Answers to review's comments

Dear reviewer, thank you for your valuable comments. Below we have addressed all of you concerns.

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.

In the elaboration of the K-means clustering method for assessment of the ocean general circulation model (GETM in particular case), we used two essential variables – salinity and temperature. These variables "integrate" temporal and spatial dynamics of the water basin that has been modelled. Usually temperature and salinity are measured simultaneously. To form error space of the model, the assumption is that different variables are measured simultaneously. Already now we had more than one million data pairs. In the interpretation of the error clusters, we limited ourselves to the main physical features of the Baltic Sea. It is known that the circulation models have problems in reproducing the highlighted dynamics "*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)"*. These features were clearly shown by the K-means clustering method, which shows the applicability of the method in assessment of the model quality.

In the assessment of atmospheric models, the set of simultaneously measured variables could be pressure, temperature, humidity, (wind speed), which forms 4-dimensional error space. Indeed, then the number of error quadruplets is much larger. In the marine biogeochemical models, essential variables are nitrate, phosphate and dissolved oxygen, which are usually measured simultaneously. These variables somehow "integrate" biology and chemistry of the model. In the coupled physical and biogeochemical models, it is natural to form 5-dimensional error space (temperature, salinity, nitrate, phosphate and dissolved oxygen) for the assessment of the model system, as biogeochemistry depends on the physics, also. For different models, geographical region and time period, the number of multidimensional error points could be very large.

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.

We tried to perform agglomerative clustering for the whole dataset. The outcome was that too much computer memory and computational time was needed. We performed agglomerative clustering for the surface layer data, only. There was no significant difference between the results of the K-means algorithm and hierarchical clustering algorithm (except computational resources) for the surface layer data. Thus, we consider that using K-means algorithm is computationally more feasible than using hierarchical clustering.

In cases of the K-means algorithm we have to select the number of clusters, while in case of hierarchical clustering the distance between clusters should be predefined (Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, 745 pp.), which is not straightforward. In the latter case, for selection of the number of clusters, we had to plot distribution of the clusters in the error space and decide if the clusters have reasonable oceanographic meanings. Usually, dendrograms are used for the visualisation of the

results of hierarchical clustering, but in our case with the data number of O(10^5), the visualisation of the results is not straightforward. Using of different algorithms for hierarchical clustering might be more justified, but comparison of different clustering algorithms is not the scope of this paper and requires separate study.

Without clustering our intuition for error clustering by setting thresholds... would be ambiguous. Therefore, we would need ML algorithm which would learn from data.

We add several sentences concerning hierarchical clustering methods. Main disadvantage of the hierarchical clustering methods is that they require more computational time and computer resources than K-means clustering algorithm.

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).

Some preliminary assumptions are needed to perform model validation using Taylor or target diagram. These methods require that existing measurements are somehow spatially or temporally grouped, e.g. we select all measurements over certain geographical area, calculate the statistics and present it as one point in the diagrams. This procedure will be applied for different regions or depth levels so that set of points will be displayed in the Taylor diagram. When applying this method, then the information about model performance within the spatial domain is lost. Using K-means clustering algorithm, the spatial and temporal (+seasonal) analysis of the errors is new (Fig. 5), for example. In comparison, Kärna et al. (2021) (Kärnä, T., Ljungemyr, P., Falahat, S., Ringgaard, I., Axell, L., Korabel, V., Murawski, J., Maljutenko, I., Lindenthal, A., Jandt-Scheelke, S., Verjovkina, S., Lorkowski, I., Lagemaa, P., She, J., Tuomi, L., Nord, A., Huess, V., 2021. Nemo-Nordic 2.0: Operational marine forecast model for the Baltic Sea. Geoscientific Model Development 14(9), pp. 5731-5749. doi:10.5194/gmd-14-5731-2021) used conventional methods for validation of the NEMO-Nordic 2.0 circulation model. Their results on the spatial distribution of the model errors are presented on Fig. 8.

The second point how the Taylor and target diagrams differ from K-means clustering is that in case of Taylor diagram all variables are treated independently of the others. For instance, the statistics for salinity and temperature are calculated separately and form two points in Taylor and target diagram. In K-means a location of a single centroid is found, which represents model errors for interdependent salinity and temperature errors.

The K-means algorithm enables to assess the model performance over entire model domain and in time. For instance, Fig. 5. shows that at the eastern side of the Bothnian Sea, in certain case, the model overestimates temperature (cluster k=3), while being more correct in the open part of the Bothnian Sea. This information cannot be obtained if we use Taylor diagram, unless we calculate error statistics for the eastern coastal area of the Bothnian Sea and open Bothnian sea separately and present it in the Taylor diagram.

In a specific example, presented in Fig. 6a, we evaluate the model performance at the monitoring station BY15. K-means clustering approach, implemented on whole dataset, shows that below the halocline (depth>60-80m) the model underestimates salinity (errors belong to the cluster k=1) from 1966 to 1989. From 1990 to 2003 model has correct salinity, but temperature is slightly overestimated (errors belong to the cluster k=2). This information cannot be extracted from Taylor diagram, unless we calculate salinity and temperate errors for different depth intervals and different time periods, i.e. 1966-1989 and 1990-2003.

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.

We agree that the Baltic Sea is different from the other marginal seas and the ocean. Still, we cannot follow the argument by reviewer that the method we propose could not be applied to the other seas or ocean. For instance, the same metrics is used for different seas (incl. the Baltic Sea) and for the ocean in CMEMS. If the reviewers concern is small salinity variability of world ocean or the other coastal seas compared to the Baltic Sea, then this should not impact clustering of the normalized salinity errors. To validate the proposed method for the other seas is a separate task.

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),

In current stage of the elaboration of the K-means clustering algorithm for the assessment of the general ocean model quality (GETM in particular case), we form an error space using the set of **simultaneously** measured variables (temperature and salinity). Sea level, mixed layer depth, sea ice concentration and or thickness, heat fluxes are 3 dimensional (2D in space and time) fields. We use temperature and salinity, which are 4D (3D in space and time). Some of the suggested variables are not directly measurable (mixed layer depth, heat fluxes except solar radiation). It is rather difficult to obtain simultaneous measurements of sea level height, ice parameters unless we are limited to the coastal sea. River runoffs are completely different type of variable. Currents are measured at very selected locations and time and not necessarily simultaneously with temperature and/or salinity. We agree, that these variables could be included in the assessment of the models using K-means algorithm, but in future work.

or to use multiple models (at least two, another one can be CMEMS results). This way, I can determine the method's reliability.

We have used proposed K-means methods for the assessment of model quality in two papers, one is published and the other one is currently under revision. Indeed, both of the applications deal with the Baltic Sea.

In paper by Kõuts et al. (2021) (Kõuts, M., Maljutenko, I., Elken, J., Liu, Y., Hansson, M., Viktorsson, L., Raudsepp, U., 2021. Recent regime of persistent hypoxia in the baltic sea. Environmental Research Communications 3(7), 075004. doi: 10.1088/2515-7620/ac0cc4) we used proposed method for the assessment of coupled physical and biogeochemical model reanalyses data. The reanalyses data belong to the CMEMS multi-year product. The error pairs were formed for salinity and dissolved oxygen. In the paper by Kõuts et al. (2021), both the proposed method, common statistics and Taylor diagrams were used. The paper showed that more general picture of the model performance can be obtain with the proposed K-means method than with using Taylor diagram.

In paper by Raudsepp et al (under revision), we assessed the quality of the NEMO-Nordic 2.0 model performance (used in the CMEMS for near-real-time product) in reproducing surface temperature and salinity fields in comparison with ferry-box measurements along the ship track in the Baltic proper. The results showed that either model or ferry-box data cannot be trusted at the entrance area to the ports, especially in the southern Baltic Sea. This result could be intuitive, but in the study, we have shown it based on the data.

We provide reference to the paper by Kõuts et al. (2021) in revised manuscript.

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.

We have done the experiments with agglomerative hierarchical clustering and with divisive hierarchical clustering. Main concern by applying these methods is that these methods are not so robust as the K-means clustering is. In addition, these methods need much more computational resources. We have used the other K-means algorithms as suggested by reviewer 1 and found no significant differences in the results. Rule-based algorithms have assumptions that follow prior knowledge of the rules, i.e. geographical regions, or use the other machine learning algorithm to define the rules.

In conclusion, the proposed method is simple and robust, feasible in terms of computer resources required and contains information for general assessment of the model quality as well as for task oriented posterior analysis. We address the concern of the reviewer in revised manuscript.

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'?

We explain the advantages of clustering methods more clearly.

4 dimensional information is that error pairs can be mapped back to the (x,y,z,t) space for posterior analysis after clustering is done. We have interpolated the model data to the exact location and time of the measurements as they are in the database. Vertically, the 5-m bins and horizontally 25 km^2 grid are used for the analysis of the clustered errors.

We refer to the Fig. 5 in our study and Fig. 8. by Kärnä et al. (2021), as well as paper by Kõuts et al. (2021) to decide about the information that is obtained by traditional methods and the method proposed by us. Much more information can be obtained from the proposed method during postprocessing. Our aim was to show how the methods performs in obtaining general information on the model quality.

In the current context "the huge dataset" is dataset where the implementation of machine learning methods helps to extract and understand the information.

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.

It has not fallen from sky, but has been adopted from clustering literature/text books (e.g. Hastie et al., 2009), where it has been straightforwardly introduced as robust and easily understandable to wider audience.

We rewrite this section adding the introduction to conventional clustering algorithms.

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.

We move this part to the discussion.

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.

We also cite the other application of the GETM model in the Baltic Sea.

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.

The subsection 2.1 is moved to the introduction and we include short review of the Baltic Sea dynamics. Still, it is somehow unclear, why this method cannot be adopted to the other seas. In the future study, it is aimed to apply this method to the other European seas included in CMEMS.

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

Yes. We write it more clearly in the revised manuscript.

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

Yes. We rewrite it as suggested.

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 nonEuclidean measures should be clearly stated.

We include different measures in the description and justify why we use squared Euclidian distance. The square Euclidian distance is commonly used as the first choice of the measure of the distance, if not justified otherwise. We like to note, that we have normalized the salinity and temperature errors to make clustering independent on the data units. Thus, the clustering is performed to normalized errors, but the results are presented in original units. This also stated in the manuscript.

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.

The dramatic change of clusters is not due to smoothing. It is seen in Fig. 1b that number of measurements has increased at that time. This increase is mainly due to increased number of measurements in winter season. In winter, large volume inflows to the Baltic Sea occur. Model underestimates the salinity of these inflows and spreading of the water downstream in the Baltic Sea.

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).

We have provided evidence that our method provides information about model quality over entire spatial modelling domain and in time. It takes into account interdependent variables that describe the model performance in general. Also, we have shown that posterior analysis can provide information on model performance in specific area and time period. All this information cannot be obtained by using Taylor or targeted diagrams. On the other side proposed method is simple and robust. The interpretation of the results is straightforward concerning intuitive knowledge of the

modellers, but provides quantitative measures. Posterior analysis could fetch out different type of information on particular region of the model and time period of interest.

Computationally this method is feasible and can be applied on “colossal” datasets. We have provided postprocessing and interpretation of the result in different levels.

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.

We present new method here not investigate the new findings from model. That's why we select model where we know main errors a-priori. By using K means single-handedly, we have identified all known errors without examining each region and depth layer separately (see. Maljutenko and Raudsepp 2014).