We have made experiments with a) random selection of the initial centroids, b) random selection of the initial centroids in the range of min/max data rectangle and c) uniformly distributed initial centroids as described in the paper. In all cases centroids converged almost to the same locations. To use uniformly distributed initial centroids was merely suggestion to start with and check if meaningful clusters in terms of numerical model under consideration occur.

2) We agree that there is no obvious elbow in the plot of the distance. But the distance does not have a smooth exponential decay. We calculated the first and second derivative of the distance as function of the number of clusters, which suggests using of 2 or 4 clusters.

Indeed, we agree that there is no obvious cluster structure in our dataset. For us was challenging to implement clustering algorithm to the dataset where there were no obvious clusters. Our aim was to see is there meaningful clusters in the context of the application.

3) We agree that halting the K-means process after 100 iterations does not ”ensure convergence”. All our tests showed that K-means process converged after 100 iterations. Thus we used 100 iteration as an indicative number of iterations.

4) We agree that in case of “ideal” numerical model the errors should be independent and tend to the normal distribution with mean zero and variance σ^2. Error distribution in Fig. 2 has the features of Gaussian distribution, but in Table 1 for K=1, the means of dS and dT are neither zero nor equal. This already provides information about the model quality. Similar argumentation holds for the STD.

We cannot agree that approximation of a circular cluster is a good approximation, as the error distribution is skewed towards positive dT. In addition, if we assume circular clusters, then we loose relevant information about spatial and temporal quality of the model. Very roughly, if we presume that cluster k=4 is one cluster and cluster k=1,2,3 is the other cluster in Figure 5, then we do not get information that salinity is either over or underestimated while temperature is “correct” in the

southwestern Baltic. In addition, we lose information on the vertical and temporal structure of the errors.

5) We disagree that K-means clustering does not add any information on the structure of the data. The distribution of data is not Gaussian and to use arbitrary number of std for threshold is not justified. In this application we have shown that clustering provides meaningful information on the spatio-temporal distribution of the model errors. This is relevant for the interpretation of the model results and for the future improvement of the model.

6) We completely agree with this comment. We wanted to show the order of magnitude of comparison data that is needed for assessment of the model in present application. If the numerical model performance is stable, i.e. the data comes from the same distribution, then the location of the centroids does not change. But if the model quality "drifts away" from its initial quality, then the location of the centroids changes, which will be a warning signal for the uses of continuously run model like near-real-time ocean forecast model.

The sentence (line 274-275) "The rough estimate of the number of comparison points is about 100 000 for the current model, which shows relatively stable centroids and the stability of the model accuracy." is more correct.

The use of random location for initial clusters and performing multiple runs did not change the results.