Response to Reviewer#2

> Dear authors, congratulations for the performed work.
>
> The manuscript that you presented includes a really interesting and robust technique (GMM) that was used for oceanic regionalisation considering SSH and presenting accurate results. Additionally, I find this technique really promising because it can be applied considering other variables (SST, currents, SSS, chlorophyll concentration, turbidity, ...) from different databases (in situ, remote sensing, numerical models).
>
> The manuscript is really well writen, in good English, easy to follow and to understand. The figures are clear and necessary, the conclusions are in the line of the obtained results and the references are up-to-date.
>
> I have some comments that I expect could help to improve the manuscript.

We thank the reviewer for their supportive review and extremely helpful questions and comments. Their contribution was crucial in the improvement of the manuscript and is acknowledged in it (see acknowledgements). In the following text we answer each of the comments individually.

Major comment

> The mayor comment that I have is that I missed the explanation of why these regions were split (Figures 3 and 4) and if they correspond with bathymetry/hydrodynamic characteristics. The technique is really good and the results are promising, but I miss here a little about the physics of the regions, justifying why the classification method selected those regions and which specificities each one of them has that differs from the others, reinforcing and validating the obtained results.

Thank you for your recommendation. We originally tried to somewhat refrain from it because GMM is not a physical model and based on purely its results, without using other methods, we cannot explain the physical processes; besides, the physical drivers of sea level variability are the topic of another manuscript of ours, currently under review. But you are right that the manuscript lacks ocean science, so we added more discussions for some of the regions and dominant processes in them, based primarily on other people's works. See also our responses to similar comments made throughout the manuscript by Reviewer#1.

Other comments

> Line 80: I would like to ask the authors why they selected the period 1995 - 2019. The authors explained why they start in 1995, but not why they end in 2019. The selected database is now available until August 2022. I agree to have entire years, so to not considered 2022. But, why the authors did not considered 2020 and 2021?

The honest answer is: Because we started this work a long time ago, when only data until mid-2020 was available. But since now not only newer data is available, but the processing chain for the data set had been modified since we originally downloaded the data set (see description of the changes to the processing chain at https://catalogue.marine.copernicus.eu/documents/QUID/CMEMS-SL-QUID-008-032-068.pdf) we decided to re-do the experiments with the new expanded and improved data set. This has caused some changes in the classification results (e.g., needing a different number of EOFs to achieve the same number of classes or losing some of the classes), stemming from both the processing changes and from the change in the time span, visible by comparing the results with the old 1995-2019, new 1995-2019, and the new 1995-2021 data sets. You can see the new results

(replacement for Figs. 3 and 4 in the manuscript) in Figs. R1 and R2 and compare them with the results obtained with the 1995-2019 time period with the new data set on Fig. R3 (intermediate model only), as well as with the results with the old data set from the manuscript.

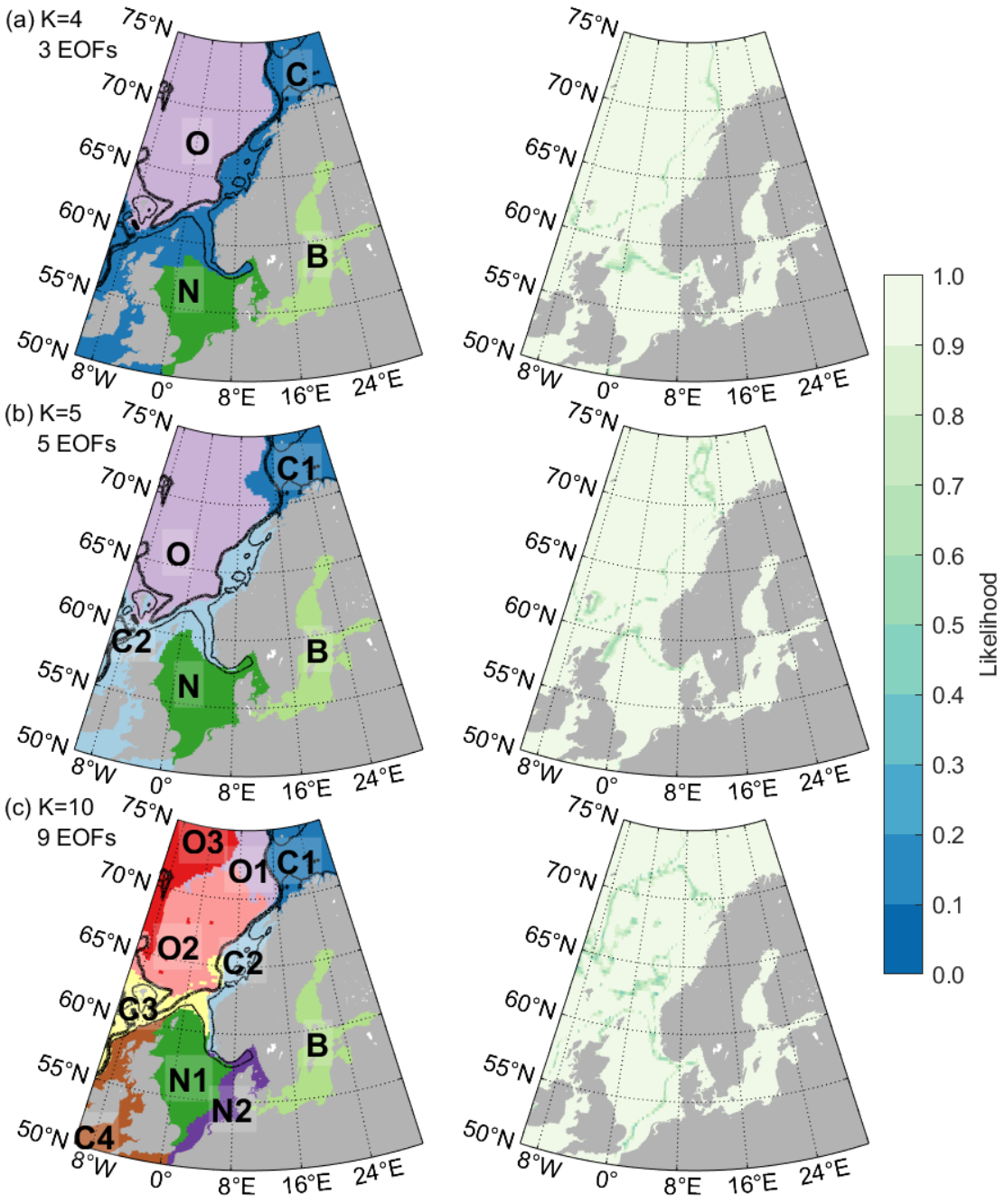


Figure R1: Replacement for manuscript Fig. 3. Classification using an ensemble of 200 Gaussian Mixture Models (left) and the respective likelihoods of the model sorting the grid points to that particular class (right). Classification is performed using 3 (a), 5 (b), and 9 (c) empirical orthogonal functions and 4, 5, and 10 classes, respectively. Letters indicate the names used to refer to the regions in the core of the text. Contour lines represent the 250 and 1000 m isobaths.

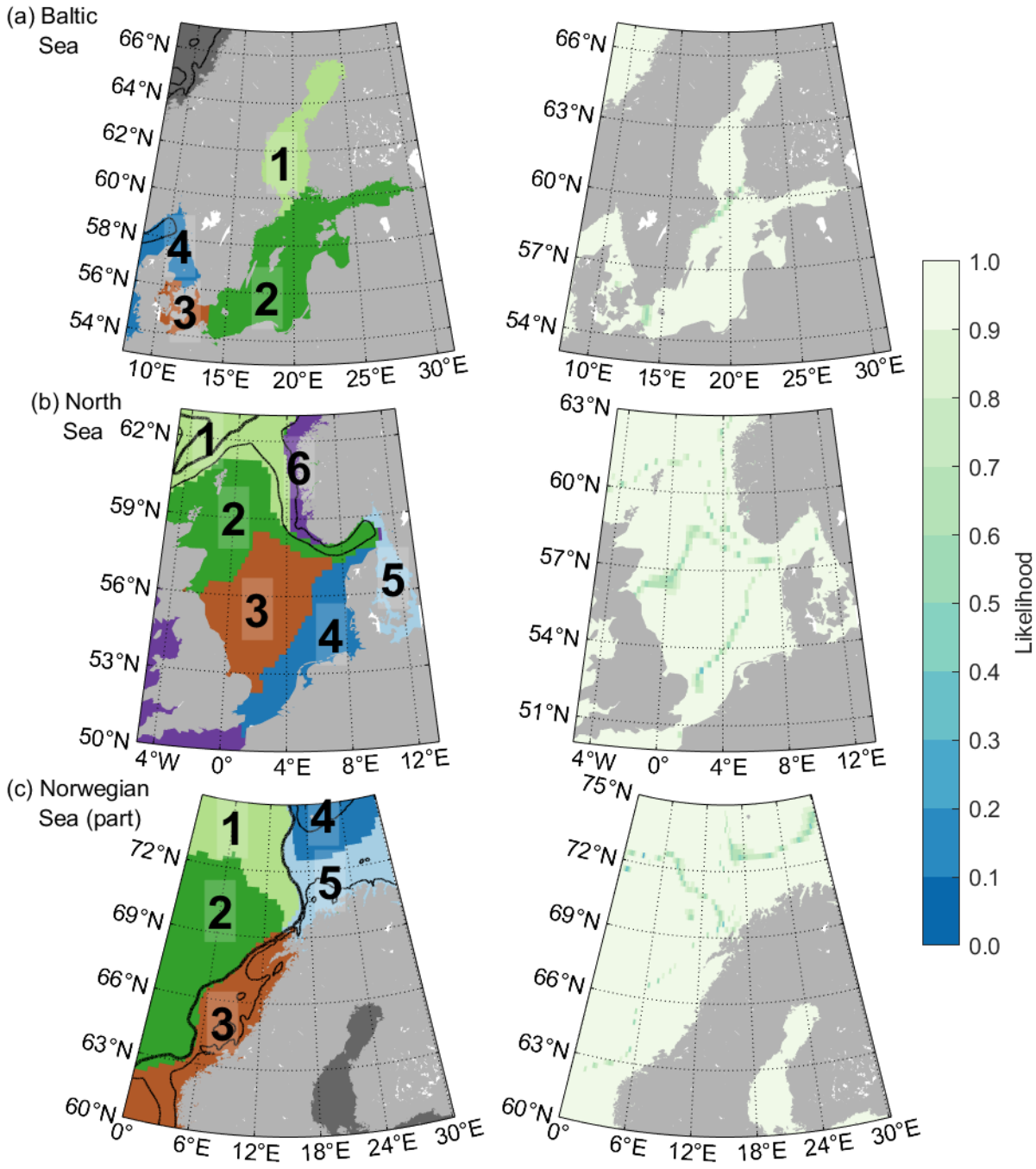


Figure R2: Replacement for manuscript Fig. 4. Classification using an ensemble of 200 Gaussian Mixture Models (left) and the respective likelihoods of the model sorting the grid points to that particular class (right) for the Baltic Sea performed using 4 EOFs (a); North Sea using 5 EOFs (b); and part of the Norwegian Sea using 6 EOFs (c). Numbers indicate the assigned classes. Contour lines represent the 250 and 1000 m isobaths.

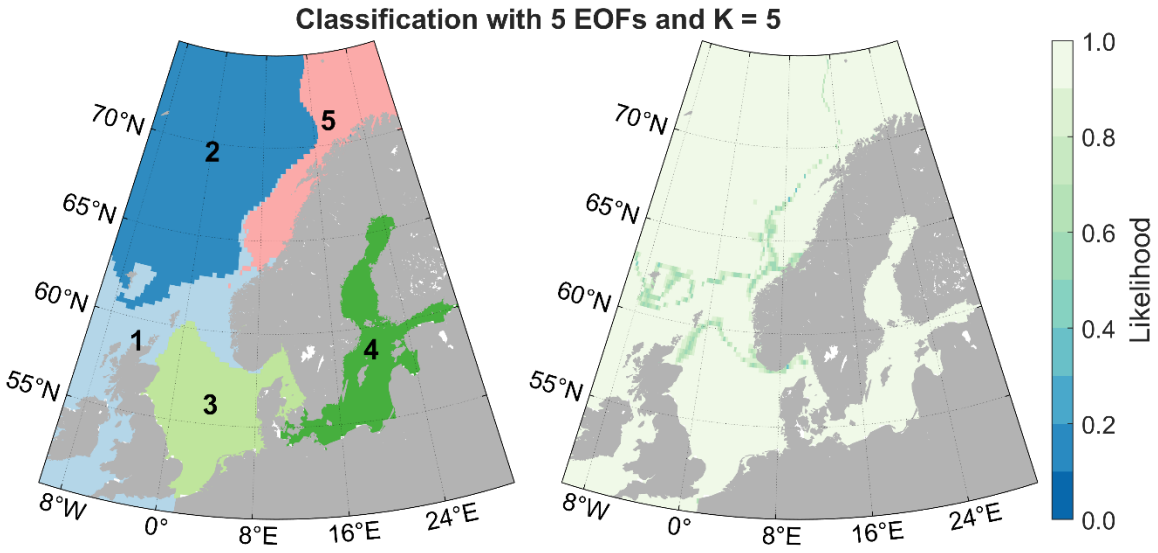


Figure R3: Classification (left) and accompanying likelihood (right) obtained with 5 EOFs and K = 5 with the new version of the data set and the same time period as used in the previously submitted manuscript (1995-2019).

We no longer discuss the former data set and assume that the newer data set is improved and thus classification results obtained with it are more realistic. The differences due to different time spans might have physical relevance and show that perhaps this method can be used to track the changes in large scale sea level patterns, but we do not focus on that in the current manuscript. When comments or our answers are affected by the change in the data set, we point that out in the reply to the specific comment.

Summary of the changes in the classification results

With the new data set, when considering the whole Northwestern European continental shelf (Fig. 3 manuscript and Fig. R1 here) the simplest model remains the same, the intermediate model has one fewer class (class O1 from the original manuscript is not visible in this model), and the complex model remains virtually the same. Class O1 is also found by this model, suggesting that the signals responsible for it are still present in the data, but now only in the higher EOFs, which is why a model using only five of them did not capture it. The area with low likelihood in the southern North Sea visible on Figure 3c in the original manuscript is also gone, suggesting that it was a result of data processing in the old data set. Additionally, with this data set, we obtain almost identical K=10 classification with 8, 9, 10, and 11 EOFs, of which we decided to display the results with 9 EOFs because it is the lowest number of EOFs (meaning the fastest model) for which the model reaches the minimal likelihood of 0.95 averaged over the whole region.

Of the sub-regions (Fig. 4 manuscript and Fig. R2) the Baltic Sea region classification remains the same and the classification of the North Sea has a difference only along the southern coast, most likely for the same reason as the difference in likelihood in the whole area model, but we now need 5 EOFs to achieve that instead of only 3. This unfortunately means that it is no longer possible to directly show this model in the abstract EOF space because it now requires 5 dimensions. Fig. 6 in the manuscript, therefore, now has only the simple model of the whole region and the discussion of the classes in the abstract space is shortened. For the Norwegian Sea we did not manage to obtain the same number of classes as before, the highest K with stable results is 6. The likelihood is also lower than previously.

> Line 84: The authors mentioned that "We also remove the seasonal cycle by subtracting the climatology calculated from the 25 years of data in order to focus on the non-seasonal variability.". And what about the trend? It is maintained? If yes, it could be possible that the existence of different trends in the study regions affect to the classification?

The trend is maintained. The spatial patterns of the sea level trend are included in the first EOF map. So yes, the existence of different trends in the study area definitely affects the classification. We added a sentence to the manuscript to explicitly state that.

> Line 133: For the Silhouette coefficient I recommend to include a reference. Here I suggested one, but it could be another one. Filaire, T., 2018. Clustering on mixed type data, a proposed approach using R. https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd 0a2569c3

Thank you for the suggested reference. Unfortunately, none of the authors have access to this article, but we found another reference we can use instead: Rousseeuw (1987).

> Line 136: What is the mean S? I don't understand very well how the S score was used. On my understanding, to define the number of components (K), normally S is iterated several times starting from K=2 to higher values, and then, the K that gives the best S value is selected. Please, if possible, include here a deep explanation.

Equation (5) in the manuscript is the formula for calculating silhouette score of one data sample (in our case one grid point) i. The mean intra-cluster distance a is the mean distance between grid point i and all other grid points in the same cluster and the mean nearest-cluster distance b is the mean distance between grid point i and all points from the nearest cluster. Then to find the best K we use the average of the silhouette score for all samples. We added this explanation to the manuscript.

> Line 193: This part is not really clear for me. I understand that 11 EOFs represented the 85% of the variability. But why considering 11 EOFs when the Silhouette coefficient for this model present the lowest values of all the run models (Figure 2)?

We only used the silhouette score to find the best number of classes for a specific number of EOFs, not for comparisons between different EOFs. One of the effects of the curse of dimensionality is that as we increase the number of dimensions (EOFs) the distance between any two data points becomes more similar and less meaningful, which also affects the silhouette score. In the end we used the model which allows the separation into the highest number of classes because we wanted to see the classes which would otherwise be not so obvious like the classes obtained with 3 EOFs are.

> Line 223: "The likelihood for the classification in the southern part of the North Sea is also significantly reduced, suggesting that the models struggle to properly classify this region, possibly because this many principal components introduce a lot of noise.". And it could not be related with the value of the S score that is the lowest of all the run models?

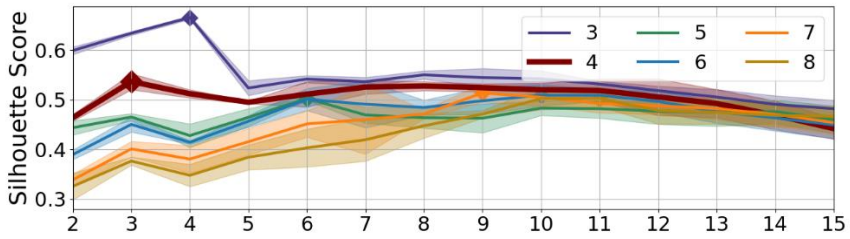
This is a very good question. Originally the answer would have been yes, both the low silhouette score and the low likelihood in that region might be a result of the difficulties of considering distance in a 11-dimensional space, but after changing the data set that area of low likelihood is gone in both 1995-2019 and 1995-2021 results, so it was most likely a result of something related to data processing.

> Figure 4: The manuscript did not include an explanation of why those K were selected. Additionally, the S score is not presented. I recommend to add this information.

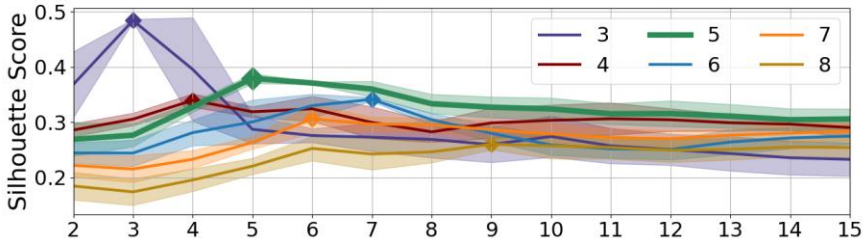
> Line 255: "Note that this number of classes is not chosen with the silhouette score.". If the S score is not used in this region, why the authors chose 4 classes?

We tested all the Ks around those determined by the silhouette score and chose to use and present the results which fulfilled the requirements for the "best" model we listed in Sect. 2.3. We added this explanation into the manuscript, as well as a note on how important it is to test the models because the silhouette score or the Bayesian Information Criterion or other metrics sometimes fail to detect the best K. Since in the end the silhouette score is not used for the number of classes in any of the subregions, we believe a figure with it would not fit well into the article, but you can see it in Fig. R4 here.

a) Baltic Sea



b) North Sea



c) Norwegian Sea

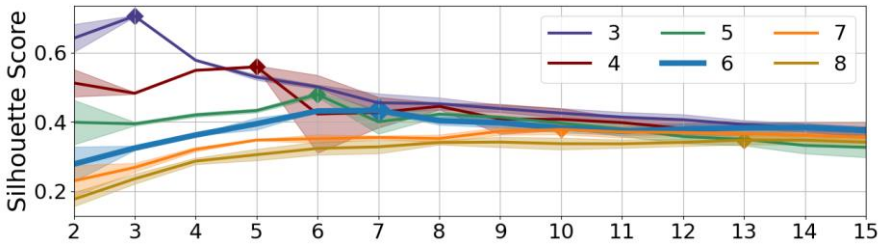


Figure R4: Silhouette score for the three sub-regions: Baltic Sea (a), North Sea (b), and Norwegian Sea (c). Differently colored lines represent the silhouette score computed with a different number of EOFs. The number of EOFs used in the manuscript is marked by a thicker line.

> Figure 5 is somewhat confusing. Can you please add more explanations to be fully understandable?

Yes, we apologize, we realize that the explanation we provided for this figure was lacking and easily misunderstood, so we added a better one into the revised manuscript. We also decided to remove the seventh EOF from the figure because it does not provide any crucial information and excluding it makes the other plots larger and more readable. Apart from assigning classes, GMM also gives the class means and covariance matrices it fits the data to. Since it fits to multivariate Gaussian distributions, for each class it gives a class mean for each EOF used to train it, which is what we show in columns b-d for three ensemble GMMs by replacing the values of EOFs at each grid point with mean values from the class assigned to that grid point. This can show us two things:

1) A comparison of the class mean EOFs with the original EOFs (column a) indicates how well the model fits to the data; and 2) The difference in class means between two classes can tell us which EOFs are responsible for that class border.

References

Rousseeuw, P. J. (1987): Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20, 53-65, https://doi.org/10.1016/0377-0427(87)90125-7.