Ecological Informatics 93 (2026) 103586

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Consistency of clustering analysis of complex 3D ocean datasets

Rebecca Millington [*], Dale Partridge, Helen R. Powley, Gennadi Lessin, David Moffat, Jerry Blackford

Plymouth Marine Laboratory, Plymouth, UK

ARTICLE INFO

Keywords:
Machine learning
Clustering
Biogeochemical model
Ocean model
Northwest European shelf
Marine ecosystem

ABSTRACT

Rapid advancement of machine learning and artificial intelligence is enabling new analysis techniques to be applied across all fields of scientific research. To aid analysis of the physical or biogeochemical characteristics of the ocean, marine systems have been subdivided into spatial regions where properties exhibit similar distributions or behaviour, such as the Longhurst provinces. Machine learning techniques enable the identification of spatial regions in a robust and transferable way. In this paper we drive clustering algorithms with a variety of input datasets to assess the consistency of resulting clusters. We compare the results of clustering analyses applied separately to physical, biogeochemical and ecological variables at different depths, using model output from a 3D hydrodynamical-biogeochemical model (NEMO-ERSEM) on the Northwest European shelf. Clustering outcomes depended on both the variables and depths input into the algorithm, although some similarities still existed in spatial patterns between each clustering analysis, e.g. clusters were smaller near the coast and relatively extensive in the open ocean. Clusters based on physical properties showed latitudinal distribution, while biogeochemical and ecological inputs resulted in a higher concentration of clusters near the coast. Results from depth-averaged and near-bottom inputs were similar and followed the limits of the shelf-edge, unlike clusters based on surface inputs. Overall, clustering algorithms offer a useful method to define spatial regions with similar characteristics, however, our results emphasise that input data choices should be carefully considered. Our results provide a knowledge foundation which can help future researchers make informed decisions when applying clustering to complex datasets.

1. Introduction

Modern ocean models output large datasets of the order of 100 s of gigabytes. This data comprises of multiple variables, most of them with three spatial (depth, longitude, latitude) and a temporal dimension. To draw useful conclusions from ocean model output, such as predictions of the future carbon cycle, the data needs detailed analysis which often involves reducing the dimensionality by averaging across time and/or space (Nguyen and Holmes, 2019). One commonly used approach is to divide an area into a number of spatial regions with similar characteristics such as physical or biogeochemical properties (Cheung et al., 2017; Longhurst, 1995; Wakelin et al., 2020) or economics (Cavan and Hill, 2022; Daewel et al., 2019). This reduces the data down to a manageable size for analysis, making it easier to characterise patterns and behaviour. In particular, the Longhurst provinces (Longhurst, 1995), which were partitioned based on fields including surface chlorophyll, nutrients, mixed layer depth and photic depth, have frequently been applied to investigate biological trends in the ocean (Leles et al., 2019; Racault et al., 2014; van Oostende et al., 2023).

Choosing the most appropriate way to subdivide an area into spatial regions for an analysis can be challenging. Regions selected following the borders of national jurisdictions often do not reflect boundaries in the physical or biogeochemical properties (e.g. ocean thermal fronts (Miller and Christodoulou, 2014)). If a region contains two subregions where a variable of interest shows opposing behaviour, trends can be missed when an average is taken across the whole region, and misleading conclusions drawn. Even when regions are selected based on past physical or biogeochemical properties, they may need to be reevaluated if the system behaviour changes in the future, for example, due to climate change. Therefore, a flexible method is needed to robustly identify spatial regions based on varying numbers of variables and for appropriate time frames.

Unsupervised clustering methods use mathematical algorithms to partition data into groups (clusters) with similar characteristics, such as

[*] Corresponding author.
E-mail address: rmi@pml.ac.uk (R. Millington).

https://doi.org/10.1016/j.ecoinf.2025.103586
Received 18 June 2025; Received in revised form 22 December 2025; Accepted 23 December 2025
Available online 27 December 2025

spatial regions with similar behaviour in multiple physical and biogeochemical properties. When clustering is 'unsupervised', the cluster properties are an outcome of the underlying data distributions, as opposed to 'supervised' clustering, where the algorithm is trained on data labels which have been predefined by the user. Unsupervised clustering has previously been applied to physical and biogeochemical ocean model output and satellite data to aid analysis of patterns at regional and global scales (Atwood et al., 2024; Jarníková et al., 2022; Sonnewald et al., 2019, 2020). Clustering has also been used to allow comparison between sparse in-situ data and model data by aggregating data in regions with similar behaviour (McGinty et al., 2023; Sun et al., 2021).

Clustering algorithms need different properties depending on the application. For some algorithms, the number of clusters is user defined (e.g. *k*-means (Cam and Neyman, 1967)) and for others, the number of clusters is emergent (e.g. DBSCAN, self-organising maps (Ester et al., 1996; Kohonen, 1982)). The algorithm K-shape is specifically designed to capture similarities in time-series regardless of temporal offsets and differences in magnitude (Paparrizos and Gravano, 2015). Each method necessarily makes assumptions around the data inputs, such as the distribution the data follows, the existence of spatial and temporal correlations within the data and the distribution of the associated noise component.

Observational and modelled biogeochemical datasets are often comprised of a large number of different variables, from across a range of depths. For computational reasons, the entire data set cannot normally be provided to the clustering algorithm as input, so some reduction in dimensionality is necessary, which may involve selecting data from a particular depth (Sun et al., 2021), using a subset of variables (Sonnewald et al., 2020), or taking the temporal mean (Kavanaugh et al., 2014; Sonnewald et al., 2019). Additional processing, such as the use of principle component analysis, can reduce the number of variables further (Zhao et al., 2020).

Spatially and temporally explicit studies generally use one set of variables, and, when satellite data is used, remain restricted to the surface layer (McGinty et al., 2023; Racault et al., 2014; Sonnewald et al., 2020). Clustering outcomes may differ across depth or depending on which variables are included in the input datasets. Furthermore, including highly correlated variables such as temperature and photosynthetically active radiation may lead to unintentionally biased clustering outcomes (Zhao et al., 2020).

Here, we use different variable and depth inputs to explore how these choices impact the distribution of the resulting clusters within our study region. We do this using the *k*-means algorithm, which has frequently been applied to marine systems (McGinty et al., 2023; Sonnewald et al., 2019; Sun et al., 2021). We investigate how varying the data input into clustering algorithms affects the consistency of clustering outcomes, and therefore how clustering algorithms can be applied to robustly create bespoke regions for data analysis. We then investigate how correlations between input variables drive clustering outcomes. Furthermore, we hypothesise that clusters from a subset of a dataset are more similar to clusters from the whole dataset than clusters from a separate dataset. Secondly, we test the expectation that when datasets are combined, the larger dataset will dominate in determining clusters.

We use the Northwest European shelf as a testbed to explore these questions. Shelf regions account for around 10 % of the global ocean area (Harris et al., 2014) and are highly productive (Bauer et al., 2013; Pauly et al., 2002). The Northwest European shelf is important for carbon cycling and food production (Kerby et al., 2012; Legge et al., 2020); at the same time it has been heavily impacted by human activity (Eigaard et al., 2016). The shelf itself is shallow, lying above 200 m, and receives considerable freshwater and nutrient inputs from land. The shelf is subject to several distinct hydrological regimes influencing biological production. For example, the northern North Sea is seasonally stratified while the shallowest areas of the English Channel are permanently mixed (Van Leeuwen et al., 2015). Clustering has previously been applied in the northeast Atlantic to investigate seasonal controls of nitrous oxide emissions (Lessin et al., 2020) and to identify trends in observed zooplankton abundance (McGinty et al., 2023).

Clustering has the potential to help answer many important questions both on the shelf and across the global oceans: for example, to find areas that are sinks and sources of carbon, to understand which areas respond in a similar way to climate change, and to identify hotspots of environmental impacts. In this paper, we answer the question: *"how consistent are the results of clustering algorithms when applied to complex 3D ocean datasets?"* and aim to build a solid knowledge foundation to enable the application of clustering for analysis of complex marine data.

2. Methods

2.1. k-means clustering algorithm

We have used the *k*-means algorithm (Cam and Neyman, 1967; Pedregosa et al., 2011) to explore the impact of different input choices on the distribution of calculated clusters. *k*-means is known to perform well with large datasets and is robust to any assumptions on data normality, such as likely found in ocean model data (Ikotun et al., 2023). A variety of commonly-used clustering algorithms, including *k*-means, have been implemented in Python in the package scikit-learn (Pedregosa et al., 2011).

k-means minimises the 'within-cluster sum of squares':

$$\sum_{i=0}^{n} min\left(\left|\left|x_i - \mu_j\right|\right|^2\right) \tag{1}$$

where μ_j is the mean value of the samples in the *j*th cluster, also known as the centroid, and x_i is the *i*th sample in a set of *n* samples. The number of clusters is input by the user. Other distance metrics, such as the cosine similarity, can be used depending on the geometry of the input data (Zhao et al., 2020). *k*-means is implemented by first choosing initial centroids for the clusters. Each sample is then assigned to the cluster with the nearest centroid, using the Euclidean distance as defined in Eq. (1). Then new centroids are calculated by taking the mean of the samples assigned to each cluster. These two steps are repeated until the difference between the old and new centroids is less than a predefined threshold, and therefore the clusters found are not changing between each iteration. The Python package scikit-learn includes an implementation of the *k*-means++ algorithm for robust initialisation (Pedregosa et al., 2011). The *k*-means++ algorithm outputs a set of initial conditions for the *k*-means algorithm, which allow quicker convergence of the algorithm and a reproducible final set of clusters.

2.2. Biogeochemical model output

In this study, output from the European Regional Sea Ecosystem Model (ERSEM) was used, giving a dataset on a spatially and temporally regular grid. ERSEM is a marine biogeochemical-ecosystem model which includes four phytoplankton groups, three zooplankton groups and bacteria as well as the cycling of carbon, nitrogen, phosphorus, silicate and oxygen (Butenschön et al., 2016). ERSEM was run coupled to the hydrodynamic model Nucleus for European Modelling of the Ocean (NEMO) (Madec et al., 2017) via the Framework for Aquatic Biogeochemistry Models (Bruggeman and Bolding, 2014). The run covered the years 1981–2017 over the 7 km resolution Atlantic Margin Model domain, which covers the Northwest European shelf (Fig. 1) (Edwards et al., 2012). Similar model setups have previously been applied to investigate nitrous oxide emissions, near-bed oxygen under climate change and the fate of terrigenous organic carbon inputs from rivers (Galli et al., 2024; Lessin et al., 2020; Powley et al., 2024). The region to the east of Denmark (>10E, 55N to 60N) was not included in the following analysis as it was overly influenced by the Baltic Sea boundary conditions.

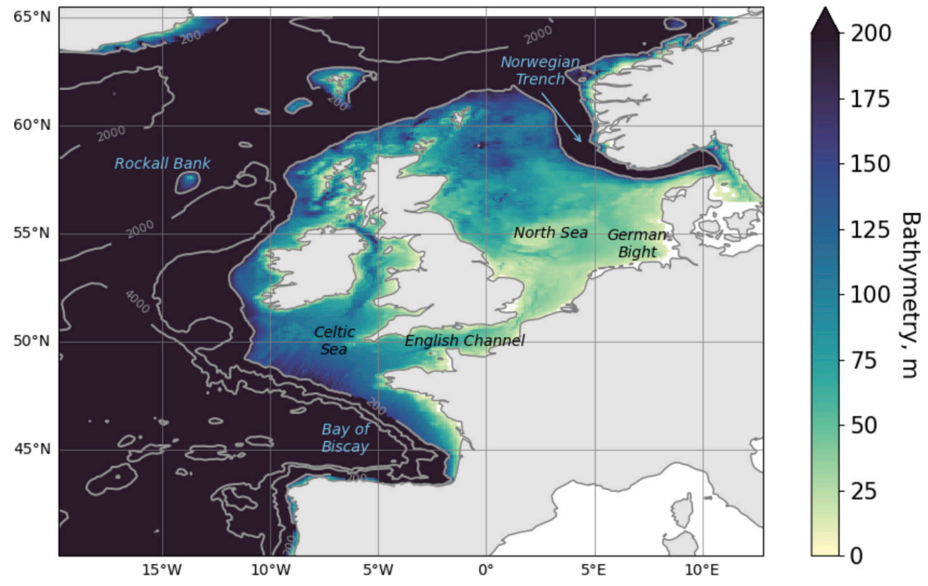

Fig. 1. Map of the model domain, including the Northwest European shelf, with key areas labelled. Colours show the depths down the 200 m, to delineate the bathymetry of the shelf. The 200 m, 2000 m and 4000 m isobaths are marked with grey lines.

When analysing multiple variables, *k*-means can only process one time point. We have used a multi-year average over five years from 2000 to 2004 to remove inter-annual variability, thus excluding the effects of long-term trends.

2.3. Standardisation

The variables input into a clustering algorithm may have different units and different magnitudes. For example, for a subset of biogeochemical model data including phosphate, nitrate, silicate, oxygen, dissolved inorganic carbon (DIC), alkalinity, ammonium, absolute values of DIC and alkalinity concentrations were much larger than the other variables (Fig. 2c). This could result in cluster selection being dominated by a subset of these variables, for example the variable with the largest magnitude, as the version of *k*-means applied here computes clusters using a Euclidean distance metric. As such, the variables were standardised to ensure that the magnitude and the range of the data was not the main influencing factor in the delineation of clusters, and instead

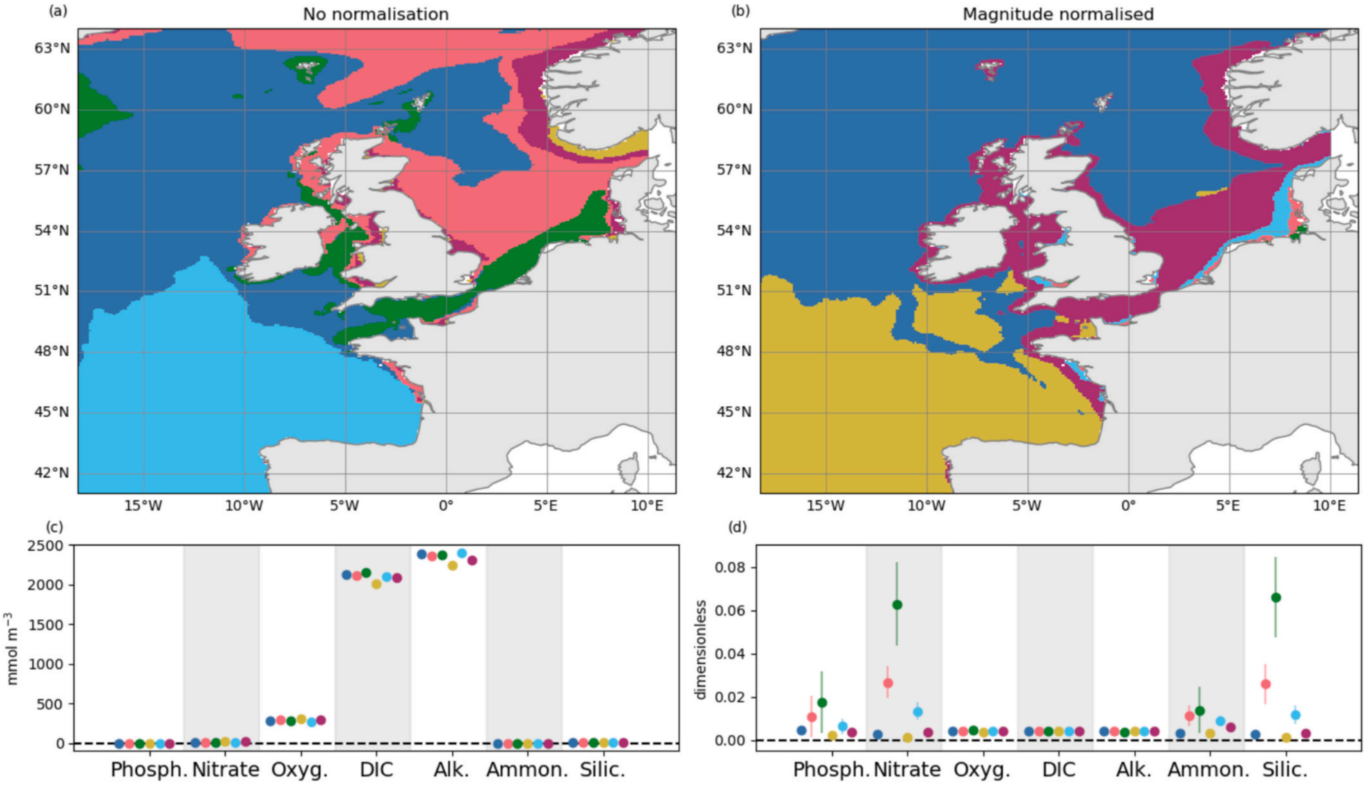

Fig. 2. Clustering output depends on the method of standardisation used. Map of clusters from biogeochemical variables at the surface with a) no standardisation and b) the magnitude normalised (Eq. (2)). Mean and standard deviation for each cluster with c) no standardisation and d) the magnitude normalised.

the spatial distribution of each variable became the factor which determined cluster membership. The distributions of the variables were transformed to have overall more similar characteristics, such as a mean of zero, and a variance of one, which makes variables more comparable and results in clusters not being dominated by a subset of the input variables.

Two methods of standardisation were tested, to explore how different levels of transformation affected the resulting clusters. These methods were (i) standardising the magnitude only and (ii) standardising the magnitude and the spatial variation.

(i) To standardise the magnitude, each variable, v, was divided by its spatial norm:

$$v \rightarrow \frac{v}{|v|}. \tag{2}$$

so that all variables had similar magnitudes (Fig. 2b, d). When k-means clustering was applied to input data with its magnitude standardised, the resulting clusters were different to clusters from input data without standardisation (Fig. 2). In the input data with standardised magnitude, nitrate and silicate had higher spatial variability than other variables (Fig. 2d). When clustering is applied to variables with different spatial variability, the variables with higher spatial variability may dominate in determining the clusters found. For example, when clustering was performed on phosphate and nitrate together, the clusters found were the same as clusters found based on nitrate individually (fig. S1), as nitrate had higher spatial variability.

(ii) In the case that both the magnitude and variability were standardised, the spatial mean of the variable was subtracted from the variable, which was subsequently divided by the spatial standard deviation of the variable,

$$\mathrm{v} \rightarrow \frac{v - \mathrm{mean(v)}}{\mathrm{stdev(v)}}. \tag{3}$$

Method (ii) was used to standardise the data throughout the analysis.

2.4. Metrics

2.4.1. Silhouette score

The silhouette score can be used as a metric for the quality of clustering, by identifying how distinct individual clusters are (Rousseeuw, 1987). A set of clusters may be considered to be ‘good’ when the points within a cluster are close to other points within the cluster and far from points in different clusters. The score can have values between −1 and 1, with lower values indicating less separated clusters. For the ith sample, the silhouette score, s_i, is

$$s_i = \frac{b - a}{max(a, b)}, \tag{4}$$

where a is the mean distance between a sample and all other points in the same cluster and b is the mean distance between a sample and other points in the next nearest cluster. The next nearest cluster is the cluster with the smallest mean distance from the sample, that is not the cluster the sample is a member of, but the next best fit. This measures how well one point is clustered. To get a measure of overall clustering quality, the mean silhouette score is calculated for all individual samples in the dataset.

2.4.2. Rand index

Similarity between two sets of clusters can be assessed by calculating the Rand index (Rand, 1971). The index ranges from 0 to 1; a value close to one means that the two sets of clusters are very similar and a value close to zero means that the two sets are very different. It is calculated by comparing each possible pair of points using

$$\text{Rand index} = \frac{\mathrm{a + b}}{\text{Total number of pairs}} \tag{5}$$

where a is the number of pairs where both points from the pair are in the same cluster in both sets of clusters, and b is the number of pairs where the points from the pair are in different clusters in both sets of clusters. The denominator is the total number of pairs – for n points, this will be n (n-1)/2 pairs.

2.5. Number of clusters

The number of clusters to be calculated by the k-means algorithm is chosen by the user, rather than being an emergent property of the clustering process as is the case with some other clustering algorithms, for example DBSCAN and self-organising maps (Ester et al., 1996; Kohonen, 1982). The dataset which clustering is performed on may contain features which mean that a particular number of clusters is best suited to the data. For example, if the values of variables in a dataset are purely determined by whether a point is on- or off-shelf, two clusters will suit the data best. The silhouette score provides a metric to test assumptions on the distribution of input data (Section 2.4.1). Other considerations may be necessary when choosing the number of clusters such as the minimum number required to resolve a particular feature or the maximum number that there is the capacity to analyse.

The silhouette score for different numbers of clusters was calculated to investigate which number of clusters provided higher quality output. k-means clustering was applied to subsets of seven surface biogeochemical variables (phosphate, nitrate, silicate, oxygen, dissolved inorganic carbon (DIC), alkalinity, ammonium). Clusters were calculated using one to seven variables, from all possible combinations of seven biogeochemical variables, i.e. seven subsets containing one variable, 21 subsets containing two variables etc. On each set of variables, k-means was applied 19 times to output between two and 20 clusters, inclusive. The silhouette score was then calculated for the output clusters.

On average, for most subsets of input data, dividing the dataset into a lower number of clusters led to a higher silhouette score which indicates

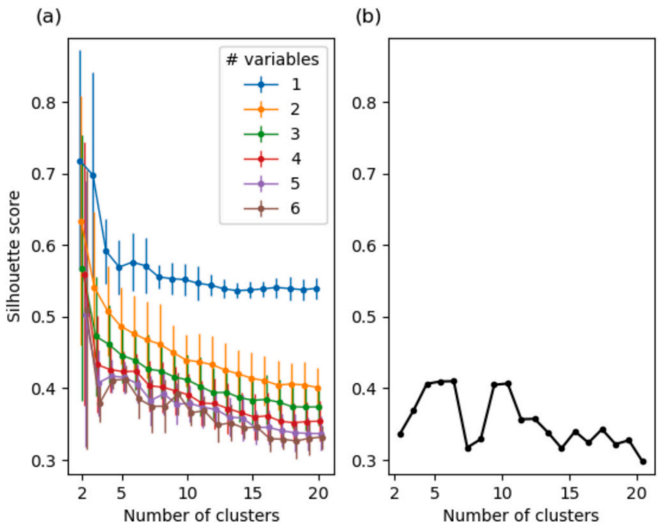

Fig. 3. Choosing the number of clusters when applying the k-means algorithm. a) Using the silhouette score to assess the quality of clustering outcomes for two to twenty clusters when one to six variables are used. Colours show the number of variables used as input, points show the mean and error bars the standard deviation. b) Silhouette score for two to twenty clusters when one set of seven variables is used.

more separation between clusters (Fig. 3a). However, within a particular set of variables, there may be peaks at larger numbers of clusters depending on the characteristics of the actual data being used. For example, with one permutation of 7 variables, there was a maximum in the silhouette score at 6 clusters (Fig. 3b). When more variables were included, the silhouette score was lower, suggesting that different variables follow different natural partitions into clusters (Fig. 3a). Too few clusters might provide insufficient detail for analysis, while too many might not sufficiently reduce the complexity of the data for its consequent analysis. Calculating the silhouette score can aid with choosing the most suitable number of clusters to compute. As a result of this analysis, we have used 6 clusters throughout the rest of the study.

2.6. *Consistency of clusters based on different input datasets*

To investigate the consistency of clustering outcomes from input datasets formed of different variables and depths, we have used subsets of the full output from a biogeochemical model as input to the *k*-means algorithm.

2.6.1. *Calculation of sets of clusters from different input variables*

Clustering analysis was carried out across three sets of variables, to compare how clusters varied by the type of input used. The sets of variables were chosen to group together the physical, biogeochemical and ecological characteristics of the system. There is a wide scope for alternative groupings and the sets chosen are not meant to represent comprehensive choices but rather correspond to some of the commonly observed environmental variables. Within the sets, the variables used were:

- Physics: temperature, salinity, mixed layer depth and photosynthetically active radiation (PAR).
- Biogeochemical: phosphate, nitrate, silicate, oxygen, dissolved inorganic carbon (DIC), alkalinity, ammonium
- Ecological: phytoplankton biomass, zooplankton biomass, dissolved organic carbon (DOC), detrital particulate organic carbon (POC), bacteria biomass.

Phytoplankton and zooplankton biomass were calculated as the sum across four and three modelled functional groups, respectively. DOC represents the sum of three pools with different labilities. POC was calculated as the sum of three different size classes.

Clustering analysis was also performed on input data comprising combinations of the sets of variables: physics and biogeochemical; physics and ecological; biogeochemical and ecological; all sets.

2.6.2. *Comparison of sets of clusters from different input variables*

The Rand index was used to compare the results of clustering performed on different sets of variables to test their similarity. As well as comparisons between each variable set (physics, biogeochemistry, ecological), comparisons were made with clustering outcomes from combinations of the variable sets (physics-biogeochemistry, physics-ecological, biogeochemistry-ecological, all sets).

2.6.3. *Comparison of sets of clusters from different input depths*

Clustering analysis was carried out using model data from different depths, to compare how clusters differed with the depth of the input data. Ocean properties vary across depth, with physics determining the water column structure, which in turn impacts biogeochemical and ecological characteristics. The surface of the ocean has the most observations and is the only part of the ocean observed by satellite. It is ecologically important with high light levels, but only studying the surface gives an incomplete picture of the system. Depth-averaged data captures the whole water column. The bottom of the ocean is of ecological importance in particular for the benthic communities which are highly biodiverse, but at risk from human impacts such as climate change, mining and trawling. Variables in the ecological set typically have higher concentrations towards the surface and very low concentrations below the euphotic layer. Therefore, the magnitude of these variables will not be dominated by the bathymetry when depth integrated. Data was used from the surface, bottom and depth-averaged across the water column, as well as depth-integrated data for ecological variables only.

The set of physics variables included the mixed layer depth, which is a widely used and biologically meaningful metric, as it defines the main area of biological activity. When physics variables were calculated at different depths, the mixed layer depth remained constant, regardless of the depth used.

Depth-averaged values were calculated as a sum over the water column divided by the depth. Depth-integrated values were calculated as a sum over the water column. Surface and bottom values were taken from the top and bottom layer of model output, respectively.

Surface values are often emphasised over other depths in ocean science due to the bias towards monitoring surface waters. To test how representative the surface layer is of the water column, clusters from biogeochemical variables at depths between 0 and 100 m were calculated at 10 m intervals and compared to clusters from the surface and the bottom. To calculate clusters at intermediate depths, interpolation from the original model grid was carried out using nctoolkit (Wilson and Artioli, 2023). In some areas, the bottom is shallower than the depths chosen, so data from the bottom was used rather than interpolation. For example, when calculating clusters at 100 m, in any areas shallower than 100 m, data from the bottom layer was used.

The Rand index (Section 2.4.2) was used to compare clustering outcomes from different depths to test their similarity.

2.7. *Role of input variables in determining clustering outcomes*

Clustering outcomes may depend on some input variables more than on others. To explore the relative role of each variable, we calculated different sets of clusters: one overall set with all variables present as well as sets with each variable removed for physical, biogeochemical and ecological variables. We then made comparisons using the Rand index (Eq. (5)) between the set of clusters with all variables and the sets of clusters from data with one variable removed. For a dataset with 7 variables, 7 comparisons were made. We expect a variable that is influential in determining the shape and distribution of overall clusters, when removed, to result in a lower Rand index, while removing a less influential variable to result in a higher Rand index.

To understand whether similarity with other variables affected how influential a variable was, the spatial correlation was calculated for each pair of variables using the R-squared score (the coefficient of determination). The R-squared score was calculated by squaring the Pearson correlation coefficient, which is implemented in Scipy (Virtanen et al., 2020). To test the relationship between variable spatial correlations and the Rand index from clustering outcomes, the R-squared scores were compared to the Rand indices for each variable. The mean R-squared score for a variable was calculated by taking the average of the R-squared scores for that variable correlated to each other variable in the set. The mean R-squared score for each variable was correlated to the corresponding Rand index comparing clusters from the complete set with clusters from the set with that variable removed. We expected the most correlated variables to contribute less to cluster outcomes, as these variables have the most redundancy within the dataset, and the least correlated variables to be more influential. Therefore, the mean R-squared scores would be positively correlated to the Rand indices.

3. Results

3.1. *Comparison of clusters based on different sets of variables*

Unsupervised clustering was performed using *k*-means with subsets of biogeochemical model output. We explored how the choice of variables affected the resulting cluster distributions by calculating clusters using sets of physical, biogeochemical and ecological variables at the surface (Fig. 4). Cluster outputs had similar silhouette scores suggesting similar quality of clusters (Table S1). The clusters found in coastal regions were spatially restricted compared to the clusters in shelf and off-shelf regions, where variability occurs over larger spatial scales. Cluster boundaries generally did not follow the shelf edge. Biogeochemical variables produced clusters with mean values furthest from the domain average, linked to biogeochemical clusters also varying the most in overall surface area (Table S1).

Physics-based clusters had the strongest latitudinal dependence with less partitioning in coastal areas (Fig. 4a, b). The average temperature and photosynthetically active radiation (PAR) of the clusters decreased north to south. Cluster 2 in the south of the domain had the highest temperature and PAR. Cluster 3 grouped points off-shelf in the Northwest area of the domain which had above average mixed layer depth. Cluster 5 had very low salinity and grouped areas near the coast which are influenced by freshwater output by rivers.

Biogeochemical clusters had less partitioning off-shelf and more coastal clusters than physics clusters (Fig. 4c, d). Coastal clusters (3,5,6) tended to have values further apart from the domain average, such as high nutrients and low alkalinity, especially cluster 5, where nitrate and silicate values were over 15 standard deviations greater than the mean (Fig. 4c). The northern coastal cluster (6) had lower dissolved inorganic carbon (DIC) and alkalinity than the southern coastal cluster (3). Much of the Northwest European shelf was grouped into one cluster (1), with higher than average ammonium. Two very large clusters grouped the remaining shelf and off-shelf regions, split by latitude. Cluster 2 grouped the northern part of the North Sea with off-shelf areas, likely due to the impact of water mass advection onto the shelf (Fig. 4b)(Holt and Proctor, 2008), and had close to average values. The southern cluster had lower nutrients, oxygen and DIC, and higher alkalinity than the average.

Coastal ecological clusters had high concentrations of most variables (Fig. 4e, f, clusters 3, 4 and 6). Cluster 3 had particularly high zooplankton and phytoplankton content and cluster 4 had particularly high particulate organic carbon (POC). Cluster 5, covering the Atlantic margin of the domain, had lower than average values for all variables. This was a consequence of low phytoplankton concentrations, applied as Atlantic boundary conditions to ensure numerical stability, being

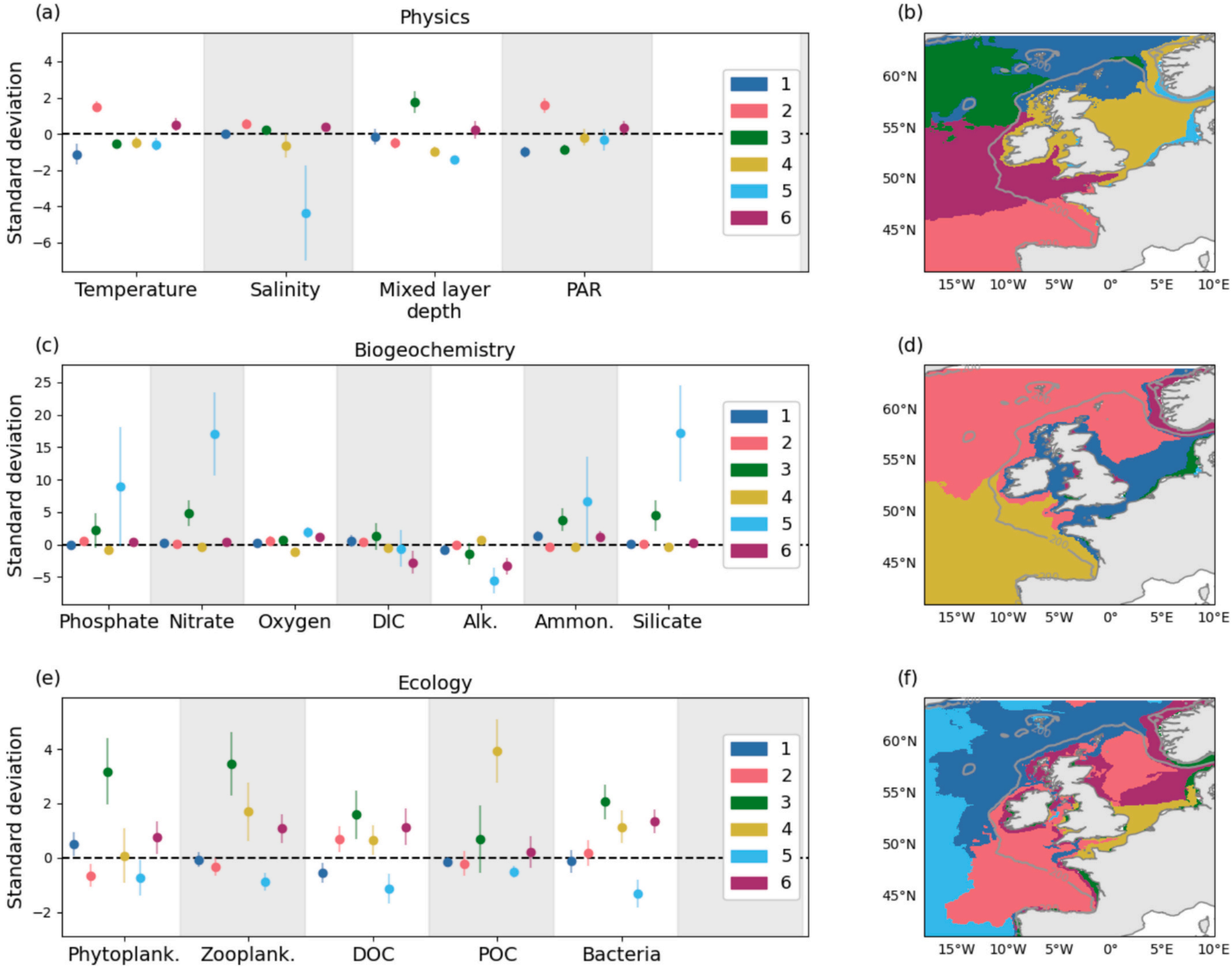

Fig. 4. Choice of variables included in input data affects cluster spatial patterns. Clusters based on *k*-means algorithm performed on multi-year mean surface NEMO-ERSEM model output. Standardised mean and standard deviation of variables grouped by cluster and spatial map of clusters from a,b) physics c,d) biogeochemical and e,f) ecological variables. For each set of variables, six distinct clusters have been identified, shown by different colours. a,c,e) The horizontal black dashed line marks the mean value across the domain for each variable, standardised to zero. The 200 m depth contour is marked with a grey line on the spatial maps. Note the difference in y-axis scales on figures (a), (c) and (e).

advected into the domain (Fig. 4e).

Sets of clusters were compared using the Rand index (Eq. (5)). Comparisons between clusters from the individual sets of variables had similar Rand indexes (0.71 and 0.73, Table 1). Clusters from ecological variables were least similar to clusters from other sets of variables, as demonstrated by the lower Rand index (Table 1). This was partially due to the cluster at the edge of the domain as a consequence of the low phytoplankton boundary condition (Fig. 4c, d).

With clusters computed from the combination of datasets (e.g. physics + biogeochemistry), it might be expected that the dataset with a larger number of variables would dominate in determining the output clusters. However, clusters based on physics were more similar to clusters based on biogeochemistry + ecology variables than clusters from only ecology variables (Table 1). This is partially due to the cluster attributed to the boundary condition in ecology-only clusters not being present in the biogeochemistry + ecology clusters (fig. S2).

3.2. *Comparison of clusters from different depths*

Clusters were computed using physics, biogeochemical and ecological data at the surface, the bottom and depth-averaged across the water column. In addition, clusters were calculated from depth-integrated data for ecological variables.

Bottom and depth-averaged physics clusters were similar to each other while clusters from the surface were different (Fig. 5, Table 2). Clusters at the surface did not follow the shelf edge in contrast to clusters based on depth-averaged and bottom data. At the surface, physics clusters had strong latitudinal dependence with little partitioning in coastal areas (Fig. 5a). All depths grouped an area in the Northwest of the domain, which in the surface data was due to a greater mixed layer depth (Fig. 4a).

All sets of biogeochemical clusters reflected the bathymetry, but clusters from the surface did not follow the depth contour marking the edge of the shelf (Figs. 4, 6a-c). There were more coastal clusters from the surface than the bottom and depth-averaged, which had more off-shelf clusters (Fig. 6a-c). Depth-averaged and bottom clusters followed the shelf edge and captured the Rockall bank (Fig. 6b, c, Fig. 1). The Norwegian trench was grouped differently in each case, sharing a near-coastal cluster with surface data, clustered with an off-shelf region with depth-averaged data, and clustered with the shelf with bottom data.

The Rand index showed that clusters from the bottom and depth-averaged data were more similar to each other than they were to clusters from the surface (Table 2). When used to compare clusters from depths between 0 and 100 m to clusters from the surface, the Rand index decreased sharply to around 0.8 over 0-40 m (Fig. 6d). When clusters from 0 to 100 m were compared to clusters from the bottom, the Rand index increased over 0-50 m. For both the surface and the bottom clusters, the Rand index was relatively constant with clusters from 50 to 100 m. Clusters from a depth of 60 m were as similar to clusters from the surface as to clusters from the bottom.

Clusters from ecological data at the surface and depth-integrated were similar to each other, but were different to clusters from bottom data (Table 2, Fig. 7). The depth-integrated clusters grouped an area of the Bay of Biscay off the shelf and near the coast of Norway (Fig. 7b, dark pink), which was not found with any other depths. Similarly, a cluster grouping the southern North Sea and English Channel was found at all depths other than depth-integrated. Clusters from depth-averaged data were very similar to clusters from the bottom (Table 2, Fig. 7). The surface and bottom were more different when ecological data was used than when physics or biogeochemical data was used (Table 2).

3.3. *Understanding how underlying correlations in input data drive clustering outcomes*

When clusters from the default set of physics variables were compared with cluster outcomes with one of the variables excluded, the exclusion of mixed layer depth was found to make the biggest difference while omitting salinity made the least difference (Table 3). For biogeochemical variables, the removal of ammonium and DIC were found to make the biggest difference (Table 3). Removing phosphate made the least difference, closely followed by oxygen, silicate and nitrate. For ecological variables, the removal of phytoplankton was found to make the biggest difference, whilst removing zooplankton made the least difference, closely followed by bacteria and dissolved organic carbon (DOC) (Table 3). Removing one variable made less difference with ecological clusterings than with biogeochemical and physics clusterings.

Overall, the mean spatial correlation of a variable was found to be positively correlated to the Rand index of clusterings ($R^2 = 0.34$, $p = 0.02$). This offers support for our hypothesis that more influential variables are the ones with less redundancy in a dataset. Out of the physics variables, temperature and PAR were strongly correlated, but otherwise correlations between variables were weak (Fig. S3a). For biogeochemical variables, nitrate and silicate were strongly correlated, while DIC had the least correlation to other variables (Fig. S3b). For the ecological variables, zooplankton and bacteria were both strongly correlated with several other variables while POC had few, weak correlations (Fig. S3c).

4. Discussion

k-means clustering performed on physical, biogeochemical and ecological variables resulted in sets of clusters which highlighted different features of the system (Fig. 4). For example, results from physics inputs tended to have more divisions of the open ocean, whereas biogeochemical and ecological inputs split the domain into more clusters on the shelf (Figs. 5–7). Differences with depth also varied; the surface was more similar to the bottom with clusters from physical than ecological inputs (Table 1). This suggests that rather than using a ‘one size fits all’ approach to region selection and aiming to redefine the Longhurst provinces (Longhurst, 1995), consideration should first be given to the question being answered. Clustering can help address questions such as how ocean productivity will respond to climate change, where on the Northwest European shelf is more exposed to nutrient inputs from the coast, or which areas follow a similar hydrodynamic regime. Each of these questions will need a different set of variables as input data. In all cases, clustering algorithms offer a technologically accessible method to calculate spatial regions specific to the use case, especially as many clustering algorithms are readily available in packages for popular programming languages such as Python and R (Pedregosa et al., 2011; R Core Team, 2020). However, the sensitivity of results to input choices means that some investment in understanding the clustering method is needed before application.

Clustering allows us to easily investigate how results from the surface are different to other depths. Studies of spatial trends rarely consider

Table 1
Comparison of sets of clusters calculated from different input sets of variables. Rand index (Eq. (5)) was calculated to compare the sets of clusters output from the *k*-means algorithm. A high/low Rand index means the two sets of clusters are similar/different.

	Physics	Biogeochem	Ecology	Physics + Biogeochem	Physics + ecology	Biogeochem + ecology	All
Physics (4)	–	0.73	0.73	0.79	0.83	0.77	0.80
Biogeochem (7)	0.73	–	0.71	0.80	0.77	0.81	0.81
Ecology (5)	0.73	0.71	–	0.70	0.77	0.71	0.74

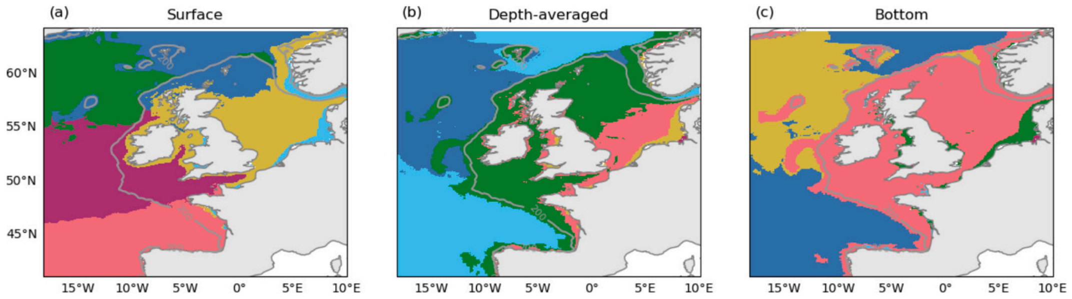

Fig. 5. Clusters from physics data at the surface show latitudinal distribution, while clusters from depth-averaged and bottom data follow bathymetry. Spatial maps of clusters from a) surface layer b) depth-averaged and c) bottom layer. For each depth, six distinct clusters have been identified, shown by different colours. The 200 m depth contour is marked with a grey line.

Table 2
Comparison of sets of clusters calculated from input data at different depths using the Rand index. Clusters were calculated using data from the surface, bottom, depth-averaged and depth integrated where relevant, with sets of physical, biogeochemical and ecological variables. A high/low Rand index means the two sets of clusters are very similar/different.

	Depth-integrated	Depth-averaged	Bottom
Physics			
Surface		0.74	0.72
Depth-averaged		-	0.88
Biogeochemistry			
Surface		0.72	0.71
Depth-averaged		-	0.90
Ecology			
Surface	0.74	0.64	0.62
Depth-integrated	–	0.67	0.65
Depth-averaged	0.67	–	0.94

more than one layer of the ocean, due to constraints on data availability and analysis capacity (Higgs et al., 2024; Leles et al., 2019; Sonnewald et al., 2019). In particular, as satellite data is not collected beyond the top few millimetres of the ocean surface, studies analysing satellite data are restricted to the surface (Kavanaugh et al., 2014; Racault et al., 2015, 2017). Clusters from the surface were found to sharply decrease in similarity to other clusters between 0 and 40 m, below which cluster similarity remained relatively constant (Fig. 6d). This is in line with the top ~40 m of the ocean often being well mixed and therefore similar to the surface (De Boyer Montégut et al., 2004). While clusters might appear to follow bathymetry, especially when depth-averaged or bottom data is used as input (Figs. 5–7), other factors such as the stratification regime, which is closely correlated to bathymetry, may be responsible for driving these patterns (Van Leeuwen et al., 2015).

Variable selection should be carefully considered and account for the research question posed. Biogeochemical variables can be highly correlated spatially, such as nitrate and silicate (Brzezinski, 1985) (Fig. S3b), which could lead to biases in the results. Dissolved inorganic carbon (DIC) did not exhibit any strong correlations with other variables and altered the distribution of clusters more when removed from the variable set than ammonium and silicate, which exhibited strong correlations with each other (Table 3, fig. S3). This suggests that variables which are highly correlated may be less influential in determining clustering outcomes than variables which are less correlated (Zhao et al., 2020). Only relevant variables should be included, as if variables irrelevant to the analysis are included, the resulting clusters may not follow the pertinent features of the system. In some circumstances, additional preprocessing may be appropriate, for example when one variable is particularly important, such as in systems prone to hypoxia, where organisms will be very sensitive to a small change in oxygen (Galli et al., 2024; Rubalcaba et al., 2020). This additional knowledge can be used to weight variables appropriately, for example by using temperature transformed to a biologically relevant temperature index as an input (Hobday et al., 2018; Wilson et al., 2024).

4.1. Considerations and caveats

Some patterns in the clustering output initially appear surprising, e.g. unsupervised clustering of surface ecological data grouped the northern North Sea and Bay of Biscay into the same cluster (Fig. 4e,f). These two regions appear superficially similar based on the data provided to the clustering algorithm, as both regions have low phytoplankton and zooplankton biomasses. However, in reality, due to geographical separation, the ecosystems will be formed of different species inhabiting different climates and are unlikely to behave similarly. Therefore, clusters must not be overinterpreted when they are based on limited input data; additional input data such as temperature (Fig. 4a,b) could help in differentiating the two regions.

Furthermore, surface and depth-integrated ecological clusters both identified a region around the boundary of the domain (Fig. 7a,b), which was not found in any of the other analyses. This cluster is likely due to the low biomass boundary condition used for phytoplankton and suggests that clustering algorithms may highlight areas where the model or observations exhibit bias. The cluster is not present at all depths and makes it hard to compare the similarity of clusters between variable sets (Table 1). However, this demonstrates that clustering can be used to identify the extent of the domain influenced by boundary conditions, which could then be excluded from further analysis.

While the consistency of clustering outcomes between a range of inputs has been investigated here, the combinations explored are far from exhaustive. Choices were made to explore potentially interesting differences between inputs, such as between the surface and bottom layers. The partitioning of variables into sets of physical, biogeochemical and ecological inputs is not definitive and some variables could be included in more than one set, such as dissolved organic carbon (DOC) which could fit within the sets of both biogeochemical and ecological variables.

Model data was used for this study as there is no set of observations sufficiently complete for this analysis. Clustering is useful for model data analysis, but also has applications to observational data, for example, grouping sparse observations from areas with similar behaviour (McGinty et al., 2023) or using gridded observational data (Sun et al., 2021). Clustering has the potential to inform researchers where to focus sampling effort, e.g. sites spanning different behaviour or regions that are more sparsely sampled (Cazenave et al., 2021). Many of the conclusions drawn from model data will be relevant to forming clusters from observational datasets, such as the need to consider which variables and depths are used as input and the similarities and differences between clusters from physical, biogeochemical and ecological variables.

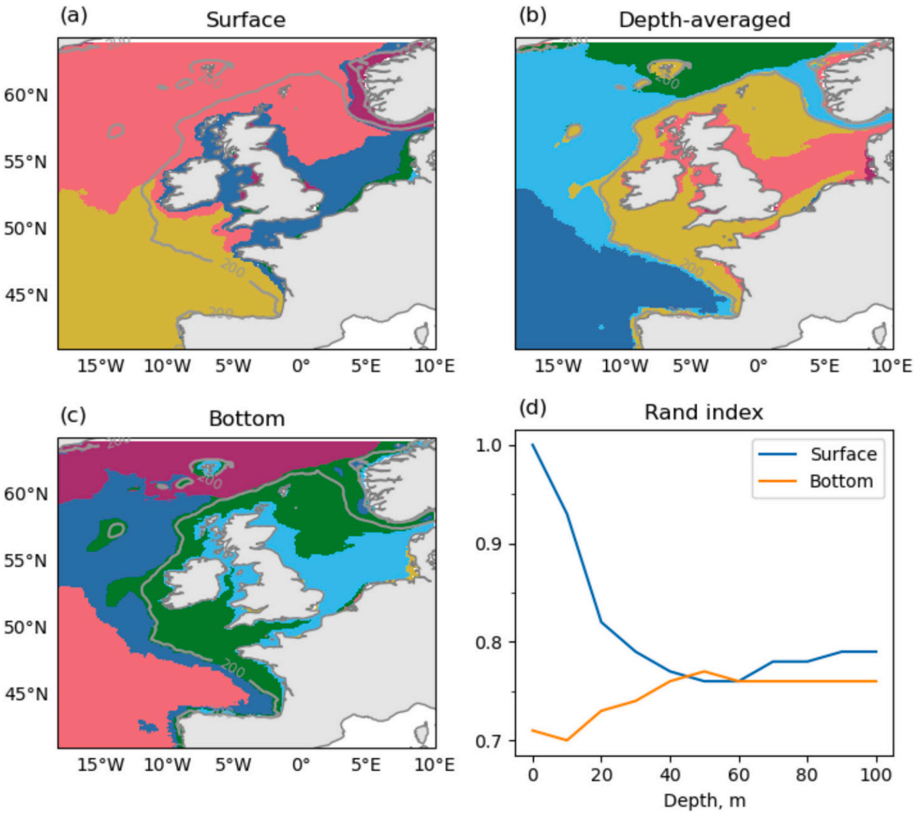

Fig. 6. Clusters from biogeochemical data at different depths. Spatial map of clusters from a) surface layer b) depth-averaged c) bottom layer. For each depth, six distinct clusters have been identified, shown by different colours. The 200 m depth contour is marked with a grey line. d) The Rand index was used to compare the surface and bottom to depths between 0 and 100 m, at 10 m intervals.

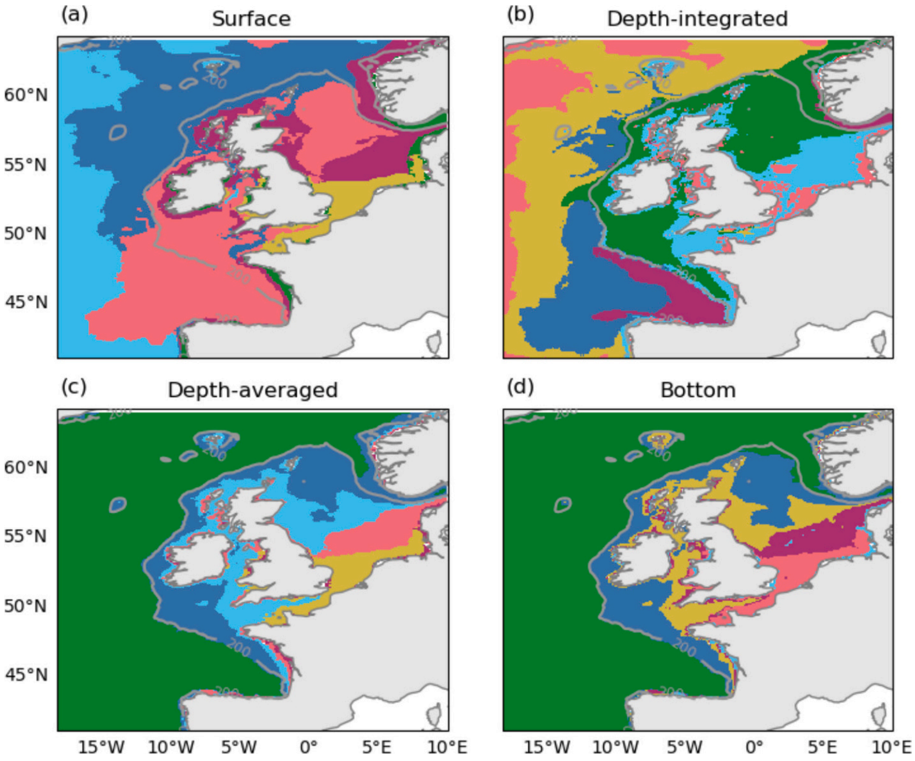

Fig. 7. Clusters from ecological data at different depths. Spatial map of clusters from a) surface layer b) depth-integrated, c) depth-averaged and d) bottom layer. For each depth, six distinct clusters have been identified, shown by different colours. The 200 m depth contour is marked with a grey line.

Table 3
Comparison of sets of clusters calculated from input datasets with all variables and with one variable removed. Bold headings show the variable set used and column headings show the variable which has been removed. Rand index (Eq. (5)) was used to compare the sets of clusters. A lower Rand index means the clusters computed without this variable are more different to the clusters computed with this variable, suggesting that the variable is important in determining the clusters.

Physics variables							
Variable removed	Temperature	Salinity	Mixed layer depth	PAR			
Rand index	0.83	0.87	0.81	0.84			
Biogeochemical variables							
Variable removed	Phosphate	Nitrate	Oxygen	DIC	Alk.	Ammon.	Silicate
Rand index	0.89	0.87	0.88	0.83	0.85	0.83	0.88
Ecological variables							
Variable removed	Phytoplank.	Zooplank.	DOC	POC	Bacteria		
Rand index	0.86	0.94	0.92	0.90	0.93		

4.2. Future work

In our work, we have focused specifically on applying the *k*-means algorithm (Cam and Neyman, 1967), but other unsupervised clustering algorithms may be better suited to certain applications, such as where an emergent number of clusters is needed to follow features in the data (Ester et al., 1996; Kohonen, 1982). Furthermore, the *k*-means algorithm can be applied with different distance metrics, e.g. cosine similarity rather than Euclidean (Zhao et al., 2020), which may be more appropriate for some data distributions. Differences in algorithm structure or metric choice may result in different sensitivities to those presented in this study, and should be investigated.

Inclusion of temporal variation in the inputs to clustering algorithms requires further consideration. The input data used in this study was the mean of a five-year time-series, but for many applications, temporal variation, both seasonally and on longer time scales, will be important (Racault et al., 2012, 2014; Wakelin et al., 2020). Some algorithms are designed to analyse time-series, such as K-shape (Paparrizos and Gravano, 2015), which allows regions with a similar seasonal cycle to be grouped by accounting for temporal offsets in time-series and reducing the importance of variable magnitude. Some studies have included the seasonal cycle in their input data (Jarníková et al., 2022; Lessin et al., 2020; McGinty et al., 2023), but often at the cost of the number of variables considered. We suggest that clusters from inputs including the seasonal cycle could be compared to clusters using summary metrics as inputs (e.g. maximum, minimum, standard deviation), to investigate whether summary information on temporal variation input into the clustering algorithm would be sufficient without providing the entire time-series.

Further applications of clustering include facilitating the analysis of long-term temporal trends. For example, a time-series of model predictions can be averaged spatially over a cluster to investigate long-term trends due to climate change. Alternatively, cluster definitions calculated from present day model output can be applied to model predictions to project the spatial extent of the clusters in the future and therefore examine the spatial evolution of similar environments or ecosystems, highlighting potential regime shifts.

Unsupervised clustering has significant potential as a tool to inform marine spatial planning and ecosystem-based management, enabling deeper understanding of spatial patterns and insights into the impacts of interventions, such as designation of a new marine protected area or deployment of an offshore wind farm, by comparing the distribution of clusters before and after the intervention takes place (Azzellino et al., 2013). As ever, the clustering approach should be guided by a clear set of questions, and the factors driving cluster selection must be carefully investigated, which will ensure adoption of optimal area definitions for informed decision-making.

5. Conclusion

Unsupervised clustering provides a useful tool for oceanic data analysis. Here we have shown how clustering can provide bespoke spatial regions depending on the inputs chosen. The distribution of spatial regions depended on the variables and depths included in the input data. For example, clusters from physics variables at the surface were distributed latitudinally, while data inputs from the bottom or depth-averaged tended to lead to clusters following the shelf-edge. Careful consideration of input choices in the context of the research question will lead to the most insightful outcomes, while poor choices of inputs may provide insufficient information and spurious clustering outcomes, such as clusters purely defined by model boundary conditions or clusters grouping empirically unrelated ecosystems.

To aid the uptake of clustering by researchers, we have made code, data and analysis scripts openly available at Doi: https://doi.org/10.5281/zenodo.17227522.

Authors contribution

Conceptualization: JB, DP, RM. Analysis: RM, DP. Visualization: RM, DP. Writing – original draft: RM. Writing – review and editing: all authors.

CRediT authorship contribution statement

Rebecca Millington: Writing – review & editing, Writing – original draft, Visualization, Formal analysis. **Dale Partridge:** Writing – review & editing, Visualization, Formal analysis, Conceptualization. **Helen R. Powley:** Writing – review & editing. **Gennadi Lessin:** Writing – review & editing. **David Moffat:** Writing – review & editing. **Jerry Blackford:** Writing – review & editing, Conceptualization.

Funding sources

This work was funded by the UK National Capability Project FOCUS NE/X006271/1, and the NERC Single Centre National Capability Programme Atlantic Climate and Environment Strategic Science (AtlantiS) NE/Y005589/1Atlantis. This work benefitted from internal funding at Plymouth Marine Laboratory.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the NERC Earth Observation Data Acquisition and

Analysis Service (NEODAAS) for access to compute resources for this study. The authors would like to thank the three anonymous reviewers for their feedback which led to improvements in the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103586.

Data availability

All data and the code used for data processing, analysis and visualization is available at Doi: 10.5281/zenodo.17227522.

References

Atwood, E.C., Jackson, T., Laurenson, A., Jönsson, B.F., Spyrakos, E., Jiang, D., Sent, G., Selmes, N., Simis, S., Danne, O., Tyler, A., Groom, S., 2024. Framework for regional to global extension of optical water types for remote sensing of optically complex transitional water bodies. Remote Sens. 16 (17), 3267. https://doi.org/10.3390/rs16173267.

Azzellino, A., Ferrante, V., Kofoed, J.P., Lanfredi, C., Vicinanza, D., 2013. Optimal siting of offshore wind-power combined with wave energy through a marine spatial planning approach. Int. J. Mar. Energy. 3–4, e11–e25. https://doi.org/10.1016/j.ijome.2013.11.008.

Bauer, J.E., Cai, W.-J., Raymond, P.A., Bianchi, T.S., Hopkinson, C.S., Regnier, P.A.G., 2013. The changing carbon cycle of the coastal ocean. Nature 504 (7478), 61–70. https://doi.org/10.1038/nature12857.

Bruggeman, J., Bolding, K., 2014. A general framework for aquatic biogeochemical models. Environ. Model Softw. 61, 249–265. https://doi.org/10.1016/j.envsoft.2014.04.002.

Brzezinski, M.A., 1985. The Si:C:N ratio of marine diatoms: interspecific variability and the effect of some environmental variables. J. Phycol. 21 (3), 347–357. https://doi.org/10.1111/j.0022-3646.1985.00347.x.

Butenschön, M., Clark, J., Aldridge, J.N., Allen, J.I., Artioli, Y., Blackford, J., Bruggeman, J., Cazenave, P., Ciavatta, S., Kay, S., Lessin, G., van Leeuwen, S., van der Molen, J., De Mora, L., Polimene, L., Sailley, S., Stephens, N., Torres, R., 2016. ERSEM 15.06: A generic model for marine biogeochemistry and the ecosystem dynamics of the lower trophic levels. Geosci. Model Dev. 9 (4), 1293–1339. https://doi.org/10.5194/gmd-9-1293-2016.

Cam, L.M.L., Neyman, J., 1967. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press.

Cavan, E.L., Hill, S.L., 2022. Commercial fishery disturbance of the global ocean biological carbon sink. Glob. Chang. Biol. 28 (4), 1212–1221. https://doi.org/10.1111/gcb.16019.

Cazenave, P.W., Dewar, M., Torres, R., Blackford, J., Bedington, M., Artioli, Y., Bruggeman, J., 2021. Optimising environmental monitoring for carbon dioxide sequestered offshore. Int. J. Greenh. Gas Con. 110, 103397. https://doi.org/10.1016/j.ijggc.2021.103397.

Cheung, W.W.L., Jones, M.C., Lam, V.W.Y., Miller, D., Ota, Y., Teh, L., Sumaila, U.R., 2017. Transform high seas management to build climate resilience in marine seafood supply. Fish Fish. 18 (2), 254–263. https://doi.org/10.1111/faf.12177.

Daewel, U., Schrum, C., Macdonald, J.I., 2019. Towards end-to-end (E2E) modelling in a consistent NPZD-F modelling framework (ECOSMO E2E_v1.0): application to the North Sea and Baltic Sea. Geosci. Model Dev. 12 (5), 1765–1789. https://doi.org/10.5194/gmd-12-1765-2019.

De Boyer Montégut, C., Madec, G., Fischer, A.S., Lazar, A., Iudicone, D., 2004. Mixed layer depth over the global ocean: an examination of profile data and a profile-based climatology. J. Geophys. Res. Oceans 109 (C12), 2004JC002378. https://doi.org/10.1029/2004JC002378.

Edwards, K.P., Barciela, R., Butenschön, M., 2012. Validation of the NEMO-ERSEM operational ecosystem model for the north west European continental shelf. Ocean Sci. 8 (6), 983–1000. https://doi.org/10.5194/os-8-983-2012.

Eigaard, O.R., Bastardie, F., Breen, M., Dinesen, G.E., Hintzen, N.T., Laffargue, P., Mortensen, L.O., Nielsen, J.R., Nilsson, H.C., O'Neill, F.G., Polet, H., Reid, D.G., Sala, A., Sköld, M., Smith, C., Sørensen, T.K., Tully, O., Zengin, M., Rijnsdorp, A.D., 2016. Estimating seabed pressure from demersal trawls, seines, and dredges based on gear design and dimensions. ICES J. Mar. Sci. 73 (suppl_1), i27–i43. https://doi.org/10.1093/icesjms/fsv099.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Knowledge Discovery and Data Mining, vol. 96. AAAI Press, pp. 226–231. https://api.semanticscholar.org/CorpusID:355163.

Galli, G., Wakelin, S., Harle, J., Holt, J., Artioli, Y., 2024. Multi-model comparison of trends and controls of near-bed oxygen concentration on the northwest European continental shelf under climate change. Biogeosciences 21 (8), 2143–2158. https://doi.org/10.5194/bg-21-2143-2024.

Harris, P.T., Macmillan-Lawler, M., Rupp, J., Baker, E.K., 2014. Geomorphology of the oceans. Mar. Geol. 352, 4–24. https://doi.org/10.1016/j.margeo.2014.01.011.

Higgs, I., Skákala, J., Bannister, R., Carrassi, A., Ciavatta, S., 2024. Investigating ecosystem connections in the shelf sea environment using complex networks. Biogeosciences 21 (3), 731–746. https://doi.org/10.5194/bg-21-731-2024.

Hobday, A., Oliver, E., Sen Gupta, A., Benthuysen, J., Burrows, M., Donat, M., Holbrook, N., Moore, P., Thomsen, M., Wernberg, T., Smale, D., 2018. Categorizing and naming marine heatwaves. Oceanography 31 (2). https://doi.org/10.5670/oceanog.2018.205.

Holt, J., Proctor, R., 2008. The seasonal circulation and volume transport on the northwest European continental shelf: a fine-resolution model study. J. Geophys. Res. Oceans 113 (C6), 2006JC004034. https://doi.org/10.1029/2006JC004034.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J., 2023. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Inf. Sci. 622, 178–210. https://doi.org/10.1016/j.ins.2022.11.139.

Jarníková, T., Olson, E.M., Allen, S.E., Ianson, D., Suchy, K.D., 2022. A clustering approach to determine biophysical provinces and physical drivers of productivity dynamics in a complex coastal sea. Ocean Sci. 18 (5), 1451–1475. https://doi.org/10.5194/os-18-1451-2022.

Kavanaugh, M.T., Hales, B., Saraceno, M., Spitz, Y.H., White, A.E., Letelier, R.M., 2014. Hierarchical and dynamic seascapes: a quantitative framework for scaling pelagic biogeochemistry and ecology. Prog. Oceanogr. 120, 291–304. https://doi.org/10.1016/j.pocean.2013.10.013.

Kerby, T.K., Cheung, W.W.L., Engelhard, G.H., 2012. The United Kingdom's role in North Sea demersal fisheries: a hundred year perspective. Rev. Fish Biol. Fish. 22 (3), 621–634. https://doi.org/10.1007/s11160-012-9261-y.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43 (1), 59–69. https://doi.org/10.1007/BF00337288.

Legge, O., Johnson, M., Hicks, N., Jickells, T., Diesing, M., Aldridge, J., Andrews, J., Artioli, Y., Bakker, D.C.E., Burrows, M.T., Carr, N., Cripps, G., Felgate, S.L., Fernand, L., Greenwood, N., Hartman, S., Kröger, S., Lessin, G., Mahaffey, C., Williamson, P., 2020. Carbon on the northwest European shelf: contemporary budget and future influences. Front. Mar. Sci. 7. https://doi.org/10.3389/fmars.2020.00143.

Leles, S.G., Mitra, A., Flynn, K.J., Tillmann, U., Stoecker, D., Jeong, H.J., Burkholder, J., Hansen, P.J., Caron, D.A., Glibert, P.M., Hallegraeff, G., Raven, J.A., Sanders, R.W., Zubkov, M., 2019. Sampling bias misrepresents the biogeographical significance of constitutive mixotrophs across global oceans. Glob. Ecol. Biogeogr. 28 (4), 418–428. https://doi.org/10.1111/geb.12853.

Lessin, G., Polimene, L., Artioli, Y., Butenschön, M., Clark, D.R., Brown, I., Rees, A.P., 2020. Modeling the seasonality and controls of nitrous oxide emissions on the northwest European continental shelf. J. Geophys. Res. Biogeosci. 125 (6), e2019JG005613. https://doi.org/10.1029/2019JG005613.

Longhurst, A., 1995. Seasonal cycles of pelagic production and consumption. Prog. Oceanogr. 36 (2), 77–167. https://doi.org/10.1016/0079-6611(95)00015-1.

Madec, G., Bourdallé-Badie, R., Bouttier, P.-A., Bricaud, C., Bruciaferri, D., Calvert, D., Chanut, J., Clementi, E., Coward, A., Delrosso, D., Ethé, C., Flavoni, S., Graham, T., Harle, J., Iovino, D., Lea, D., Lévy, C., Lovato, T., Martin, N., Vancoppenolle, M., 2017. NEMO Ocean Engine. https://www.earth-prints.org/handle/2122/13309.

McGinty, N., Irwin, A.J., Finkel, Z.V., Dutkiewicz, S., 2023. Using ecological partitions to assess zooplankton biogeography and seasonality. Front. Mar. Sci. 10, 989770. https://doi.org/10.3389/fmars.2023.989770.

Miller, P.I., Christodoulou, S., 2014. Frequent locations of oceanic fronts as an indicator of pelagic diversity: application to marine protected areas and renewables. Mar. Policy 45, 318–329. https://doi.org/10.1016/j.marpol.2013.09.009.

Nguyen, L.H., Holmes, S., 2019. Ten quick tips for effective dimensionality reduction. PLoS Comput. Biol. 15 (6), e1006907. https://doi.org/10.1371/journal.pcbi.1006907.

Paparrizos, J., Gravano, L., 2015. k-Shape: efficient and accurate clustering of time series. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1855–1870. https://doi.org/10.1145/2723372.2737793.

Pauly, D., Christensen, V., Guénette, S., Pitcher, T.J., Sumaila, U.R., Walters, C.J., Watson, R., Zeller, D., 2002. Towards sustainability in world fisheries. Nature 418 (6898), 689–695. https://doi.org/10.1038/nature01017.

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., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Powley, H.R., Polimene, L., Torres, R., Al Azhar, M., Bell, V., Cooper, D., Holt, J., Wakelin, S., Artioli, Y., 2024. Modelling terrigenous DOC across the north west European shelf: fate of riverine input and impact on air-sea CO2 fluxes. Sci. Total Environ. 912, 168938. https://doi.org/10.1016/j.scitotenv.2023.168938.

R Core Team, 2020. R: A Language and Environment for Statistical Computing [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Racault, M.-F., Le Quéré, C., Buitenhuis, E., Sathyendranath, S., Platt, T., 2012. Phytoplankton phenology in the global ocean. Ecol. Indic. 14 (1), 152–163. https://doi.org/10.1016/j.ecolind.2011.07.010.

Racault, M.-F., Platt, T., Sathyendranath, S., Airba, E., Martinez Vicente, V., Brewin, R., 2014. Plankton indicators and ocean observing systems: support to the marine ecosystem state assessment. J. Plankton Res. 36 (3), 621–629. https://doi.org/10.1093/plankt/fbu016.

Racault, M.-F., Raitsos, D.E., Berumen, M.L., Brewin, R.J.W., Platt, T., Sathyendranath, S., Hoteit, I., 2015. Phytoplankton phenology indices in coral reef ecosystems: application to ocean-color observations in the Red Sea. Remote Sens. Environ. 160, 222–234. https://doi.org/10.1016/j.rse.2015.01.019.

Racault, M.-F., Sathyendranath, S., Menon, N., Platt, T., 2017. Phenological responses to ENSO in the global oceans. Surv. Geophys. 38 (1), 277–293. https://doi.org/10.1007/s10712-016-9391-1.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66 (336), 846–850. https://doi.org/10.1080/01621459.1971.10482356.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Rubalcaba, J.G., Verberk, W.C.E.P., Hendriks, A.J., Saris, B., Woods, H.A., 2020. Oxygen limitation may affect the temperature and size dependence of metabolism in aquatic ectotherms. Proc. Natl. Acad. Sci. 117 (50), 31963–31968. https://doi.org/10.1073/pnas.2003292117.

Sonnewald, M., Wunsch, C., Heimbach, P., 2019. Unsupervised learning reveals geography of Global Ocean dynamical regions. Earth Space Sci. 6 (5), 784–794. https://doi.org/10.1029/2018EA000519.

Sonnewald, M., Dutkiewicz, S., Hill, C., Forget, G., 2020. Elucidating ecological complexity: unsupervised learning determines global marine eco-provinces. Sci. Adv. 6 (22), eaay4740. https://doi.org/10.1126/sciadv.aay4740.

Sun, Q., Little, C.M., Barthel, A.M., Padman, L., 2021. A clustering-based approach to ocean model–data comparison around Antarctica. Ocean Sci. 17 (1), 131–145. https://doi.org/10.5194/os-17-131-2021.

Van Leeuwen, S., Tett, P., Mills, D., Van Der Molen, J., 2015. Stratified and nonstratified areas in the North Sea: long-term variability and biological and policy implications. J. Geophys. Res. Oceans 120 (7), 4670–4686. https://doi.org/10.1002/2014JC010485.

van Oostende, M., Hieronymi, M., Krasemann, H., Baschek, B., 2023. Global Ocean colour trends in biogeochemical provinces. Front. Mar. Sci. 10. https://doi.org/10.3389/fmars.2023.1052166.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Vázquez-Baeza, Y., 2020. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods 17 (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Wakelin, S.L., Artioli, Y., Holt, J.T., Butenschön, M., Blackford, J., 2020. Controls on near-bed oxygen concentration on the northwest European continental shelf under a potential future climate scenario. Prog. Oceanogr. 187, 102400. https://doi.org/10.1016/j.pocean.2020.102400.

Wilson, R.J., Artioli, Y., 2023. Nctoolkit: a python package for netCDF analysis andpost-processing. J. Open Source Softw. 8 (88), 5494. https://doi.org/10.21105/joss.05494.

Wilson, R.J., Kay, S., Ciavatta, S., 2024. Partitioning climate uncertainty in ecological projections: Pacific oysters in a hotter Europe. Eco. Inform. 80, 102537. https://doi.org/10.1016/j.ecoinf.2024.102537.

Zhao, Q., Basher, Z., Costello, M.J., 2020. Mapping near surface global marine ecosystems through cluster analysis of environmental data. Ecol. Res. 35 (2), 327–342. https://doi.org/10.1111/1440-1703.12060.