https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

Unsupervised classification of the Northwestern European seas based on satellite altimetry data

Lea Poropat[1], Dan(i) Jones[2], Simon D.A. Thomas[2,3], and Céline Heuzé[1]

[1]Department of Earth Sciences, University of Gothenburg, Gothenburg, Sweden
[2]British Antarctic Survey, NERC, UKRI, Cambridge, UK
[3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

Correspondence: Lea Poropat (lea.poropat@gu.se)

Abstract. From generating metrics representative of a wide region to saving costs by reducing the density of an observational network, the reasons to split the ocean into distinct regions are many. Traditionally, this has been done somewhat arbitrarily, using the bathymetry and potentially some artificial latitude/longitude boundaries. We use an ensemble of Gaussian Mixture Models (GMM, unsupervised classification) to separate the complex northwestern European coastal region into classes based on sea level variability observed by satellite altimetry. To reduce the dimensionality of the data, we perform a principal component analysis on 25 years of observations and use the spatial components as input for the GMM. The number of classes or mixture components is determined by locating the maximum of the silhouette score and by testing several models. We use an ensemble approach to increase the robustness of the classification and to allow the separation into more regions than a single GMM can achieve. We also vary the number of empirical orthogonal function maps (EOFs) and show that more EOFs result in a more detailed classification. With three EOFs, the area is classified into four distinct regions delimited mainly by bathymetry. Adding more EOFs results in further subdivisions that resemble oceanic fronts. To achieve a more detailed separation, we use a model focused on smaller regions, specifically the Baltic Sea, North Sea, and the Norwegian Sea.

1 Introduction

Sea level variability in coastal regions is a critical area of research due to its implications for coastal management, climate change assessments, and hazard mitigation (Fox-Kemper et al., 2021). Sea level also reflects ocean currents (Dangendorf et al., 2021), so understanding the patterns and classifying the ocean based on sea level data can provide valuable insights into the dynamic behavior of these regions. Furthermore, while satellite-based instruments provide us with observations covering large areas, they only exist for the last three decades at most (Ablain et al., 2015), so in order to study interannual and decadal processes, we have to rely on tide gauges, which are only available at specific point locations. Knowing the regions of coherent sea level variability allows us to estimate how broad of an area our conclusions based on tide gauges can be applied to.

The traditional ways for studying coherence in sea level is by calculating correlation, a principal component analysis (PCA) or a combination of the two. For example, Papadopoulos and Tsimplis (2006) extracted empirical orthogonal functions (EOFs) to create regional indices that represent sea levels in large areas of the world oceans and calculated the correlations with climate indices to study the teleconnection patterns. Bulczak et al. (2015) used EOFs to decompose the observed seasonal

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

sea level variability in the Nordic Seas and compare it with the steric and dynamic forcing. Iglesias et al. (2017) performed a correlation analysis between the altimetry-observed sea level anomaly in the North Atlantic and the teleconnection patterns. As most studies do, they all separated the ocean into regions only based on geographical locations, i.e., the coastlines and ocean basins, but did not attempt to further separate the basins based on the differences in observed sea level variability. There have also been a few attempts to apply more complex classification or clustering methods to sea level, e.g. Scotto et al. (2010) used agglomerative hierarchical methods to group time series in the North Atlantic Ocean based on their posterior predictive distributions for extreme values, while Barbosa et al. (2016) used wavelet-based clustering to find regions with similar sea level records in the Baltic Sea. Self-organizing maps (SOMs; Kohonen and Mäkisara, 1989; Kohonen, 1990) are a type of unsupervised neural network often used for clustering and pattern analysis in atmosphere and ocean research. They have, among many other things, been successfully applied to find the patterns of upper layer ocean circulation from altimeter observations on the West Florida Shelf (Liu and Weisberg, 2005) and in the South China Sea (Liu et al., 2008), as well as from radar data in the Northern Adriatic (Mihanović et al., 2011). However, SOMs are primarily a feature detection tool, which is also able to perform classification. Since SOMs are based on a neural network, it is harder to interpret the results.

Therefore, in this work we use another unsupervised classification method called the Gaussian Mixture Model (GMM; Bilmes, 1998) to determine the regions of coherent sea level variability. This method has already been used in oceanography to classify the ocean based on temperature and salinity profiles. Maze et al. (2017) applied it to temperature profiles in the North Atlantic to find the regions with similar vertical thermal structure, and Jones et al. (2019) did a similar study of the Southern Ocean. Rosso et al. (2020) focused only on a part of the Southern Ocean, the Kerguelen Sector, but included both the temperature and the salinity observations into the model. Thomas et al. (2021) then did a similar study of the whole Southern Ocean. They all used PCA to reduce the number of levels in the vertical, which reduces the computational cost of the classification. Here we apply the same method on satellite observed sea level, using PCA to reduce the amount of information in the temporal domain. GMM provides similar output as the self-organizing map, i.e., the classification of the area and the main pattern for each class, but since it is based on statistical distributions, it is easier to interpret the results. Because GMM gives a class for every data point, the results from it not only provide an insight into the patterns of sea level variability, but can also be used as a mask to isolate a region and focus on the dominant processes in it without being affected by the noise from everything in the neighboring areas. GMM is also probabilistic, i.e., it provides the probability distribution across all classes for each data point, which can be helpful when trying to determine the robustness of the classification, giving it an advantage over simpler approaches, such as K-means clustering (Lloyd, 1957).

In Sect. 2, we describe the methods and data used in this paper. We start with the description of the used data set and applied data processing steps (Sect. 2.1), then we continue to explain how the Gaussian Mixture Model works (Sect. 2.2), and finally detail the ensemble classification procedure (Sect. 2.3). Sect. 3 contains the results and the discussion, focusing on the classification and its dependence on the amount of information contained in the data set (Sect. 3.1), showing the results for specific subregions of our area of interest (Sect. 3.2), and then illustrating how the classification works in the abstract empirical orthogonal function domain in which it is performed (Sect. 3.3). Finally, we present our conclusions in Sect. 4.

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

2 Method

To determine the regions of coherent sea level variability we use a machine learning method called Gaussian Mixture Model (GMM). It is an unsupervised classification (or clustering) model, i.e., a model that seeks to sort data points into classes due to their similarity without any *a priori* information about the classes. We apply it to the satellite-observed sea level data in the Northwestern European coastal seas and part of the Atlantic Ocean. To increase the robustness of the classification, we use an ensemble of GMMs.

2.1 Data preparation

For the sea level variability information we use gridded reprocessed global ocean sea surface height satellite observations downloaded from Copernicus Marine Services (Pujol and Mertz, 2020). This data set uses a multi-mission mapping procedure based on an optimal interpolation technique derived from Le Traon and Ogor (1998), Ducet et al. (2000), and Traon et al. (2003), which combines the data from all available satellite missions: Sentinel-3A/B, Jason-3, HY-2A, Saral[-DP]/AltiKa, Cryosat-2, OSTM/Jason-2, Jason-1, Topex/Poseidon, Envisat, GFO, ERS-1/2 (Taburet et al., 2019). The data set has a global coverage, with 0.25° spatial and monthly temporal resolutions. The full description of the processing of the altimetry data and all the corrections applied to them can be found in Pujol et al. (2016) and Taburet et al. (2019).

As an example for our method, we select the area between 10°W and 30°E, and 50°N and 75°N (Fig. 1), which covers the coastal seas of Northwestern Europe and a part of the Northern Atlantic Ocean. It is an interesting and complex region that comprises of many different ocean floor features and includes both mid-latitudes and polar region, as well as continental shelves and deep ocean regions. It consists of the very shallow enclosed Baltic Sea, the shallow North Sea and coastal seas between Great Britain and Ireland, the Faroe Shelf, the Norwegian continental shelf and part of the Barents Sea, all shallower than 1000 m. At the other extreme in depth, this region of interest also includes the Norwegian Sea, part of the Greenland Sea, and a section of the mid-Atlantic ridge in between, all deeper than 2000 m.

While the data set covers a longer time span, we decided to use 25 years of data, from 1995 - 2019, to avoid the large areas with missing data in year 1994. In most winters there are gaps in the data in the Gulf of Bothnia and Gulf of Finland in the Baltic Sea, due to the extensive sea ice cover, which prevents sea level retrieval through altimetry (Pujol et al., 2016; Taburet et al., 2019). While these gaps might affect the results of the classification, we want to keep the whole area, so we linearly interpolate in time the gaps as the first step of data processing. 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.

While it is technically possible to use the time series directly as input for the mixture model, as it is with SOMs (Liu and Weisberg, 2005), they contain so much noise that the model is unable to converge to one best distribution of classes. It also makes the model one order of magnitude slower, from approximately 1 to 13 seconds for a single GMM, which makes testing large ensembles much more time consuming. Therefore, before applying the unsupervised classification method, we perform a principal component analysis (PCA) on the altimetry data set to reduce its dimensionality. We obtain the empirical orthogonal function (EOF) maps, which contain the spatial component of the dataset, and the accompanying principal component time

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

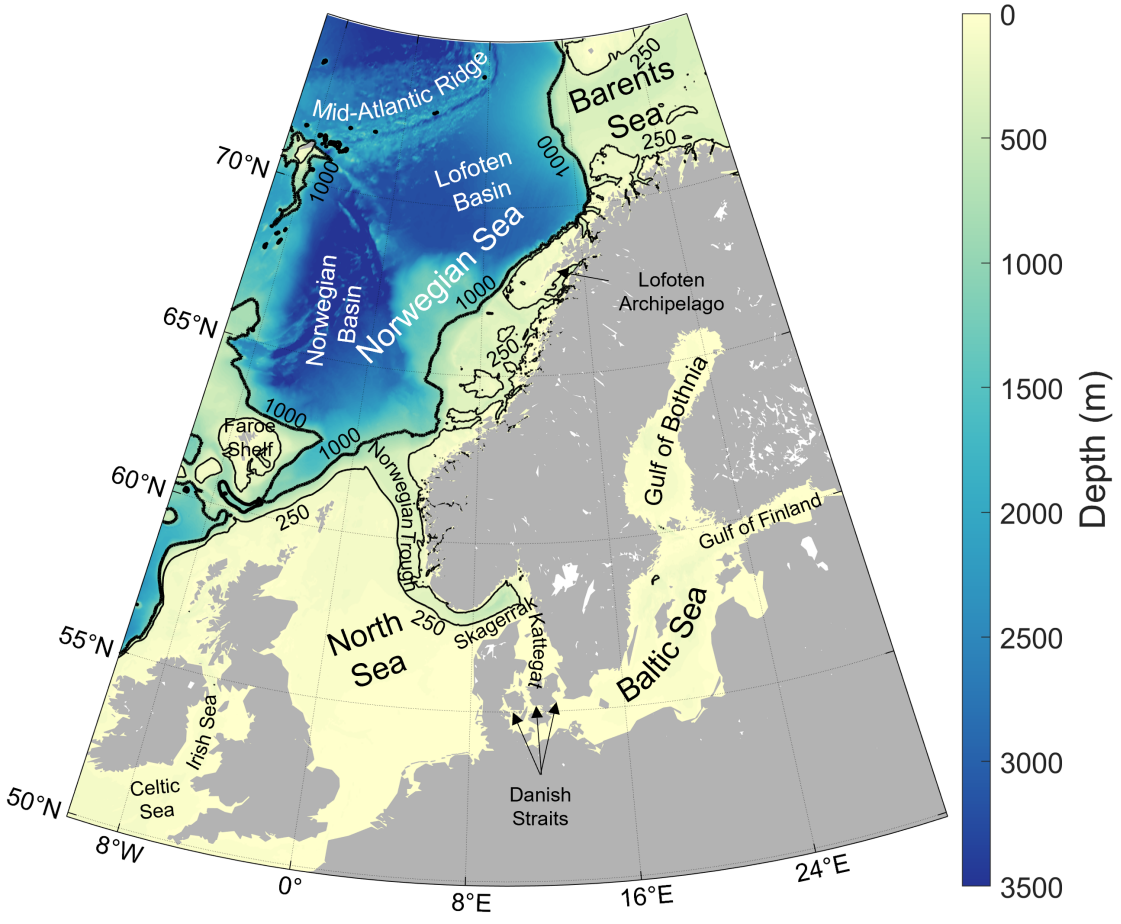

Figure 1. Our region of interest and its bathymetry, from the General Bathymetric Chart of the Ocean GEBCO 2022, along with place names referred to in the manuscript. Black contours represent the 250 (thin line) and 1000 m (thick) isobaths.

series, as described in Björnsson and Venegas (1997), and use the EOF maps as input for the machine learning classification model. In this way the input data is reduced from 300 monthly grids to only 3-11 EOFs, which explain 75-85% of observed variability, and each grid point that we wish to classify is explained with 3-11 values instead of by the whole 25 years long time series. The decision on how many EOFs are included is based on whether we are trying to achieve a simpler classification, in which case less information is enough, or study the finer details, which requires higher degree EOFs.

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

2.2 Gaussian mixture model

We want to objectively identify patterns appearing in the EOF maps of the satellite-observed sea level and use them to define regions of similar sea level variability. A powerful method for this task is the Gaussian Mixture Model, an unsupervised machine learning classification approach that provides the probability that a location belongs to each of the classes. Since it is an unsupervised method, it does not need any prior information about the classes. It is based solely on sea level variability, without any geographical information, which allows us to determine physically coherent regions even if they are not adjacent. Finally, since this method provides the probability distribution across all of the classes, it enables us to distinguish clearly-coherent regions from boundaries.

GMM is based on the assumption that any probability density function (PDF) can be described with a model of weighted sums of Gaussian PDFs, which represent the components of the mixture model. In our case, the PDF describing the sea level EOFs can be represented with a weighted sum of Gaussian PDFs:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \lambda_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}), \tag{1}$$

with K components, where

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma_k}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_k})^\top \boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_k})\right) \tag{2}$$

is the multivariate Gaussian distribution in D dimensions with mean $\boldsymbol{\mu_k}$ and covariance matrix $\boldsymbol{\Sigma_k}$. The weighting coefficients λ_k must satisfy $0 \leq \lambda_k \leq 1$ and $\sum_k \lambda_k = 1$ (Bishop, 2006). The classification occurs in the abstract, D-dimensional EOF space, where each dimension represents one of the EOFs and each data point $\boldsymbol{x}$ is a grid cell described by D EOF values. The aim of the GMM is to fit the PDF model from Eq. (1) to the observed probability density function of the EOFs by maximizing the likelihood of the observations using the Expectation-Maximization method (Dempster et al., 1977; Bishop, 2006), which is referred as model training. This boils down to finding the best estimates for the parameters λ_k, $\boldsymbol{\mu_k}$, and $\boldsymbol{\Sigma_k}$. After the observed PDF has been decomposed into a sum of K Gaussian mixture model densities defined by mean $\boldsymbol{\mu_k}$ and covariance matrix $\boldsymbol{\Sigma_k}$ with $\lambda_k = p(\mathbf{c} = k)$ being the component *a priori* density for class $\mathbf{c}$, we then use Bayes' theorem:

$$p(\mathbf{c} = k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathbf{c} = k)p(\mathbf{c} = k)}{p(\boldsymbol{x})}, \tag{3}$$

where $p(\boldsymbol{x}|\mathbf{c} = k)$ is defined by Eq. (2) and $p(\boldsymbol{x})$ is given in Eq. (1), to obtain the *a posteriori* probability of a location belonging to class $\mathbf{c}$:

$$p(\mathbf{c} = k|\boldsymbol{x}) = \frac{\lambda_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}{\sum_{k=1}^{K} \lambda_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}. \tag{4}$$

Finally, the location is labeled with class k for which the posterior probability is the largest. The mean values which define each class in our case give us information about which principal components and processes associated with them are dominant in that region. A detailed description of Gaussian Mixture models can be found in e.g., Bilmes (1998) or Bishop (2006), while

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

e.g., Maze et al. (2017) or Thomas et al. (2021) provide similar, oceanography oriented explanations. The computation was performed using the open-source Python library Scikit-learn (Pedregosa et al., 2011).

In addition to deciding on the amount of information included in the data by selecting the number of EOFs D, the only input parameter to the GMM is the number of mixture components or classes, K, which needs to be specified before applying the model. It is not an easy task to determine the appropriate K. Sometimes it is possible to select a number by relying on theory behind the processes we are studying, e.g., when using GMM to find the fronts in the Antarctic Circumpolar Current like in Thomas et al. (2021), the number of fronts is known, but that is not always possible. There are multiple methods to objectively determine the optimal K, of which we used the silhouette score. Silhouette score for each sample is computed as:

$$S_i = \frac{b - a}{max(a, b)}, \tag{5}$$

where a is the mean intra-cluster distance and b is the mean nearest-cluster distance. To determine the best number of classes, we use the mean S. S ranges between -1 and 1, where higher values correspond to better distinguished classes. Other studies, such as Maze et al. (2017) or Thomas et al. (2021), used the Bayesian Information Criterion (BIC) for this purpose. In our case however, the K selected based on the BIC was usually too large (see next subsection), while the silhouette score provided a better estimate. Note that, while the silhouette score provides a good estimate, it is a metric that does not always work perfectly, so it is best to test the model with multiple options until we find the optimal number of classes.

2.3 Ensemble classification

The initial class means in the GMM algorithm are determined by the simpler k-means clustering method, which depends on random initialization. To test whether the model converges, we do not specify random seed, so the initial parameters are different every time. Due to the size and complexity of the area and the sea level variability, each time the model is trained the results can be slightly different. To mitigate that and increase the robustness of the results, we use an ensemble prediction. GMM provides not only the classification, but the probability for a point to belong to each of the classes, so the most fitting way to do the ensemble classification is to use soft voting. With this method, the ensemble takes into account the probabilities from each model that a point belongs to each class, the class with the largest sum of probabilities wins and the grid point is finally assigned to that class. We also obtain the likelihood that a grid point will belong to that class, i.e., a combination of the number of models that sorted it into that class and the probability they provided, which tells us how difficult it was for the model to sort that particular location.

For most unsupervised classification models, including GMM, the main problem with using an ensemble is that, since the classes are not known *a priori*, they are not numbered in any particular way, so class 1 of one ensemble member can correspond to class 7 of another. Since there are some differences between model runs, it is also possible that a class appears in some model runs but not others. To be able to compare the classes from all ensemble members, we match them based on the correlation between the class means (a *D*-dimensional vector), which results in a list of classes that appeared at least once in any of the models of the ensemble. This list of classes is substantially longer than the predetermined number of components *K*, but after voting many of the classes get voted out because they only appear in a few of the models.

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

Another problem that appears when using an ensemble is that, while GMMs themselves usually do not result in very small classes, after soft voting, some classes can lose most of their points to neighboring classes, ending up with only a few data points. That is avoided by setting up a minimal class size threshold, excluding those classes that have a number of points below the threshold, and re-sorting all the grid points belonging to the excluded classes to the class with next highest vote. In such cases, there are usually two classes with very similar probability sums, so the resulting likelihood is not considerably reduced.

Therefore, there are three parameters pertaining to the ensemble: the minimal correlation for the class means to be considered "same", the minimal class size, and the number of ensemble members N. The other parameters that we need to set are the GMM's only intrinsic parameter; the number of classes K; and the amount of information included into the model set by the number of EOFs D. To determine which combination works best, we use three criteria: (1) the model converges, i.e., multiple experiments with the same parameters find the same classes; (2) the ensemble keeps the same number of classes prescribed to the individual GMM; and (3) the average likelihood inside the classes is as high as possible for the desired level of subdivision, while still allowing low likelihoods on class borders. After testing several options, we find that the ensemble works best if we use a minimal correlation for matching classes of 0.98. With smaller correlation, classes that are not similar enough could be merged, while with larger, even a slight difference in geographical distribution of a class in different ensemble members results in the ensemble seeing them as different classes, neither of which receives enough votes, so only the next best class wins. The minimal class size should be chosen based on the smallest area we are trying to capture, so we select a size of 100 grid points, which allows the model to sort the small basins such as the Gulf of Bothnia or channels such as the Kattegat into a separate class if necessary. We use 200 ensemble members for all our experiments. In some cases with small number of classes, we would have achieved the same results with less, but since training a single GMM is fast, using an ensemble with 200 members does not take too much time and it increases the robustness of the results. Using more ensemble members does not improve the results.

In the end, we obtain an ensemble classification with a new class number K_E that is usually similar to, but not necessarily exactly the same as, the *a priori* class number K, along with a likelihood that a particular point belongs to the selected class. The likelihood would correspond to 1 if all ensemble members chose that particular class with the probability of 1. Both fewer models assigning that class and models assigning it with a lower probability, i.e., the models not being certain that the grid point belongs there, reduce the likelihood.

3 Results

3.1 Classification depending on the number of empirical orthogonal functions

After greatly reducing the dimensionality of the data with principal component analysis, the question arises as to how much data we should keep. As can be seen in Fig. 2, the optimal number of classes based on the silhouette score grows as we add more information, i.e., more EOF maps (colors), to the mixture. We can therefore decide on both (1) the number of EOFs and (2) the number of classes based on how many details we need to retain for our application. Note that according to the silhouette

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

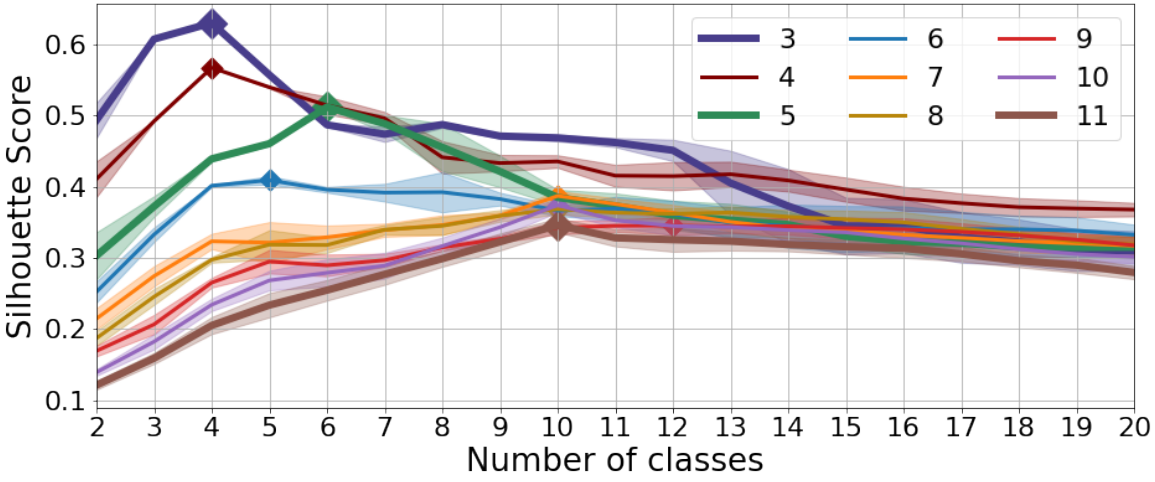

Figure 2. Silhouette Score for different numbers of classes calculated for Gaussian Mixture Models using different numbers of empirical orthogonal functions (differently colored lines). The class for which a respective model has the highest silhouette score is marked with a diamond. Silhouette score is computed for 100 models using the same parameters and the figure represents their average values (central lines) and one standard deviation (shaded areas). The models are fitted to randomly selected 90% of the grid points from the region shown in Fig. 1. Thicker lines and larger markers represent the models presented in this paper.

score, sometimes adding another principal component does not increase the number of classes the model is able to support, but it could still change which classes the model decides to include with this new added information.

Fig. 3 shows the classification obtained using 3 (simplest model), 5 (intermediate), and 11 (complex) EOF maps, which contain 75%, 80%, and 85% of the observed variability, with an ensemble of 200 GMMs. All classifications are created using the number of classes recommended by the silhouette score: 4, 6, and 10 respectively, which were indeed the numbers with which the model works best. Since the ensemble classification is able to modify the number of classes if the chosen one does not work well by discarding the classes that are only rarely selected by individual models, the fact that the ensemble maintains the selected number of classes is an additional proof that the number was good. The black dots on all figures mark the location of grid points that were excluded from the training dataset, i.e. the grid points that were held out for testing. It can be seen that even though the model does not consider the geographical information at all, it properly sorted almost all the grid points, both from the training set and the test set, into suitable geographically connected areas.

In the simplest model based on only three EOF maps (Fig. 3a), the GMM splits the area into only four classes: Baltic Sea (B), North Sea (N), which includes most of the transition area towards the Baltic called Skagerrak and Kattegat, the remaining continental shelf areas, including the northernmost part of the North Sea and the shallow Barents Sea (C), and the deep open ocean (O). The border between the latter two classes follows almost perfectly the continental shelf border, which can be seen from the 1 000 m isobath. The border between the Baltic and the North Sea classes is also related to the geographical properties

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

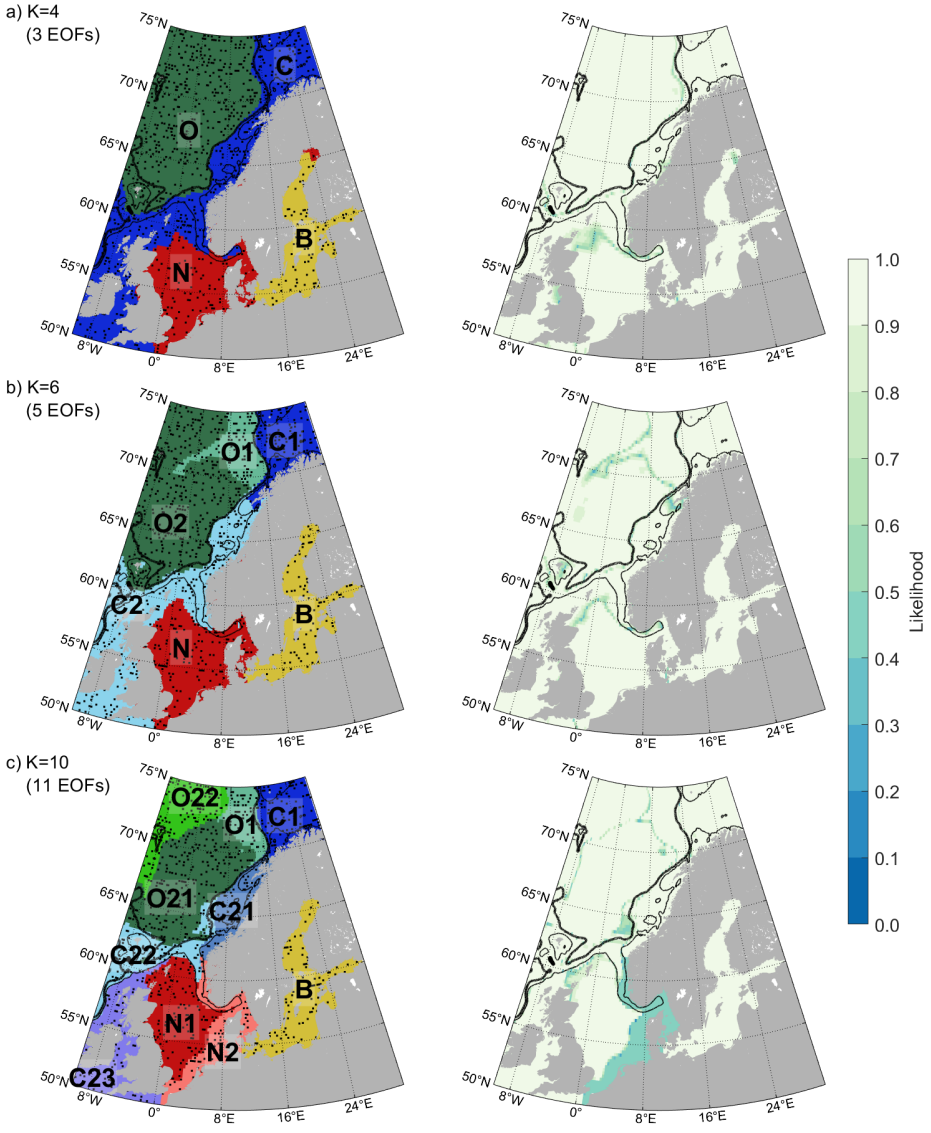

Figure 3. Classification using an ensemble of 200 Gaussian Mixture Models (left) and the respective likelihoods of the model sorting the grid points to that particular class (right). Classification is performed using 3 (a), 5 (b), and 11 (c) empirical orthogonal functions and 4, 6, and 10 classes, respectively. 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. Black dots represent the data points used for validation.

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

and is set at the narrowest region connecting them, the Danish Straits. The only border that does not seem to be directly caused by bathymetry is between the majority of the North Sea (N) and the remaining coast (C). That border is also the hardest one to classify, which can be seen from the likelihood (as low as 0.26); it is the only area where a significant number of models created a slightly different border between the classes. Interestingly, the model classifies the Bothnian Bay, the northernmost part of the Baltic Sea, together with the North Sea. This requires further investigation of the class means of the EOFs to determine whether this has a physical background or is an artefact possibly related to the interpolation of the missing values in the Bothnian Bay, which will be discussed in Sect. 3.3. There are also a few points along the Norwegian coast that get assigned to the North Sea class, suggesting that perhaps this class contains a particular mode of variability in the shallow coastal areas.

Adding two more EOFs that together contribute 5% of variability information (intermediate model, Fig. 3b) does not change the classification significantly. The class borders from the simplest classification remain principally the same, and the new EOFs allow further subdivision of the coastal and open ocean classes into two classes each. The split into C1 and C2 is based on the bathymetry difference just south of the Lofoten Archipelago in Norway, while the split into O1 and O2 is based on the mid-Atlantic ridge. Adding these two extra EOFs resulted in removing the Bothnian Bay and the separate coastal areas from the North Sea class because the model now has enough information to clearly distinguish them.

Finally, when using 11 EOF maps (the most complex model, Fig. 3c), we end up with 10 classes in our region. The class borders due to bathymetry remain the same. There is further subdivision of both coastal and open ocean areas, and the border of the North Sea -itself subdivided, is shifted northward. 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. Models maintaining the basic classification and further subdividing some of the classes after adding new information into the model is not a characteristic of the GMM. The GMM could completely change some or all the classes if the number of classes is different, so the fact that this is not happening here must be based on the characteristics of the ocean. The ocean is first coarsely divided into regions determined by the bathymetry, and then each of those regions can be further subdivided based on other aspects of the sea level variability such as finer resolution bathymetric features (e.g. the subdivision of class O around the mid-Atlantic ridge) or water masses (e.g. the Barents Sea separating from class C early on).

Adding even more EOFs (not shown) does not result in a finer subdivision; it only causes either the ensemble to reduce the number of classes to 10 or even introduces so much noise that multiple ensembles with the same parameters produce different results. With that in mind, if we would like to obtain a more detailed subdivision, it is better to do the classification for a smaller region. The complexity of sea level patterns also differs significantly from region to region, so narrowing our focus to a smaller area would allow us to use the principal components specific to that area, increasing the amount of information in fewer EOF maps, thus reducing the noise and allowing a better classification.

3.2 Reducing the size of the region

Here, we apply the ensemble GMM to three subregions of our area of interest: the Baltic, the North and the coastal part of the Norwegian Sea (Fig. 4). By using EOF maps calculated solely for these regions, the model input contains only the data

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

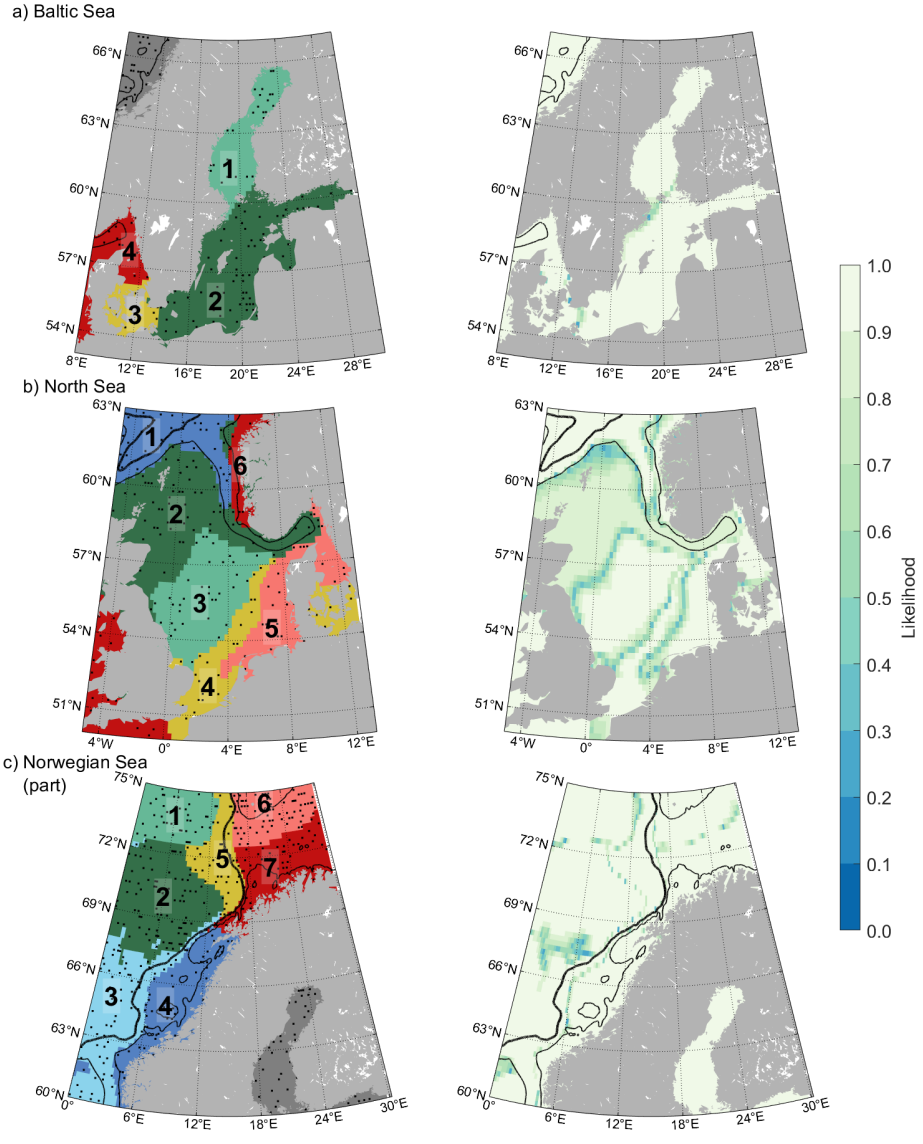

Figure 4. Classification using an ensemble of 200 Gaussian Mixture Models (left) and the respective likelihoods of the model sorting the grid points to that particular class (right) for the Baltic Sea performed using 4 EOFs (a); North Sea using 3 EOFs (b); and part of the Norwegian Sea using 7 EOFs (c). Numbers indicate the assigned classes. Contour lines represent the 250 and 1000 m isobaths. Black dots represent the data points used for validation.

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

relevant for them, without the noise coming from EOFs significant only elsewhere, which allows the models to find more region-specific patterns and increase the number of classes they are able to find. By comparing the results from the subregions with results from Sect. 3.1, as well as by comparing the classification in the overlapping areas of the three subregions, we can also check whether the ensemble GMM classification is robust and finds the same patters of sea level variability regardless of the size or shape of the area of interest.

We can see that the EOF maps computed for the whole northwestern European coastal area (Fig. 5a) are rather smooth in the Baltic Sea compared to the rest of the region, which is why the models based on them do not usually divide the Baltic Sea further than the basin-scale. This uniformity is also reflected in the EOFs calculated for the Baltic Sea separately: the first four explain 93.7% of the variability. The ensemble classification model based on them is able to distinguish two more classes (Fig. 4a) than in the most complex model based of EOF maps for the whole area of interest (Fig. 3c). The Baltic (class B in Fig. 3c) is now split into three classes: the Gulf of Bothnia (class 1), the Western Baltic (class 3) and the remainder of the Baltic Sea (class 2). The Danish Straits, a series of narrow channels connecting the Baltic Sea with Kattegat, are now sorted together with the Western Baltic class (class 3), while Kattegat and Skagerrak form one class connected to the North Sea (class 4). The likelihood is very close to 1 virtually everywhere except at the border between classes 1 and 2 and between 2 and 3, meaning that the majority of the ensemble members selected the same classes. Note that this number of classes is not chosen with the silhouette score. Although using the recommended number of classes gives reasonably good results (not shown), using this different K works better, i.e., has higher likelihood, especially in the transition zone. This demonstrates that while the silhouette score is a good tool to give an estimate of the number of mixture components, it does not always give the best result. One should always try the model with several options to find the best solution to their specific classification problem.

The North Sea has the most complex sea level variability patterns of the whole considered area. We would need more than 20 EOFs to achieve the same level of explained variance as we achieved with only four for the Baltic. It is however enough to use only three of them, explaining 75% of the sea level variability, to split the North Sea into four distinct classes (Fig. 4b), compared to only two in Fig. 3c. Class N1 from Fig. 3c is here further split into classes 2 and 3, and N2 into classes 4 and 5. The region-specific classification model also finds class 1, which mostly corresponds to class C22 from Fig. 3c, and it combines parts of classes C21 and C23 included here into class 6. The larger classes are separated as zones in the north-south direction, as expected from other works, e.g. Dangendorf et al. (2014) or Sterlini et al. (2016), who found a difference in the sea level variability between the northern and the southern North Sea. Some of the class borders (1 and 2) are also based on bathymetry, following the Norwegian Trough. Interestingly, part of the Norwegian coast included here is sorted into the same class as the western coast of Great Britain, which suggests that the model most likely sees some processes relevant for western coasts and the regional atmospheric pressure and wind patterns. The likelihood is lower than in the Baltic Sea model, but nevertheless close to one across the area, with the exception of class borders, which the ensemble members do not agree on so well. The classes in the area overlapping with the Baltic Sea model match, demonstrating that in both cases the models find the same patterns of variability despite being based on different EOFs.

Finally, the classification obtained by considering 7 EOF maps calculated for the Norwegian coast (Fig. 4c) is quite similar to that obtained by using the whole region (Fig. 3c). Classes 1, 2, 3, 4, and 5 from Fig. 4c generally correspond to classes

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

O22, O21, C22, C21, and O1 from Fig. 3c, although there are some differences in class borders, particularly in the northern boundary of classes 3/C22 and 2/O21. Because of the shift of the northern boundary of class 3, this is the only class containing both deep ocean and continental shelf areas. Additionally, the region-specific model splits the Barents Sea opening based on its depth (classes 6 and 7), which the large model is not able to do.

3.3 Empirical orthogonal functions

To learn more about how the GMM determines the classes, we can take a look at the empirical orthogonal functions (EOFs) because GMMs perform the classification based on them. We compare the EOF maps (Fig. 5a) with the class means obtained by the three models for the whole area of interest (Fig. 5b, c, and d). The border between the Baltic (class B) and the North Sea (class N) is visible on most EOF maps (column a), except the second one. The border based on the continental shelf break is most visible in EOF 2 and 3. The only border in the simplest model (column d) that is not based on bathymetry, i.e. the border between the North Sea and the rest of the continental shelf, just south of 60°N, is determined by the gradient in the first EOF map in that location.

When we look at the class means from the more complex models (Fig. 5, columns b and c), they start to resemble the original EOF maps more closely. Both the coastal area outside of the North and the Baltic Seas and the open ocean are rather uniform in the first three EOFs, which is why the simplest model is unable to divide them further. However, adding the fourth EOF map introduces a clear border in the continental shelf between the Barents Sea (C1) and the rest of the coastal shelf (C2), while the fifth EOF adds the class in the northern part of the open ocean (O1). Most EOFs have a gradient in the Bothnian Bay, which has a seasonal sea ice cover. As detailed in Taburet et al. (2019), whether to mask a pixel is decided based on sea ice concentration. As they aim to eliminate as few pixels as possible, only high concentrations are masked (value not explicitly provided in their paper). Therefore, interestingly, it is possible that the EOFs are indirectly detecting the sea ice concentration variability. As this gradient results in a magnitude similar to that of the North Sea for EOF2, and EOF3 to some extent, this is probably the reason why the simplest model sorted that area with the North Sea. Since other EOFs do not have much similarity between the Bothnian Bay and the North Sea, adding them results in the model correctly sorting that area into the Baltic Sea class. Additionally, the large positive values along the southern Norwegian coast in the EOF 5 are most likely responsible for separating these coastal points from the North Sea class, where the simplest model possibly incorrectly sorted them, and including them in the C2 class.

Finally, not all EOF maps included in the most complex model are shown, but we can see that the sixth one is responsible for the British-Irish class (C23) created in this model, while the seventh one for the creation of the southern North Sea class (N2). But we can also see that for the higher order EOFs, the class means become smoother, indicating that the model learns less and less from each new EOF added to it, until it reaches the point when adding new EOFs introduces only noise and prevents the model from finding reasonable classes.

To achieve useful results, we need to find a balance between interpretability and accuracy. Simpler models with fewer EOFs tend to be easier to interpret, in that they have clearer boundaries between regions, but they fail to capture the full variability of the data. More complex models capture more of the variability of the data, but they tend to be harder to interpret, in that they

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

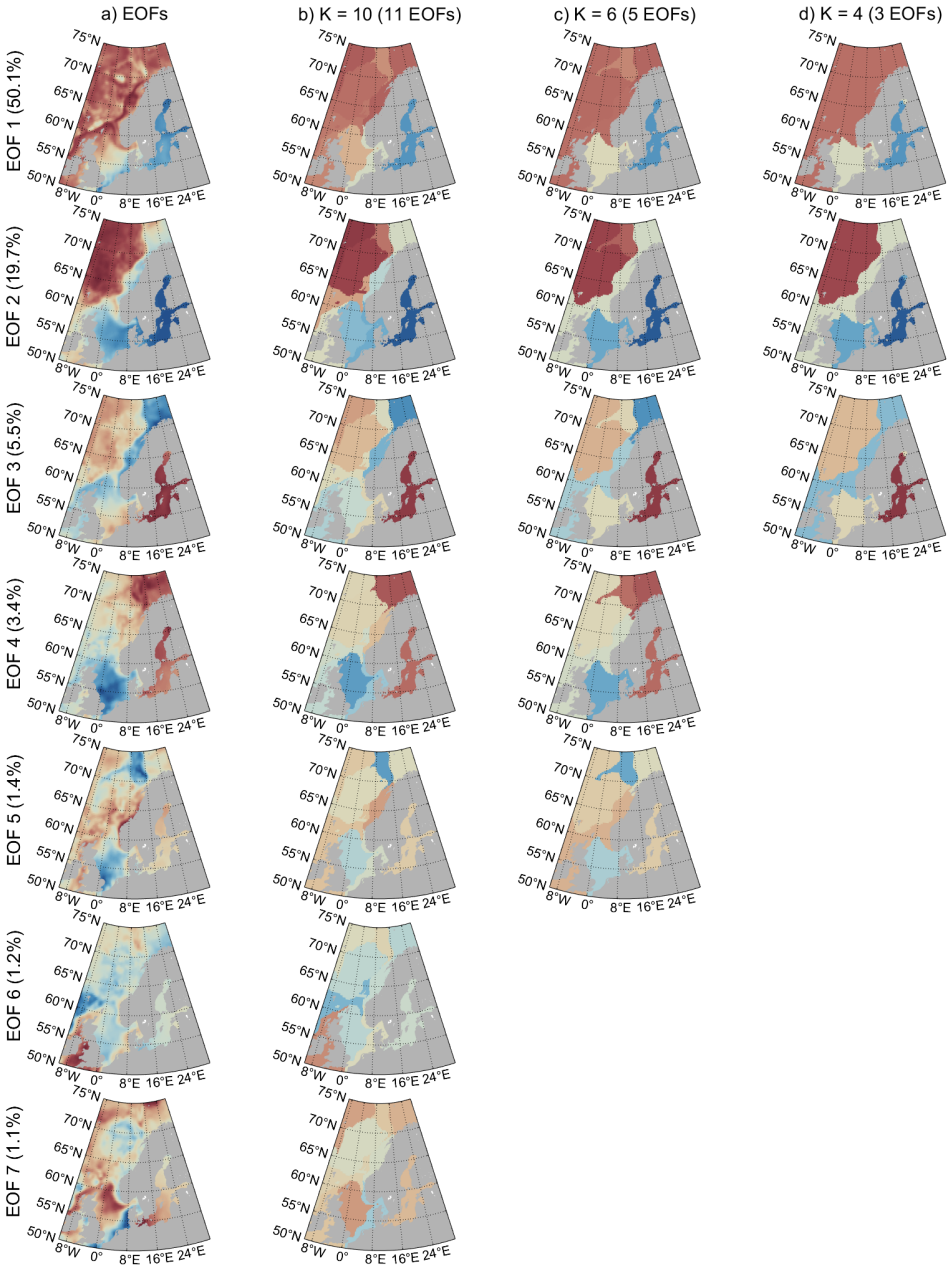

Figure 5. Empirical orthogonal function maps of satellite-observed sea level used as input for the Gaussian Mixture Model (a) and class means for the classification into 10 (b), 6 (c), and 4 (d) classes performed using 11, 5 and 3 empirical orthogonal functions. Column (b) contains only the first 7 EOFs, while columns (c) and (d) show the results for all EOFs used in the classification. The color scale is the same in each row.

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

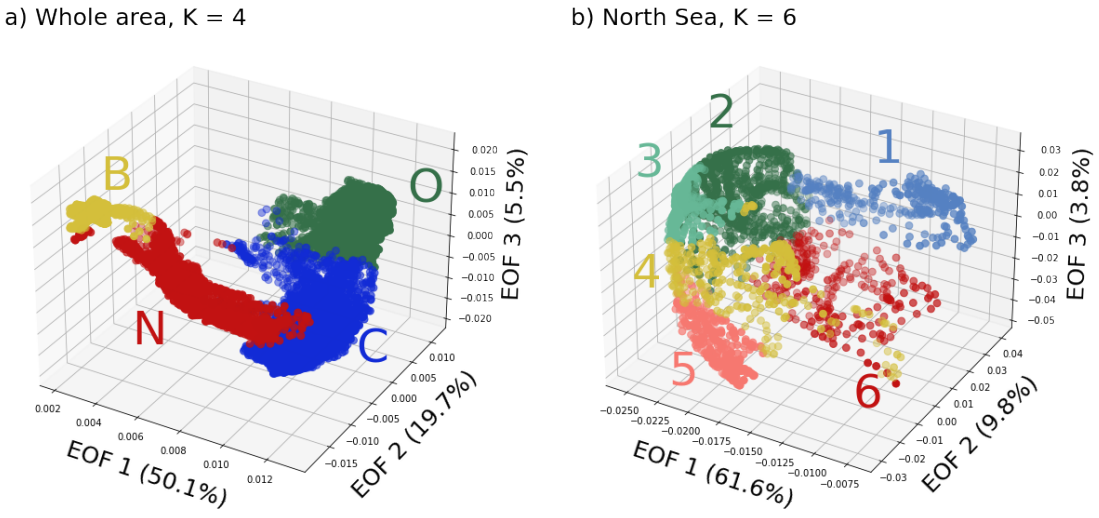

Figure 6. The classification analysis in the abstract EOF space for the two models that use only three EOFs and can therefore be displayed: (a) classification of the whole region of interest into 4 classes shown in Fig. 3a; and (b) classification of the North Sea region into 6 classes shown in Fig. 4b. Each point represents a three-dimensional vector of EOFs that describe a single grid point and the three axes are the three EOFs. Class assignments are indicated using colors.

feature more ambiguous boundaries between regions. The compromise between interpretability and accuracy is not universal and should be tailored to the application at hand. In our case, the balance is struck when the classification is able to highlight novel ideas about the spatial coherence of sea level variability in our study region. Ultimately, unsupervised classification methods can be useful as “hypothesis generation tools" (Kaiser et al., 2022).

Since this paper includes two models based on only three EOFs, these models can be directly depicted in the abstract EOF space to see how the classes are distributed (Fig. 6). The simplest model for the whole region (Fig. 6a and Fig. 3a) has a generally clear separation between the classes and our conclusions about which EOF is responsible for which class border from the beginning of this subsection are confirmed. We can also see where the grid points belonging to the Bothnian Bay and the Norwegian coast are in the abstract space and how they were sorted into class N instead of classes B and C, where the other, more detailed models, sorted them into. The North Sea region (Fig. 6b and Fig. 4b) looks a lot more complex in the abstract EOF space, with less clear boundaries between the classes. We can see that classes 2-5, i.e., the classes inside the North Sea,

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

are generally close to each other in the EOF space, while class 1, which mainly contains the continental shelf outside of the North Sea, and class 6, which contains the western coasts, are both much more widespread, but also more clearly separated from other classes.

4 Conclusions

Gaussian Mixture Modeling, an unsupervised classification approach based on the assumption that all probability density functions can be described as a weighted sum of Gaussian PDFs, can be used to find regions of coherent sea level variability based on satellite altimetry data. Here, we focused on the northern European coastal shelf area and a small adjacent part of the Atlantic Ocean, but the method is applicable in any region. While it is technically possible to use the time series of the sea level data directly as input for the GMM, that approach makes the fitting extremely slow and introduces too much noise for the model to converge towards a single classification solution. Using the empirical orthogonal function maps, the spatial part obtained with a principal component analysis, as input allows us to include most of the observed variability but with greatly reduced dimensionality and noise level.

After reducing the dimensionality, the GMM is able to separate our region of interest into a relatively small number of classes. However, if we want to use more than six mixture components, the models start to diverge, with results varying slightly between individual model runs. Since the models generally find the same patterns despite some differences between them, we show here that we can use an ensemble approach to find the most common classification by applying soft voting, i.e., selecting the class which most models chose with a high probability. The ensemble also gives a likelihood of a model assigning this particular class for each grid point, which tells us how robust the classification is and how difficult it was for the models to classify each area. By comparing the class means with the EOF maps used as input for the GMM ensemble, we showed that we can usually see which class border is based on which EOF, making the model explainable to some extent, and thus directly useful for scientific analysis.

The simplest classification of our entire region of interest, i.e., the classification based on only 3 EOF maps, mostly follows the bathymetry and the coastlines. After including more EOF maps, this basic separation remains the same, but the models are able to also find class borders that are based on ocean dynamics. The largest number of classes with which we can achieve robust results for our region is 10, and we need to use 11 EOF maps as input. This model finds two classes along the Norwegian coast, two in the North Sea, and only one for the whole Baltic Sea. Since the complexity of these three regions vary significantly between them, we show that we can achieve a much more detailed classification if we focus on each region separately. In this way the GMM separate the Baltic into three regions, with virtually all models selecting the same classes. Since the North Sea is very complex, it is split into four classes with high likelihood. The classes belonging to the overlapping area from these two models match.

This classification method is not based on any arbitrary threshold or even on the geographical information such as longitude and latitude, so it is applicable to other ocean regions. It could also be used for finding patterns of sea level variability on different temporal scales, both shorter, such as mesoscale eddies or storm surges, or longer, such as decadal changes or trends.

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

It is not limited to altimetry observations; it could also easily be applied to *in situ* observations or to model data, to study past and future sea level variability that is changing in response to climate change. It can be used on its own, to gain more insight into the patterns of sea level variability, or just as a step in data processing, to create a mask for separating the ocean into regions, which can then be further examined with other methods. Finally, the method is not limited to sea level, it could be used for any other variable.

Code and data availability. In this work we used satellite altimetry observations from Copernicus Marine Service (https://data.marine.copernicus.eu/product/SEALEVEL_GLO_PHY_L4_MY_008_047/description, downloaded on 2022-02-10). Bathymeric information is from the General Bathymetric Chart of the Ocean GEBCO (https://doi.org/10.5285/e0f0bb80-ab44-2739-e053-6c86abc0289c, accesed on 2022-11-07). Code for the ensemble classification with GMMs is available at https://github.com/leapor/GMMensemble.git (DOI for it will be acquired at a later date).

Author contributions. LP and CH conceived the idea for this work. LP planned and performed the experiments, with the help of ST and DJ. LP, CH, and DJ analysed the results. LP wrote the manuscript, while all authors commented on it and provided critical feedback.

Competing interests. No competing interests are present.

Acknowledgements. This work is part of the research project "Would the Northern European Enclosure Dam really protect Sweden from sea level rise? (NEEDS)" funded by the Swedish Research Council for Sustainable Development (Formas, dnr 2020-00982, awarded to CH). DJ was supported by UK Research and Innovation (grant no. MR/T020822/1). ST was funded by UK Research and Innovation (grant no. EP/S022961/1).

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

References

Ablain, M., Cazenave, A., Larnicol, G., Balmaseda, M., Cipollini, P., Faugère, Y., Fernandes, M. J., Henry, O., Johannessen, J. A., Knudsen, P., Andersen, O., Legeais, J., Meyssignac, B., Picot, N., Roca, M., Rudenko, S., Scharffenberg, M. G., Stammer, D., Timms, G., and Benveniste, J.: Improved sea level record over the satellite altimetry era (1993–2010) from the Climate Change Initiative project, Ocean Science, 11, 67–82, https://doi.org/10.5194/os-11-67-2015, 2015.

Barbosa, S., Gouveia, S., and Alonso, A.: Wavelet-based clustering of sea level records, Mathematical Geosciences, 48, 149–162, https://doi.org/10.1007/s11004-015-9623-9, 2016.

Bilmes, J. A.: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, International Computer Science Institute, Berkley, California, 1998.

Bishop, C. M.: Pattern Recognition and Machine Learning, Springer, "New York", 2006.

Björnsson, H. and Venegas, S.: A Manual for EOF and SVD Analysises of Climatic Data, McGill University, 1997.

Bulczak, A. I., Bacon, S., Naveira Garabato, A. C., Ridout, A., Sonnewald, M. J. P., and Laxon, S. W.: Seasonal variability of sea surface height in the coastal waters and deep basins of the Nordic Seas, Geophysical Research Letters, 42, 113–120, https://doi.org/10.1002/2014GL061796, 2015.

Dangendorf, S., Calafat, F. M., Arns, A., Wahl, T., Haigh, I. D., and Jensen, J.: Mean sea level variability in the North Sea: Processes and implications, Journal of Geophysical Research: Oceans, 119, https://doi.org/10.1002/2014JC009901, 2014.

Dangendorf, S., Frederikse, T., Chafik, L., Klinck, J., Ezer, T., and Hamlington, B.: Data-driven reconstruction reveals large-scale ocean circulation control on coastal sea level, Nature Climate Change, 11, 514–520, https://doi.org/10.1038/s41558-021-01046-1, 2021.

Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society. Series B (Methodological), 39, 1–38, http://www.jstor.org/stable/2984875, 1977.

Ducet, N., Le Traon, P. Y., and Reverdin, G.: Global high-resolution mapping of ocean circulation from TOPEX/Poseidon and ERS-1 and -2, Journal of Geophysical Research: Oceans, 105, 19 477–19 498, https://doi.org/10.1029/2000JC900063, 2000.

Fox-Kemper, B., Hewitt, H., Xiao, C., Aðalgeirsdóttir, G., Drijfhout, S., Edwards, T., Golledge, N., Hemer, M., Kopp, R., Krinner, G., Mix, A., Notz, D., Nowicki, S., Nurhati, I., Ruiz, L., Sallée, J.-B., Slangen, A., and Yu, Y.: Ocean, Cryosphere and Sea Level Change, p. 1211–1362, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.011, 2021.

GEBCO Compilation Group: GEBCO_2022 Grid, https://doi.org/10.5285/e0f0bb80-ab44-2739-e053-6c86abc0289c, accessed: 2022-11-07, 2022.

Iglesias, I., Lorenzo, M. N., Lázaro, C., Fernandes, M. J., and Bastos, L.: Sea level anomaly in the North Atlantic and seas around Europe: Long-term variability and response to North Atlantic teleconnection patterns, Science of The Total Environment, 609, 861–874, https://doi.org/10.1016/j.scitotenv.2017.07.220, 2017.

Jones, D. C., Holt, H. J., Meijers, A. J. S., and Shuckburgh, E.: Unsupervised Clustering of Southern Ocean Argo Float Temperature Profiles, Journal of Geophysical Research: Oceans, 124, 390–402, https://doi.org/https://doi.org/10.1029/2018JC014629, 2019.

Kaiser, B. E., Saenz, J. A., Sonnewald, M., and Livescu, D.: Automated identification of dominant physical processes, Engineering Applications of Artificial Intelligence, 116, 105 496, https://doi.org/10.1016/j.engappai.2022.105496, 2022.

Kohonen, T.: The self-organizing map, Proceedings of the IEEE, 78, 1464–1480, https://doi.org/10.1109/5.58325, 1990.

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

Kohonen, T. and Mäkisara, K.: The self-organizing feature maps, Physica Scripta, 39, 168, https://doi.org/10.1088/0031-8949/39/1/027, 1989.

Le Traon, P.-Y. and Ogor, F.: ERS-1/2 orbit improvement using TOPEX/POSEIDON: The 2 cm challenge, Journal of Geophysical Research: Oceans, 103, 8045–8057, https://doi.org/10.1029/97JC01917, 1998.

Liu, Y. and Weisberg, R. H.: Patterns of ocean current variability on the West Florida Shelf using the self-organizing map, Journal of Geophysical Research: Oceans, 110, https://doi.org/10.1029/2004JC002786, 2005.

Liu, Y., Weisberg, R. H., and Yaochu, Y.: Patterns of upper layer circulation variability in the South China Sea from satellite altimetry using the self-organizing map, Acta Oceanologica Sinica, 27, 129–144, 2008.

Lloyd, S.: Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM, IEEE Trans. Inform. Theor.(1957/1982), 18, 1957.

Maze, G., Mercier, H., Fablet, R., Tandeo, P., Lopez Radcenco, M., Lenca, P., Feucher, C., and Le Goff, C.: Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean, Progress in Oceanography, 151, 275–292, https://doi.org/10.1016/j.pocean.2016.12.008, 2017.

Mihanović, H., Cosoli, S., Vilibić, I., Ivanković, D., Dadić, V., and Gačić, M.: Surface current patterns in the northern Adriatic extracted from high-frequency radar data using self-organizing map analysis, Journal of Geophysical Research: Oceans, 116, https://doi.org/10.1029/2011JC007104, 2011.

Papadopoulos, A. and Tsimplis, M. N.: Coherent Coastal Sea-Level Variability at Interdecadal and Interannual Scales from Tide Gauges, Journal of Coastal Research, 2006, 625 – 639, https://doi.org/10.2112/04-0156.1, 2006.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Pujol, M.-I. and Mertz, F.: Global Ocean Gridded L 4 Sea Surface Heights And Derived Variables Reprocessed 1993 Ongoing, https://doi.org/10.48670/moi-00148, accessed: 2022-02-10, 2020.

Pujol, M.-I., Faugère, Y., Taburet, G., Dupuy, S., Pelloquin, C., Ablain, M., and Picot, N.: DUACS DT2014: the new multi-mission altimeter data set reprocessed over 20 years, Ocean Science, 12, 1067–1090, https://doi.org/10.5194/os-12-1067-2016, 2016.

Rosso, I., Mazloff, M. R., Talley, L. D., Purkey, S. G., Freeman, N. M., and Maze, G.: Water Mass and Biogeochemical Variability in the Kerguelen Sector of the Southern Ocean: A Machine Learning Approach for a Mixing Hot Spot, Journal of Geophysical Research: Oceans, 125, e2019JC015 877, https://doi.org/10.1029/2019JC015877, 2020.

Scotto, M. G., Alonso, A. M., and Barbosa, S. M.: Clustering Time Series of Sea Levels: Extreme Value Approach, Journal of Waterway, Port, Coastal, and Ocean Engineering, 136, 215–225, https://doi.org/10.1061/(ASCE)WW.1943-5460.0000045, 2010.

Sterlini, P., de Vries, H., and Katsman, C.: Sea surface height variability in the North East Atlantic from satellite altimetry, Climate Dynamics, 47, 1285–1302, https://doi.org/10.1007/s00382-015-2901-x, 2016.

Taburet, G., Sanchez-Roman, A., Ballarotta, M., Pujol, M.-I., Legeais, J.-F., Fournier, F., Faugere, Y., and Dibarboure, G.: DUACS DT2018: 25 years of reprocessed sea level altimetry products, Ocean Science, 15, 1207–1224, https://doi.org/10.5194/os-15-1207-2019, 2019.

Thomas, S. D. A., Jones, D. C., Faul, A., Mackie, E., and Pauthenet, E.: Defining Southern Ocean fronts using unsupervised classification, Ocean Science, 17, 1545–1562, https://doi.org/10.5194/os-17-1545-2021, 2021.

https://doi.org/10.5194/egusphere-2023-1468
Preprint. Discussion started: 12 July 2023

Traon, P. Y. L., Faugère, Y., Hernandez, F., Dorandeu, J., Mertz, F., and Ablain, M.: Can We Merge GEOSAT Follow-On with TOPEX/Poseidon and ERS-2 for an Improved Description of the Ocean Circulation?, Journal of Atmospheric and Oceanic Technology, 20, 889 – 895, https://doi.org/10.1175/1520-0426(2003)020<0889:CWMGFW>2.0.CO;2, 2003.