ICES Journal of Marine Science, 2026, Vol. 83, Issue 1, fsaf227
https://doi.org/10.1093/icesjms/fsaf227
Received: 11 August 2025; **revised:** 1 December 2025; **accepted:** 3 December 2025
Original Article

Big data approaches reveal large-scale spatial patterns in marine epifauna

Keith M. Cooper*, Matthew Curtis, Anna-Lena Downie, Stefan G. Bolam

Centre for Environment, Fisheries and Aquaculture Science (CEFAS), Lowestoft Laboratory, Lowestoft, Suffolk, United Kingdom

*Corresponding author. entre for Environment, Fisheries and Aquaculture Science (CEFAS), Lowestoft Laboratory, Lowestoft, Suffolk, NR33 0HT, United Kingdom. E-mail: keith.cooper@cefas.gov.uk

Abstract

Comprehensive maps of biological characteristics are increasingly employed to support the management of human activities in marine environments. However, their development is often constrained by insufficient data coverage across broad spatial scales. Here, we apply a big data approach by integrating marine epifaunal data from 2 m beam trawls across the UK shelf and wider North Sea to enhance understanding of the ecological characteristics of epifaunal assemblages in sedimentary habitats. We analyse spatial patterns in univariate metrics (taxon richness, total abundance) and assemblage taxonomic structure. Taxon richness was found to peak in the northern North Sea and English Channel, while abundance hotspots occurred along the Norwegian coast, southern North Sea, and inshore regions around the UK and Ireland. We also identify 11 distinct assemblage types, each exhibiting their own taxonomic epifaunal composition, with strong biogeographic structure. We identify the main environmental drivers shaping these patterns and discuss how the creation of such maps, and the added insight gained from them, may be used to facilitate the management of anthropogenic activities.

Keywords: epifauna; 2 m beam trawl; clustering; random forest; spatial maps

Introduction

Marine epifauna—such as echinoderms (e.g. brittle stars *Ophiothrix* spp.), crustaceans (e.g. brown shrimp *Crangon crangon*), molluscs (e.g. scallops *Pecten maximus*), cnidarians (e.g. hydroids), and demersal fish (e.g. plaice *Pleuronectes platessa*)—play an important role in ecosystem functioning—facilitating nutrient cycling, providing nursery areas, and modifying sediment biogeochemistry (Murillo et al. 2020), and linking benthic and pelagic systems by providing biomass for large predators (Griffiths et al. 2017; Chen et al. 2021). Their often-high abundances and rapid turnover rates support the flow of energy through marine food webs (Newcombe & Taylor 2010). As essential trophic links between benthic primary producers and higher-order consumers such as carnivorous invertebrates and commercially important fish species, epifaunal communities contribute substantially to secondary production, representing up to 75% of the total annual production in some habitats (Taylor 1998; Kramer et al. 2015).

Interest in the distribution, structure and diversity of epifaunal communities is long-standing (Rees et al. 2007). Their distributions having been described in the North Sea (Jennings et al. 1999; Zühlke et al. 2001; Callaway et al. 2002), the English Channel (Kaiser et al. 1999), and Irish Sea and Bristol Channel (Ellis et al. 2002). However, whilst integrated assessment has been conducted for the infauna (Cooper and Barry 2017; Bolam et al., 2023; Cooper et al., in press), there is no single study that describes epifaunal assemblage patterns across the UK continental shelf and adjacent North Sea. Understanding spatial distributions and variability at such scales requires large, integrated empirical datasets. Historically, efforts to develop reliable maps biologically characterising the seabed at these scales have been constrained by limited data access and insufficient IT infrastructure (Stelzenmuller et al. 2015). To address these limitations, we apply a big data approach—integrating large disparate datasets from multiple sources and time periods, harmonised via metadata filtering and standardised taxonomy—enabling robust, large-scale mapping of epifaunal communities. Here, we use "big data" in the broader sense of large and complex datasets, rather than implying continuous real-time data streams (Guidi et al. 2020).

For the infauna, Cooper and Barry (2017) showed that integrating seabed sampling data from governmental and non-governmental sources (e.g. marine aggregates, offshore wind, oil and gas) (*OneBenthic* https://rconnect.cefas.co.uk/onebenthic_portal/) could produce a robust biological baseline. In a management context, the availability of such maps (Cooper et al. 2019; O'Brien et al. 2022) is appealing as differences in the component taxa between different assemblages will undoubtedly manifest through differences in sensitivity to, and functional responses from, anthropogenic pressures (Bolam et al. 2023). Understanding these patterns and their environmental drivers is therefore an important prerequisite for effective conservation and spatial ecosystem-based management (Reiss et al. 2010).

In this paper, we demonstrate that empirical epifaunal abundance data obtained from 2 m beam trawls can be integrated using a large-scale data approach to improve understanding of spatial patterns in community structure metrics (e.g. taxon richness, abundance) and taxonomic composition (i.e. assemblage biotopes) across the UK shelf and wider North Sea. We use a random forest modelling framework to predict these patterns based on environmental conditions and produce continuous maps of their variation. We discuss the implications of these patterns and drivers for informing man-

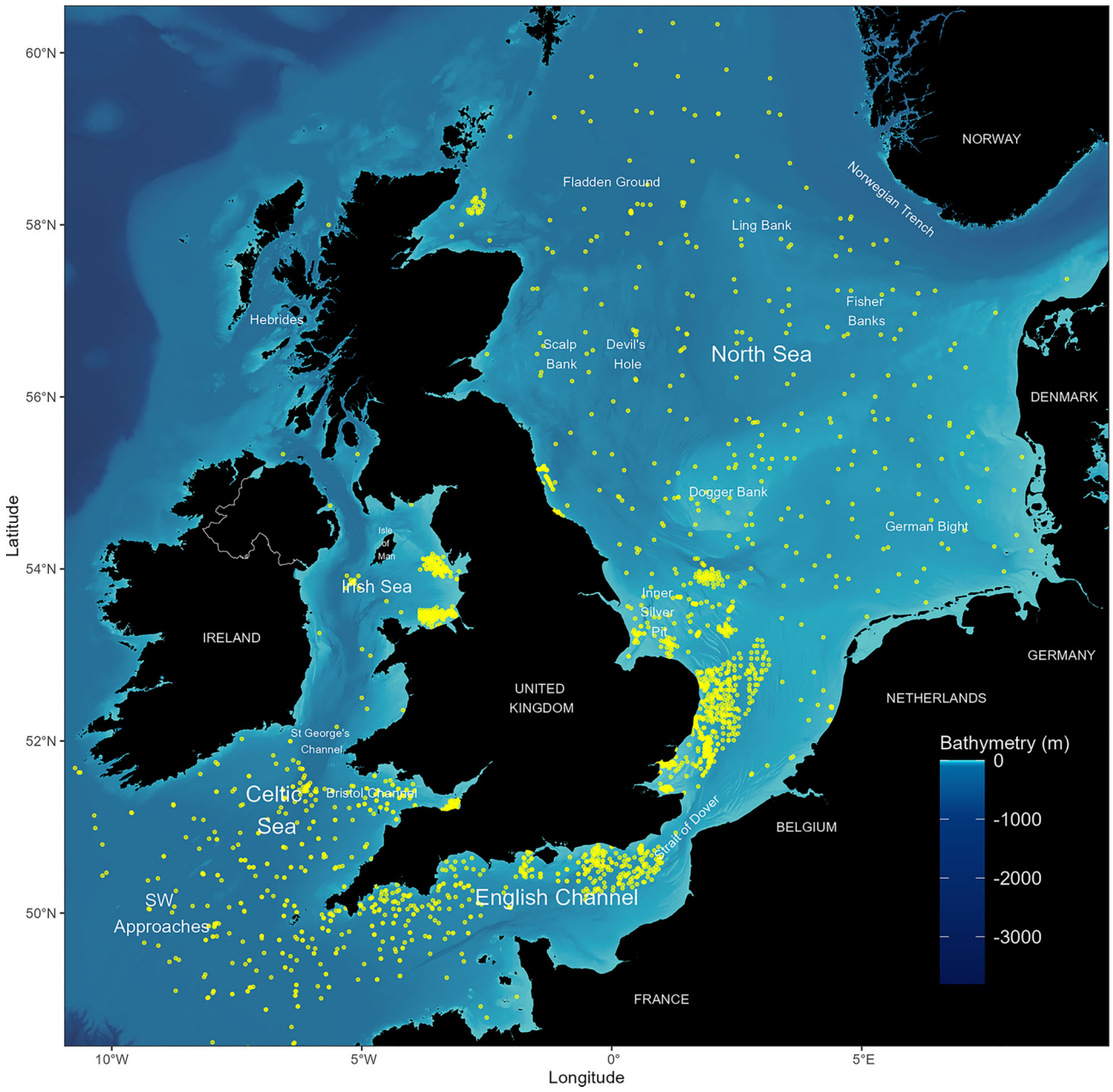

Figure 1. Locations of all 2 m beam trawl samples contained in *OneBenthic*. Background bathymetry from GEBCO Grid (The GEBCO_2023 Grid | GEBCO).

agement decisions related to licencing and mitigating anthropogenic activities that affect epifaunal assemblages.

Methods

The dataset

The empirical epifaunal data used in this study were sourced from the *OneBenthic* trawl database (https://rconnect.cefas.co.uk/onebenthic_portal/). *OneBenthic* brings together publicly available disparate benthic datasets (biological abundance/biomass and sediment particle size) in a cloud-based PostgreSQL database. The epifaunal dataset accessed on 26th March 2025 includes 3 799 trawl samples from 110 surveys collected over 36-years (1987–2023). Its spatial extent covers UK shelf waters and regions of northeast Atlantic countries including France, Belgium, The Netherlands, Germany, Denmark, and Norway (Fig. 1).

The *OneBenthic* database uses taxonomic information from the World Register of Marine Species (WoRMS Editorial Board 2025) enabling outputs with standardised nomenclature. WoRMS data were accessed via in R (R Core Team 2025) using the *worms package* (Chamberlain & Ooms, 2023), with each taxon uniquely identified by the aphiaID field.

An essential step for integrating data from disparate sources is to review associated metadata—such as sampling gear, mesh size, tow speed and length, and sample location—to refine the dataset and enhance comparability. From the full dataset of 3 799 samples, we selected a subset of 2 383 meeting the following criteria: sampled using a 2 m beam trawl fitted with either a 4 mm or 5 mm cod end mesh

(Supplementary Fig. S1), towing speed less than 2 knots, tow length less than 2 km, and sample location outside areas licensed for anthropogenic activities (e.g. dredging/extraction, disposal, renewables). This last criterion was applied to minimise the influence of recent or ongoing human activities that could alter epifaunal community structure, ensuring that the dataset more accurately reflects natural spatial patterns and baseline conditions. We recognise that demersal fishing is widespread across the region and acknowledge that excluding all areas exposed to fishing pressure was not feasible given its pervasive extent; instead, we aimed to exclude only those samples clearly flagged as potentially not representative of typical benthic conditions for their habitat type (i.e. those from within licensed areas). As sampling occurred year-round (Supplementary Fig. S1), no seasonal restriction was applied.

Raw taxon records were reviewed to exclude pelagic species, taxa better sampled with grabs (e.g. polychaetes), burrowing species, high level entries (e.g. Pisces), and juvenile or larval stages. These exclusions were made to ensure that the dataset reflected taxa that are reliably and consistently sampled by 2 m beam trawl, and to avoid including groups whose presence or abundance would not be accurately represented by this gear type. Demersal fish species, and those feeding on demersal prey, were retained because they are effectively sampled by beam trawl and form an integral part of the epifaunal community structure in these habitats. Colonial taxa which are recorded as "present" (e.g. hydroids, sponges) were included and assigned abundance of 1 (in accordance with how they are treated for grab samples (McIlwaine et al. 2025)). Taxa were aggregated to the family level to address identification and/or sample processing inconsistencies which inherently occur with combining disparate data sources (Cooper and Barry 2017).

Prior to analysis, spatial autocorrelation was assessed using the R package *emon* (Barry et al. 2017), which calculates empirical semi-variograms and fits Gaussian models by least squares (Cressie 1993) (Supplementary Fig. S2). Significant spatial autocorrelation was detected between samples less than 2 km apart, leading to the removal of 983 samples (Supplementary Fig. S3). The final dataset used for spatial analyses contained 1 400 samples.

Community metrics

Prior to the deriving community metrics, raw epifaunal abundance data were assessed for outliers using modified z-scores (Saleem et al. 2021). Based on the N-sqrt criterion, whereby N-sqrt is the square root of total abundance of each sample, 106 samples were flagged as potential outliers, typically due to extremely high counts of certain taxa (e.g. *Crangon* spp., *Ophiothrix* spp.). To address this, a cap of 1 000 individuals per taxon per sample was applied. Following outlier handling, two univariate metrics of community structure were derived for each sample: Taxon richness, as described by the total number of taxa (S) and the square root of the total abundance of individuals (N). These metrics were selected because they are commonly used, easily interpretable, and provide a meaningful context for interpreting the resulting multivariate spatial patterns in the benthic assemblage sampled using the 2 m beam trawl (see Section 2.3).

Multivariate taxonomic structure

Epifaunal abundance data were used to identify the spatial distribution of discrete faunal groups based on multivariate taxonomic structure, following the approach of Cooper and Barry (2017). Square-root transformed abundance data were clustered using the using the k-means algorithm (MacQueen, 1967) implemented via the R function *kmeans*. This method partitions samples by minimising the within-cluster sum of squares across all variables. K-means clustering was selected for its suitability for large datasets and its successful application in comparable infaunal assessments (Cooper and Barry 2017; Bolam et al. 2023; Cooper et al. (in press)). The number of cluster groups was determined using an elbow plot to balance biological detail and broader spatial patterns. The final number of clusters is generally chosen based on where along the profile of the plot the gradient starts to decrease.

To quantify relative similarity among the resulting faunal groups, absolute distances between K-means cluster centres were calculated across all variables using the R function *dist*. The resulting dissimilarity matrix was used to generate a dendrogram via group-average hierarchical clustering (R function *hclust*) and visually represent the observed difference between k-means cluster groups. Each group was then assigned a code (and colour) based on the dendrogram structure to reflect relatedness.

To describe the biological characteristics of each epifaunal group, we examined both the cluster centres and the results of a SIMPER analysis (Primer v7®; Clarke and Gorley 2015) to determine the taxa which best characterised each group. Additionally, mean univariate measures of taxon richness (S) and total abundance (N) of each group, along with the proportions of taxa by major phyla, were calculated.

Modelling

Environmental predictors

A range of raster layers representing environmental variables known to influence epifaunal distributions were assembled for use as model predictors (Table 1). Data were sourced from Bio-ORACLE (https://www.bio-oracle.org/; Assis et al. 2018; Tyberghein et al. 2012) and Mitchell et al. (2019). Bio-ORACLE layers were downloaded using the Download Manager with the following settings: Dataset version: Bio-ORACLE v3; Period of layers = Present-day conditions (2000–2010); Depth of layers = Benthic layers; Layers = Mean and Range.

From Mitchell et al. (2019), available data products included sediment composition (% Mud, % Sand, % Gravel; https://doi.org/10.14466/CefasDataHub.63), and additional environmental predictors such as water depth, wave velocity, current speed, and suspended inorganic particulate matter (SPM—summer, winter and mean); https://doi.org/10.14466/CefasDataHub.62.

Five additional environmental layers representing seafloor topography were derived from the bathymetry layer (water depth) using SAGA GIS tools for QGIS (v.3.2; Conrad et al. 2015). These included variables such as topographic slope length and steepness (LS-Factor) and Relative Slope Position (RSP, Böhner & Selige 2006). The LS-Factor combines slope gradient over slope length to estimate erosion potential (Desmet & Govers, 1996) and is analogously applied in the marine context to reflect the potential stability of sediment deposits and the likelihood of exposed hard substrata. The RSP

Table 1. Raster variables used in modelling and for explaining patterns in biodiversity cluster distribution.

Variable	Units	Detail	Source
Depth	m	Mitchell 2019 et al. 2019 Bathymetry (m)	Mitchell et al. (2019)
Valley depth	m	SAGA Valley depth	derived metric from Depth layer
Closed depressions	–	SAGA Closed depressions	derived metric from Depth layer
Ch. network distance	m	SAGA Channel network distance	derived metric from Depth layer
LS-factor	–	SAGA slope length and steepness factor	derived metric from Depth layer
Rel. slope pos.	–	SAGA Relative slope position	derived metric from Depth layer
Wave velocity	m s^{-1}	Mitchell 2019 Wave velocity	Mitchell et al. (2019)
Current speed	m s^{-1}	Mitchell 2019 Current speed	Mitchell et al. (2019)
Gravel	%	Mitchell 2019 Gravel fraction (%)	Mitchell et al. (2019)
Mud	%	Mitchell 2019 Mud fraction (%)	Mitchell et al. (2019)
Mean SPM	g m^{-3}	Mitchell 2019 Average suspended matter	Mitchell et al. (2019)
Summer SPM	g m^{-3}	Mitchell 2019 Summer suspended matter	Mitchell et al. (2019)
Winter SPM	g m^{-3}	Mitchell 2019 Winter suspended matter	Mitchell et al. (2019)
Diss. Iron	mmol m^{-3}	Bio-Oracle v.3 Dissolved Iron [depthMean] Baseline 2000–2018.	https://www.bio-oracle.org/
Diss. Oxygen	mmol m^{-3}	Bio-Oracle v.3 Dissolved Molecular Oxygen [depthMean] Baseline 2000–2018.	https://www.bio-oracle.org/
Nitrate	mmol m^{-3}	Bio-Oracle v.3 Nitrate [depthMean] Baseline 2000–2018.	https://www.bio-oracle.org/
Bottom temp.	°C	Bio-Oracle v.3 Ocean Temperature [depthMean] Baseline 2000–2019.	https://www.bio-oracle.org/
Bottom temp. range	°C	Bio-Oracle v.3 Ocean Temperature [depthMean] Baseline 2000–2019.	https://www.bio-oracle.org/
pH	–	Bio-Oracle v.3 pH [depthMean] Baseline 2000–2018.	https://www.bio-oracle.org/
Phosphate	mmol m^{-3}	Bio-Oracle v.3 Phosphate [depthMean] Baseline 2000–2018.	https://www.bio-oracle.org/
Salinity mean	ppt	Bio-Oracle v.3 Salinity [depthMean] Baseline 2000–2019.	https://www.bio-oracle.org/
Salinity range	ppt	Bio-Oracle v.3 Salinity [depthMean] Baseline 2000–2019.	https://www.bio-oracle.org/
Silicate	mmol m^{-3}	Bio-Oracle v.3 Silicate [depthMean] Baseline 2000–2018.	https://www.bio-oracle.org/
Chlorophyll	mmol m^{-3}	Bio-Oracle v.3 Chlorophyll [depthSurf] Baseline 2000–2018.	https://www.bio-oracle.org/
KD PAR (Light)	m^{-1}	Bio-Oracle v.3 Diffuse Attenuation Coefficient PAR [depthSurf] Mean 2000–2020.	https://www.bio-oracle.org/
Phytoplankton	mmol m^{-3}	Bio-Oracle v.3 Total Phytoplankton [depthSurf] Baseline 2000–2020.	https://www.bio-oracle.org/

describes the relative position along a slope (from 0 at the base to 1 at the crest) and can reflect differences in hydrodynamic conditions associated with the vertical position of seabed features (Böhner & Selige 2006). To ensure consistency across all inputs, Bio-ORACLE raster layers were cropped and resampled to match the spatial extent and pixel resolution of those from Mitchell et al. (2019).

Response variables (univariate metrics and assemblage clusters)

Full-coverage maps of community metric (S, N) and assemblage clusters were generated using random forest modelling-an ensemble method that builds a large number of decision trees (typically 500–1 000) from random subsets of the samples and predictor variables (Breiman 2001; Cutler et al. 2007). A random forest modelling approach was selected for its capacity to predict both numeric and categorical response variables, and its ability to account for complex interactions and nonlinear relationships between response and predictor variables (Rodriguez-Galiano et al. 2012). Models were implemented in R using the *randomForest* package (Liaw & Wiener, 2002) with the default settings and 1 000 trees.

Regression trees were applied to the continuous response variables—taxon richness (S) and total abundance (N)—to model how these metrics vary in space. A regression tree is a decision-tree-based model that predicts a numeric outcome by repeatedly splitting the data based on environmental variables to minimise prediction error. The random forest approach builds many such trees (here, 1 000), each trained on a random subset of the data and predictors, and then averages their predictions to improve accuracy and reduce overfitting.

For the categorical response variable—epifaunal assemblage cluster—we used classification trees, which operate similarly but are designed to predict discrete categories instead of numeric values. Each tree assigns a sample to a cluster based on environmental conditions, and the final prediction is based on the majority vote across all trees (i.e. the most frequently predicted cluster). In addition, we calculated the class-specific probability for each prediction, reflecting the proportion of trees that assigned a given sample to a particular cluster—providing a measure of confidence in the classification.

Preliminary models including all environmental variables were run to identify the most informative predictors and remove those exhibiting high covariance. Redundant variables-those with strong correlations with other predictors or poor explanatory relationships with the response variable-were excluded. In cases of high correlation, the variable deemed less mechanistically linked to epifaunal assemblages was removed. Covariance among predictors was assessed using values extracted from raster layers at sample locations, with correlation analysis used to visualise relationships and guide variable selection.

While Random Forest models are generally robust to multicollinearity (Huang and Boutros 2016), simpler models with fewer predictors are easier to interpret and yield more variance importance measures. Highly correlated predictors can obscure the contribution of individual variables by acting interchangeably within component trees.

Model Performance and Predictive Confidence

We evaluated model performance and robustness using cross-validation with repeated subsampling, a common technique to assess predictive accuracy and generalisability (Mitchell et al. 2018). Specifically, we generated ten independent train-test splits of the data. In each split, 75% of the samples were randomly selected for model training and 25% reserved for testing. Stratified sampling was used to ensure that all response variable levels (e.g. cluster classes) were proportionally represented in both sets.

For the continuous response variables—taxon richness (S) and abundance (N)—we used Random Forest regression models. Final predictions were computed as the mean of the ten model runs at each spatial location (i.e. grid cell). To assess spatial prediction uncertainty, we calculated the coefficient of variation (CV) at each location, defined as the standard deviation divided by the mean of predictions across runs. Model performance was summarised using the coefficient of determination (R^2), reporting both the mean and standard deviation across the ten validation runs.

For the categorical response variable—the epifaunal assemblage clusters—we used Random Forest classification models. At each location, the final predicted class was determined by majority vote: the class most frequently predicted by the ten models. To quantify prediction confidence, we generated three complementary confidence layers: i) Class frequency: the number of times (out of 10) that the predicted class was selected; ii) Mean class probability: the average predicted probability for the most frequently selected class; iii) Combined confidence index: the product of (i) and (ii), providing a single value that reflects both consistency and probability-based certainty.

Classification model performance was evaluated using standard metrics derived from confusion matrices: sensitivity (true positive rate), specificity (true negative rate), and balanced accuracy (the average of sensitivity and specificity). These metrics were reported for each assemblage class and overall, averaged across the ten validation runs.

The final model outputs of the classification model therefore include:(a) predicted maps of epifaunal cluster distributions (based on majority-vote classification across ten models), and (b) spatial confidence maps that indicate prediction reliability at each location, where higher values denote greater confidence in the class assignment.

Explaining patterns

Two broad approaches were used to explain the drivers of epifaunal spatial patterns, tailored separately to the continuous community metrics and the categorical assemblage clusters, as described below.

Community metrics

For the univariate metrics S and N, explanatory patterns were assessed using Variable Importance (VI) and Partial Dependence (PD) plots -derived from the random forest models. VI plots rank predictor variables by their contribution to model accuracy, calculated via Monte Carlo permutation of "out-of-bag" data, and reported as the percentage increase in mean squared error (% MSE) when the variable is permuted. PD plots illustrate the marginal effect of a predictor variable (x-axis) on the modelled response (y-axis), providing insight into the functional relationship between them. For clarity and brevity, we present PD plots for the top three predictor variables identified by the VI plots for each modelled layer.

Assemblage clusters

To explore the extent to which environmental variables explain variation in the assemblage clusters, we applied the *best* and *adonis* functions from the R *vegan* package (Oksanen et al. 2024). The *best* function identifies the subset of environmental variables that maximises rank correlation with assemblage dissimilarities, effectively identifying the predictors that best explain variation in the biodiversity data. The *adonis* function performs a permutational multivariate analysis of variance (PERMANOVA) using distance matrices to quantify how much of the variability in assemblage structure can be explained by the predictor variables. Relationships between samples and environmental predictors were visualised using distance-based redundancy analysis (dbRDA) ordination plots, where the axes represent the primary linear combinations of environmental variables that explain the greatest variation in community composition. Assemblage dissimilarities were calculated using the Bray-Curtis resemblance measures (Bray and Curtis 1957).

Prior to analyses, environmental variables (see Section 2.4.1 and Table 1) were screened for multicollinearity using Variation Inflation Factors (VIFs) calculated using the *vifstep* function in the *usdm* package (Naimi et al. 2014). The *vifstep* function applies a stepwise procedure to iteratively exclude variables with high VIF values (>2.5). This resulted in the removal of wave velocity and current.speed due to high collinearity with other predictors. To address right-skewed distributions, $\log(x + 0.1)$ transformations were applied to winter SPM, gravel, rel. slope pos., LS factor, mud and closed depressions. All selected environmental variables were then normalised to a common scale. Environmental dissimilarities were subsequently calculated using Euclidean distance, which is suitable for continuous, standardised variables and allows for the quantification of overall differences in environmental conditions between samples.

Results

Community metrics

Model description

Spatial patterns in the two univariate metrics assessed—taxon richness (S) and abundance (N)—are shown in Fig. 2a and 2b respectively. Family-level richness (S) across the 1 400 samples ranged from 1 to 63 and exhibited clear, broad-scale spatial gradients. Elevated values of S were observed in the northern North Sea, the mid- and western English Channel, the Southwest Approaches and along the southern Irish coast (Fig. 2a, left panel). In contrast, lower values of S were recorded in the southern North Sea, eastern English Channel and much of the Irish Sea, particularly in inshore areas.

The spatial distribution of total abundance (N) displayed a distinctly different—and in some regions, contrasting—pattern relative to S. Untransformed abundance values ranged from 1 to 8 263 individuals per sample. Based on the square-root-transformed abundance model, low abundances were widespread across the study area, including most of the North Sea, the Irish Sea, and the eastern and western English Channel (Fig. 2b, left panel). In contrast, elevated abundances were lo-

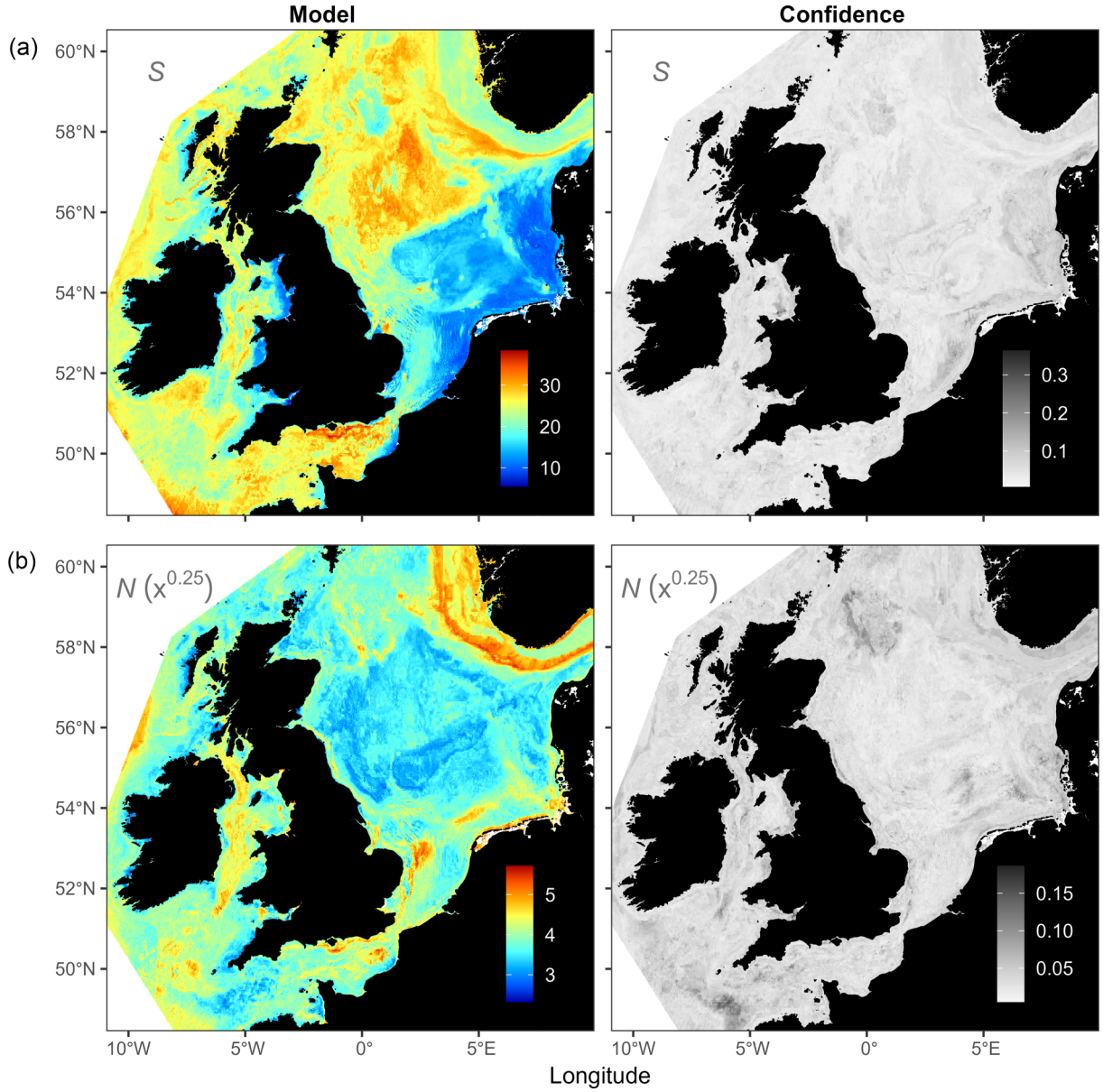

Figure 2. Spatial predictions of epifaunal taxon richness (a) and square-root-transformed total abundance (b) across the study area. Colour scales indicate predicted values per sample, with blue representing low values and red representing high values. Right-hand panels show the associated model confidence, with colour scales indicating the coefficient of variation.

calised and found along the Norwegian coast, in the German Bight off the Dutch coast, inshore areas of the mid-English Channel (both English and French coasts), the Southwestern Approaches, and a small area off southwest Wales in the Celtic Sea (Fig. 2b, left panel).

Model performance and predictive confidence

The Random Forest model for taxon richness (S) performance was characterised by having a relatively low R^2 value of 0.37 (± 0.03 SD), indicating the model explains a moderate, though limited, proportion of the variance in S. The spatial distribution of model confidence was heterogeneous, with variation observed at relatively fine spatial scales; no broad region of the study area consistently demonstrated either high or low model confidence (Figure 2a, right panel).

The Random Forest model for square-root-transformed total abundance (N) achieved a lower R^2 value of 0.13 (± 0.03 SD), reflecting relatively low explanatory power. As with S, model performance varied over small spatial scales. Regions of comparatively higher predictive confidence were evident in the western English Channel and northern North Sea (Figure 2b, right panel).

Explaining patterns

Partial dependence plots (PDPs) for the top three predictor variables provide insights into the environmental drivers of spatial variation in epifaunal community metrics. For taxon richness (S), the three most influential variables were mean salinity, wave velocity, and current speed (Fig. 3a). Mean salinity, identified as the most important predictor, showed a clear positive relationship with richness, indicating higher taxon richness in areas of elevated salinity. In contrast, wave velocity exhibited a negative relationship with taxon richness. Current

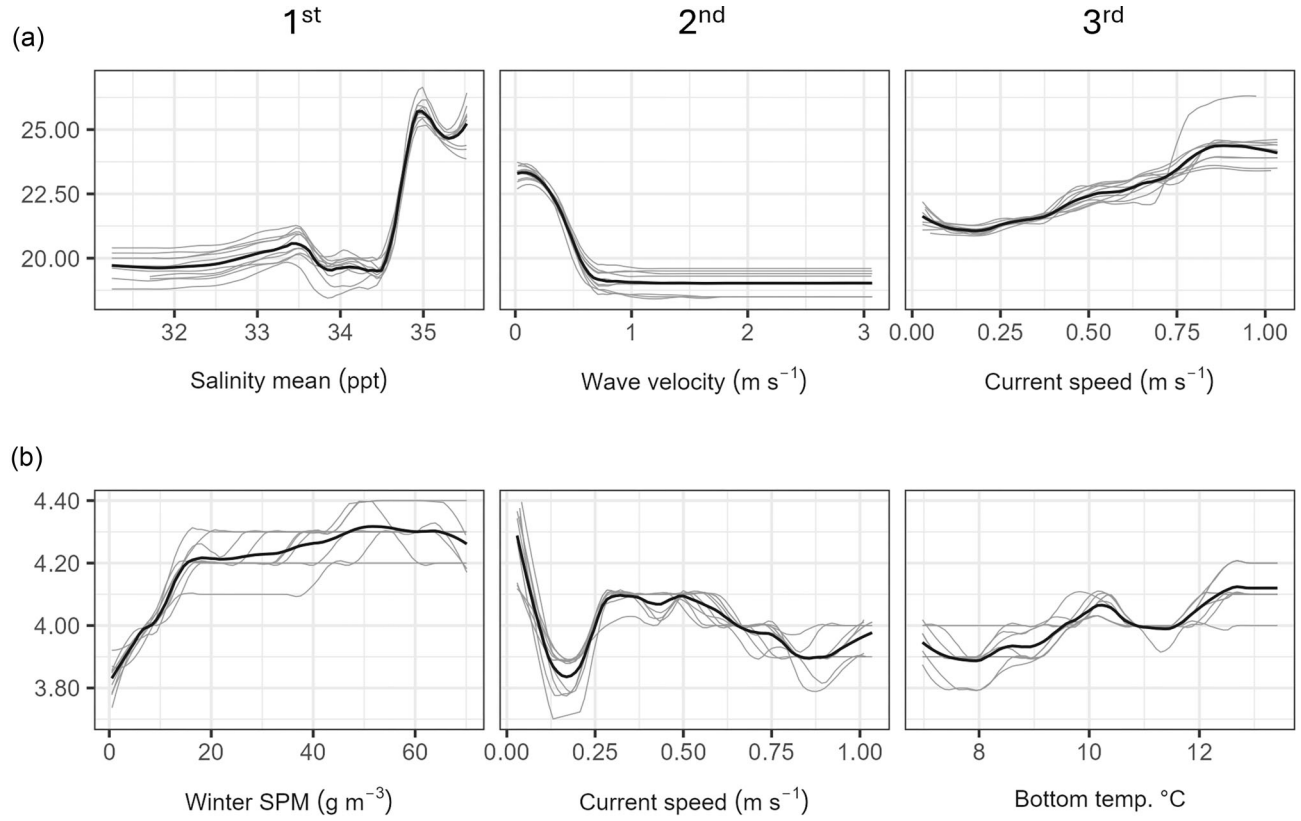

Figure 3. Partial dependence plots (PDPs) from the random forest model predicting (a) taxon richness, and (b) square-root-transformed total abundance. Each panel shows the marginal effect of a predictor variable on each metric. The 1st, 2nd, and 3rd labels above the subplots indicate the ranking of the most influential variables, with the 1st being the most important according to the mean decrease in Gini coefficient. Grey lines represent model iterations.

speed displayed a generally positive relationship, with richness increasing alongside faster bottom currents.

For observed variations in square-root-transformed total abundance (N) values, the top predictors were winter SPM, current speed, and bottom temperature respectively (Fig. 3b). The relationship with winter SPM was nonlinear: abundance peaked at moderate SPM levels but declined under very high concentrations, possibly reflecting the dual role of SPM as both a food source and a potential stressor. Current speed showed a weakly nonlinear pattern, with a decline in abundance at low speeds, followed by stabilisation and a slight increase at higher speeds. Bottom temperature was positively associated with abundance, suggesting that warmer benthic conditions may enhance epifaunal productivity or survival in certain regions.

Multivariate taxonomic assemblage structure

Clustering

K-means clustering of square-root-transformed epifaunal abundance data identified 11 distinct assemblage groups, based on a cut-off point where the rate of change in total within-cluster sum of squares diminished (Fig. 4a). While the elbow plot suggested that 13 clusters might also be justifiable, a solution of 11 clusters was selected to align with the objectives of this study—specifically, to avoid over partitioning subtle differences in assemblages across the broad spatial extent of the study region. This approach ensures that the resulting clusters are ecologically interpretable and relevant at management scales, rather than reflecting minor variations that may not be meaningful for large-scale spatial analysis or practical application. The resulting 11 cluster groups exhibit varying degrees of assemblage similarity and dissimilarity (Fig. 4b). A higher-level grouping structure was apparent, with the clusters organised into five main groups (Epi_A to Epi_E) based on a 70% similarity threshold. These groups were named to reflect their relative similarities, such that, for example, Epi_A and Epi_B are more similar to each other than to Epi_C, Epi_D, or Epi_E. One exception to this grouping structure was Epi_C, which did not contain sub-groups and appeared more distinct. Colours were applied to the cluster visualisation to reflect relative similarity relationships and facilitate interpretation of spatial patterns.

Model description

The spatial distribution of modelled epifaunal assemblage clusters across the study area is presented in Fig. 5, with associated group characteristics summarised in Table 2. Clusters in the Epi_A group (e.g. Epi_A1; navy blue) are highly localised, restricted to small inshore areas of the east coast of the UK. Despite their very limited spatial extent, they are notable for having among the highest taxon richness and abundance values observed in the study.

The Epi_B group (Epi_B1a in turquoise and Epi_B1b in grey blue) is largely confined to the English Channel. These clusters show high taxon richness relative to others, with moderate abundance, and are more broadly distributed than the Epi_A group, though still regionally restricted.

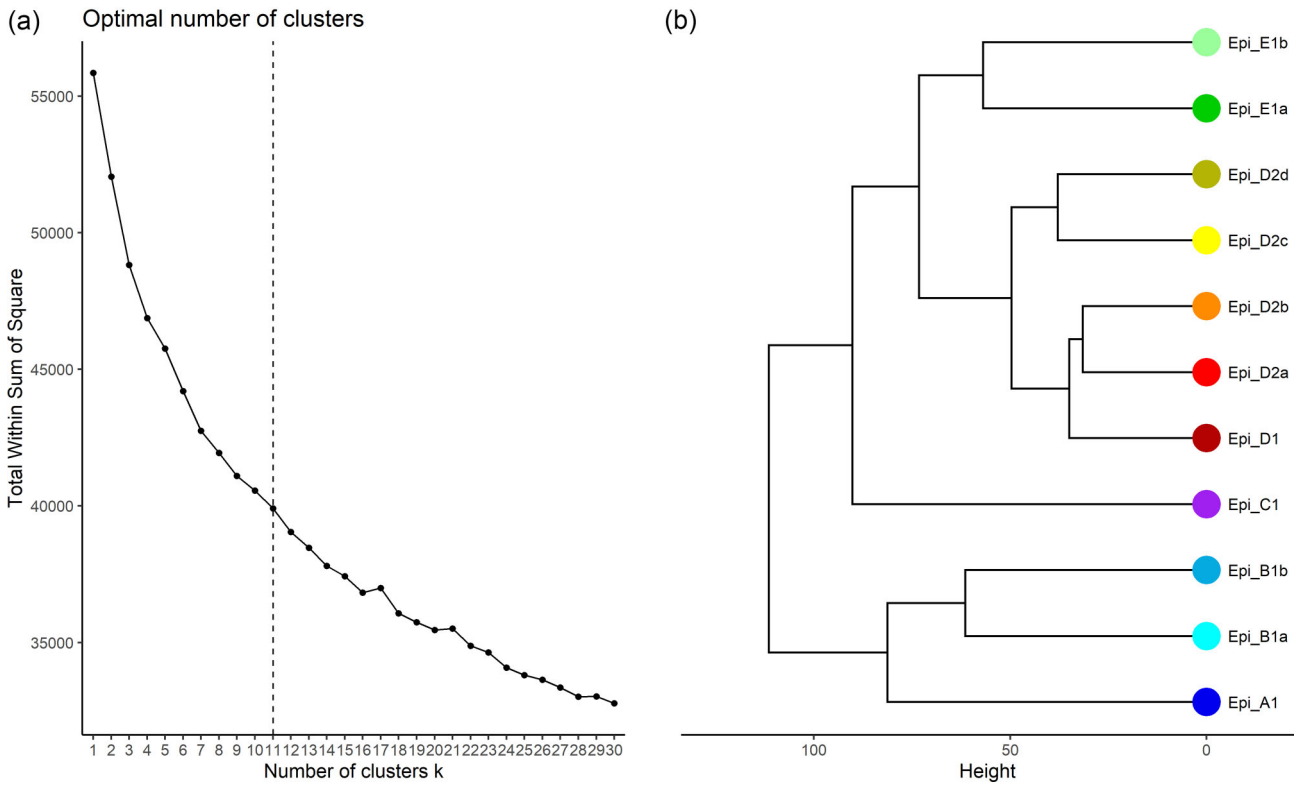

Figure 4. Outputs of k means clustering of the 2 m beam trawl square root-transformed abundance data, showing (a) elbow plot used to aid decisions regarding number of cluster groups present, and (b) dendrogram showing the relative similarity in epifaunal taxonomic multivariate structure of each cluster group. In (b), the y-axis represents Euclidean distance between cluster centroids, indicating the degree of dissimilarity in assemblage composition.

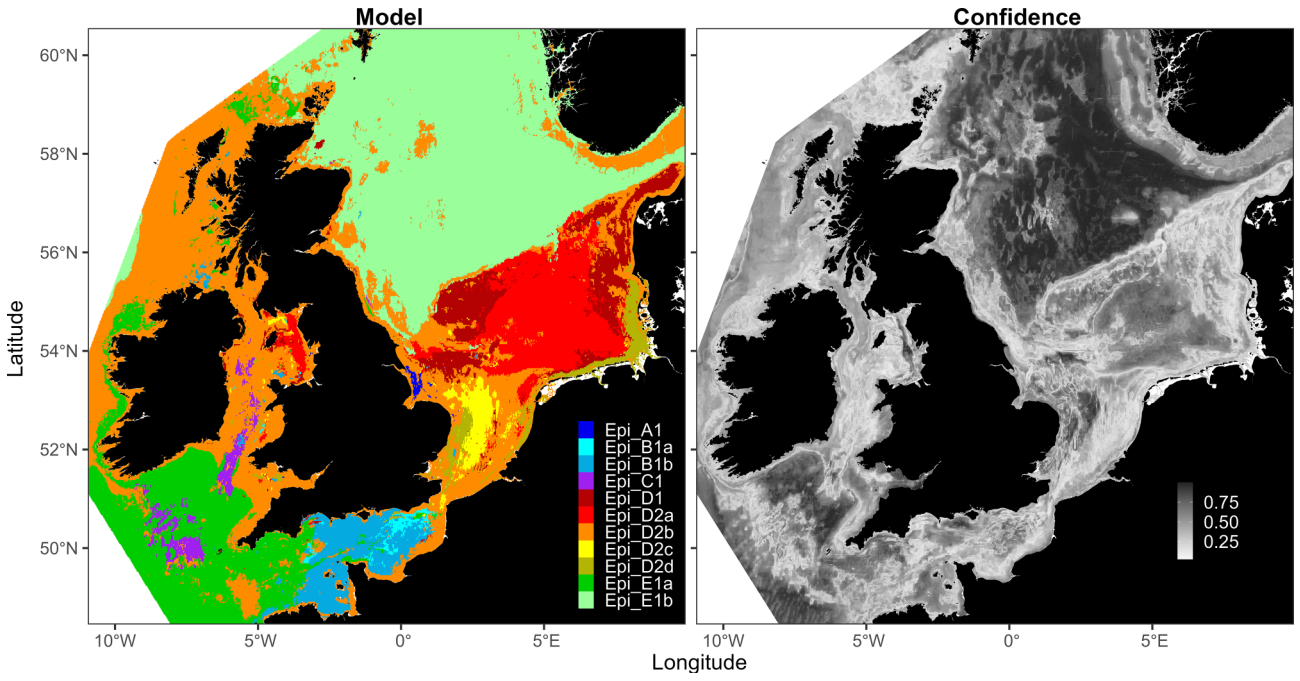

Figure 5. Cluster map showing the spatial distributions of the eleven epifaunal cluster groups across the study area (left). Model confidence is shown in the right panel, with higher values indicating higher confidence. Confidence values are expressed as probabilities (ranging from 0 to 1).

Table 2. Cluster group characteristics including mean richness, mean abundance, phyletic composition (pie charts) and characterising taxa (family level or above).

Cluster	% Area	Richness (mean)	Abundance (mean)	Phyla	Taxa
Epi_A1	0.1	30	1 621		**Pandalidae** (Art), **Polybiidae** (Art), **Asteriidae** (Ech), **Crangonidae** (Art), **Paguridae** (Art), Inachidae (Art), Solasteridae (Ech), Ophiuridae (Ech), Cancridae (Art), Calyptraeidae (Mol), Sabellariidae (Ann), Callionymidae (Cho), Gobiidae (Cho), Pleuronectidae (Cho), **Flustridae** (Bry), Porcellanidae (Art), Agonidae (Cho), Liparidae (Cho), Sertulariidae (Cni), Soleidae (Cho), Pholidae (Cho), Ascidiacea (Cho), Galatheidae (Art), Ampeliscidae (Art), Pilumnidae (Art), Sepiolidae (Mol), Actiniaria (Cni), Cottidae (Cho), Oregoniidae (Art), Calliostomatidae (Mol)
Epi_B1a	0.5	39	888		**Parechinidae** (Ech), **Pectinidae** (Mol), **Paguridae** (Art), **Inachidae** (Art), **Ophiuridae** (Ech), **Asteriidae** (Ech), **Ophiotrichidae** (Ech), **Polybiidae** (Art), **Ascidiidae** (Cho), **Galatheidae** (Art), Anomiidae (Mol), Hormathiidae (Cni), Porcellanidae (Art), Leucosiidae (Art), Gobiidae (Cho), Pandalidae (Art), Crangonidae (Art), Oregoniidae (Art), Buccinidae (Mol), Callionymidae (Cho), Sertulariidae (Cni), Styelidae (Cho), Serpulidae (Ann), Alcyoniidae (Cni), Sagartiidae (Cni), Glycymeridae (Mol), Calliostomatidae (Mol), Calyptraeidae (Mol), Balanidae (Art), Plumulariidae (Cni), Gobiesocidae (Cho), Molgulidae (Cho), Majidae (Art), Pyuridae (Cho), Alcyonidiidae (Bry), Atelecyclidae (Art), Flustridae (Bry), Trochidae (Mol), Soleidae (Cho)
Epi_B1b	3.9	29	319		**Pectinidae** (Mol), **Paguridae** (Art), **Parechinidae** (Ech), **Inachidae** (Art), **Pandalidae** (Art), **Ophiotrichidae** (Ech), **Asteriidae** (Ech), Buccinidae (Mol), **Polybiidae** (Art), Galatheidae (Art), Gobiidae (Cho), Calyptraeidae (Mol), Calliostomatidae (Mol), Serpulidae (Ann), Oregoniidae (Art), Sertulariidae (Cni), Callionymidae (Cho), Ophiuridae (Ech), Crangonidae (Art), Flustridae (Bry), Styelidae (Cho), Porcellanidae (Art), Pilumnidae (Art), Ascidiidae (Cho), Leucosiidae (Art), Muricidae (Mol), Plumulariidae (Cni), Pyuridae (Cho), Alcyoniidae (Cni)
Epi_C1	1.3	23	1 219		**Crangonidae** (Art), **Nuculidae** (Mol), **Pandalidae** (Art), **Nephropidae** (Art), **Processidae** (Art), Alpheidae (Art), Turritellidae (Mol), Polybiidae (Art), Naticidae (Mol), Goneplacidae (Art), Paguridae (Art), Gobiidae (Cho), Pleuronectidae (Cho), Munididae (Art), Astropectinidae (Ech), Aphroditidae (Ann), Amphipoda (Art), Inachidae (Art), Amphiuridae (Ech), Aporrhaidae (Mol), Sepiolidae (Mol), Pharidae (Mol), Cirolanidae (Art)
Epi_D1	5.0	20	296		**Asteriidae** (Ech), **Soleidae** (Cho), **Astropectinidae** (Ech), **Polybiidae** (Art), **Pleuronectidae** (Cho), Bothidae (Cho), Paguridae (Art), Callionymidae (Cho), Ophiuridae (Ech), Flustridae (Bry), Ammodytidae (Cho), Inachidae (Art), Alcyonidiidae (Bry), Gobiidae (Cho), Trachinidae (Cho), Sertulariidae (Cni), Corystidae (Art), Parechinidae (Ech), Alcyoniidae (Cni), Crangonidae (Art)
Epi_D2a	7.8	15	419		**Ophiuridae** (Ech), **Asteriidae** (Ech), **Paguridae** (Art), Astropectinidae (Ech), Polybiidae (Art), Loveniidae (Ech), Aphroditidae (Ann), Philinidae (Mol), Corystidae (Art), Soleidae (Cho), Crangonidae (Art), Inachidae (Art), Bothidae (Cho), Hydractiniidae (Cni), Parechinidae (Ech)
Epi_D2b	31.5	13	105		**Crangonidae** (Art), **Paguridae** (Art), **Polybiidae** (Art), **Asteriidae** (Ech), Ophiuridae (Ech), Pandalidae (Art), Gobiidae (Cho), Inachidae (Art), Sertulariidae (Cni), Pleuronectidae (Cho), Soleidae (Cho), Flustridae (Bry), Trachinidae (Cho)
Epi_D2c	1.4	20	538		**Ophiuridae** (Ech), **Crangonidae** (Art), **Gobiidae** (Cho), **Soleidae** (Cho), **Polybiidae** (Art), Trachinidae (Cho), **Paguridae** (Art), Pleuronectidae (Cho), Asteriidae (Ech), Callionymidae (Cho), Bothidae (Cho), Ammodytidae (Cho), Sepiolidae (Mol), Hydractiniidae (Cni), Agonidae (Cho), Inachidae (Art), Loveniidae (Ech), Parechinidae (Ech), Mactridae (Mol), Balanidae (Art)
Epi_D2d	1.8	19	727		**Ophiuridae** (Ech), **Crangonidae** (Art), **Paguridae** (Art), **Asteriidae** (Ech), Polybiidae (Art), Parechinidae (Ech), Inachidae (Art), Gobiidae (Cho), Pandalidae (Art), Mactridae (Mol), Soleidae (Cho), Sertulariidae (Cni), Agonidae (Cho), Naticidae (Mol), Buccinidae (Mol), Trachinidae (Cho), Sepiolidae (Mol), Callionymidae (Cho), Pleuronectidae (Cho)
Epi_E1a	13.4	28	406		**Paguridae** (Art), **Crangonidae** (Art), **Ophiuridae** (Ech), **Inachidae** (Art), **Polybiidae** (Art), **Gobiidae** (Cho), Hormathiidae (Cni), Onuphidae (Ann), **Callionymidae** (Cho), Processidae (Art), Leucosiidae (Art), Astropectinidae (Ech), Asteriidae (Ech), Turritellidae (Mol), Soleidae (Cho), Pectinidae (Mol), Aporrhaidae (Mol), Veneridae (Mol), Caryophyllidae (Cni), Galatheidae (Art), Luidiidae (Ech), Pandalidae (Art), Ophiotrichidae (Ech), Parechinidae (Ech), Atelecyclidae (Art), Bothidae (Cho), Scaphandridae (Mol), Mactridae (Mol)
Epi_E1b	33.2	28	299		**Paguridae** (Art), **Astropectinidae** (Ech), **Asteriidae** (Ech), **Buccinidae** (Mol), **Crangonidae** (Art), Echinidae (Ech), **Colidae** (Mol), Pandalidae (Art), Onuphidae (Ann), Oregoniidae (Art), Ophiuridae (Ech), **Hormathiidae** (Cni), **Sertulariidae** (Cni), Thoridae (Art), Serpulidae (Ann), Flustridae (Bry), Luidiidae (Ech), Polybiidae (Art), Suberitidae (Por), Alcyonidiidae (Bry), Loveniidae (Ech), Hydractiniidae (Cni), Epizoanthidae (Cni), Aphroditidae (Ann), Ascidiidae (Cho), Pectinidae (Mol), Alcyoniidae (Cni), Celleporidae (Bry)

Pie Chart Legend: Arthropoda (Art), Chordata (Cho), Echinodermata (Ech), Mollusca (Mol), Cnidaria (Cni), Bryozoa (Bry), Annelida (Ann), Porifera (Por)

Note: Values for richness and abundance are group sample averages, where richness is reported as the mean number of family-level taxa per trawl sample and abundance as the mean number of individuals per trawl sample. Listed taxa are those with the highest mean centroid values, where the centroid represents the average abundance of each taxon within the cluster group in multivariate space. Highlighted taxa are those identified by a SIMPER analysis as contributing to ~50% of the similarity between samples. Phyla codes are given in parenthesis (see legend at foot of table).

Table 3. Mean and standard deviation of model validation statistics for individual cluster groups and overall based on 10 random split sample runs.

Cluster	n	Sensitivity	Specificity	Balanced accuracy
Epi_A1	9	0.32 ± 0.15	0.99 ± 0.00	0.66 ± 0.08
Epi_B1a	17	0.28 ± 0.06	0.98 ± 0.01	0.63 ± 0.03
Epi_B1b	32	0.5 ± 0.09	0.95 ± 0.01	0.73 ± 0.04
Epi_C1	10	0.61 ± 0.14	0.99 ± 0.01	0.8 ± 0.07
Epi_D1	27	0.58 ± 0.09	0.98 ± 0.01	0.78 ± 0.04
Epi_D2a	27	0.63 ± 0.09	0.97 ± 0.01	0.8 ± 0.04
Epi_D2b	86	0.5 ± 0.04	0.83 ± 0.01	0.66 ± 0.02
Epi_D2c	25	0.71 ± 0.07	0.97 ± 0.01	0.84 ± 0.04
Epi_D2d	32	0.46 ± 0.07	0.97 ± 0.01	0.71 ± 0.04
Epi_E1a	44	0.79 ± 0.07	0.94 ± 0.01	0.86 ± 0.03
Epi_E1b	36	0.92 ± 0.04	0.98 ± 0.01	0.95 ± 0.02
Overall	1 055	0.6 ± 0.02	0.96 ± 0.00	0.78 ± 0.01

Table 4. Results of a "best" analysis identifying the subset of environmental variables which are most correlated with the epifaunal abundance data.

Size	Variables	Correlation (ρ)
1	Salinity mean	0.2255
2	Salinity mean, Bottom temp.	0.2378
3	Salinity mean, Bottom temp., Gravel	0.2575
4	Salinity mean, Bottom temp., Gravel, Rel. slope pos.	0.2616
5	**Salinity mean, Winter SPM, Gravel, Rel. slope pos., Mud**	**0.2641**

Note: "Size" refers to the number of environmental variables included in each subset tested. The row for size 5 is shown in bold because this subset achieved the highest correlation (i.e. represents the best model)

Epi_C1 (purple) is found mainly in offshore regions such as the mid Irish Sea and Outer Celtic Sea. It is characterised by moderate taxon richness but high mean abundance, indicating that these offshore shelf environments support dense but less diverse assemblages.

The Epi_D clusters (Epi_D1 in blood red, Epi_D2a in red, Epi_D2b in orange, Epi_D2c in yellow, and Epi_D2d in khaki) are the most spatially widespread overall, together covering nearly half of the model domain. Among these,, Epi_D2b (orange) is one of the most extensive clusters, spanning much of the UK's west coast as well as the eastern seaboard of mainland Europe. Across the D clusters, taxon richness is generally lower, with abundance varying from low to moderate.

The Epi_E group (Epi_E1a in dark green and Epi_E1b in light green) demonstrates a clear west–east contrast. Epi_E1b is the single most extensive cluster overall, occupying large areas of the northern North Sea, while Epi_E1a dominates the southwest approaches. Both E clusters show moderate richness, but typically lower abundance values in more northern waters.

Overall, the modelled clusters highlight marked biogeographic structuring of epifaunal assemblages across northwest European shelf seas, reflecting regional environmental gradients and oceanographic conditions (Fig. 5; Table 2).

Model performance and predictive confidence

The random forest spatial model used to predict seabed epifaunal assemblage distribution showed generally strong performance, with an overall sensitivity of 0.60 (± 0.02 SD), specificity of 0.96 (± 0.00 SD), and balanced accuracy of 0.78 (± 0.01 SD) (Table 3). These metrics indicate that the model is generally effective in distinguishing between different epifaunal assemblages. Assemblage-specific performance varied, with clusters such as Epi_E1b achieving high sensitivity (0.92 ± 0.04 SD) and balanced accuracy (0.95 ± 0.02 SD), indicating strong predictive capability. Conversely, clusters such as Epi_A1 and Epi_B1a showed lower sensitivity (0.32 ± 0.15 SD and 0.28 ± 0.06 SD, respectively), suggesting areas for improvement. The high specificity values across all clusters (ranging from 0.83 to 0.99; Table 3) highlight the model's effectiveness in avoiding false positives. These results suggest that while the model is generally effective in predicting epifaunal assemblages, further refinement is needed to enhance its performance for certain clusters. Confidence in the model varies spatially as shown in Fig. 5b, with notably higher confidence in the northern North Sea.

Explaining Patterns

The *best* analysis identified five variables: salinity mean, winter SPM, gravel, rel. slope pos. and mudas best explaining the spatial variability in epifaunal assemblage structure (Table 4). Correlation between the environmental and biological resemblance matrices was weak to moderate ($\rho = 0.264$), and PERMANOVA (*adonis*) results indicated that these predictors collectively explained only 7.1% of the total variation in assemblage composition. Individual contributions were as follows: salinity mean = 2.2%, winter SPM = 2.1%, gravel = 1.8%, rel. slope, pos. = 0.7% and mud = 0.3%. Despite the low proportion of explained variance, all predictors were statistically significant ($p < 0.001$). Spatial patterns shown by these predictors are shown in Supplementary Fig. S4.

The dbRDA ordination (Fig. 6) highlights environmental gradients influencing epifaunal assemblage distribution across the study area. Assemblages Epi_B1a and Epi_B1b are associated with sediments containing higher gravel content, whereas C1 aligns with higher mud content. Assemblages Epi_D2a, Epi_D2c, and Epi_D2d cluster toward the winter SPM vector, indicating a shared association with areas experiencing elevated suspended particulate matter during winter, with intermediate sediment conditions likely characterised by high sand content. In contrast, Epi_E1a and Epi_E1b occur toward the Salinity vector, suggesting a preference for more saline environments, typically away from areas of high winter SPM, and in locations where sediments are likely to be finer. While most remaining groups cluster closer to the centre of the ordination, Epi_A1a (blue) shows a tendency toward elevated winter SPM and gravel, indicating some affinity for coarser sediments under higher particulate conditions. Epi_D1 (red) and Epi_D2b (orange) remain near the centre, suggesting mixed environmental influences without a strong association with any single variable.

Discussion

By integrating disparate but comparable datasets, this study demonstrates that a big data approach can produce large-scale maps of epifaunal assemblage structure, which have the potential to inform management decisions. Notable regions of elevated epifaunal taxon richness (measured as number of taxa per sample) were identified in the northern North Sea and the English Channel, with significantly higher densities in

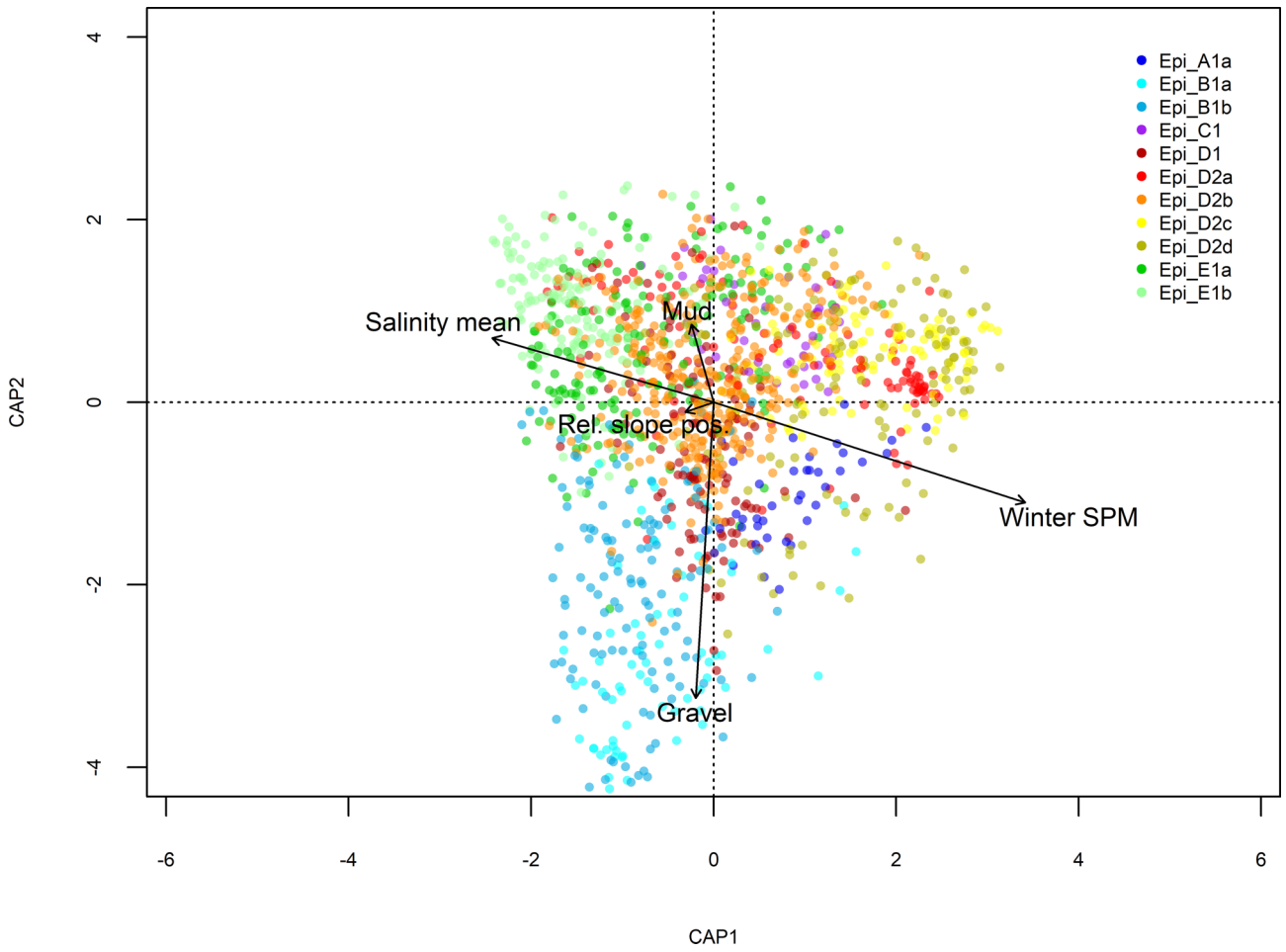

Figure 6. Distance-based redundancy analysis (dbRDA) ordination of epifaunal assemblage samples (points) constrained by environmental variables. Points are coloured by predicted assemblage type. Vectors indicate the direction and strength of environmental gradients (salinity mean, winter SPM, gravel, rel. slope pos. and mud) associated with variation in assemblage structure. The axes (CAP1 and CAP2) represent the primary canonical axes derived from principal coordinates analysis, showing the greatest proportion of variation in assemblage structure explained by the environmental predictors. The plot shows separation among assemblage types along environmental gradients, with notable influences of sediment composition and water column properties on assemblage differentiation.

the Irish Sea, southern North Sea, and along the Norwegian coast. The study distinguished five broad assemblage types (labelled A–E), each characteristic of distinct large areas of the study region, with three of these groups (B, D, and E) exhibiting more subtle internal variation that is further resolved into a total of 11 assemblage clusters. Clear structural differences were evident among assemblages inhabiting the northern North Sea, southern North Sea, Celtic Sea, eastern English Channel, and western English Channel. These differences appear to be driven by a combination of bed sediment particle size (e.g. estimated percentage contribution of gravel and mud content) and water column characteristics (e.g. salinity mean and winter SPM). Notably, the observed negative relationship between wave velocity and taxon richness suggests that higher wave energy may reduce habitat suitability or stability for diverse epifaunal communities, likely due to increased physical disturbance or sediment mobility in high-energy environments (Kaiser et al. 1999; Reiss et al. 2010). The generally positive relationship observed between current speed and taxon richness is also noteworthy. While higher current speeds might be expected to reduce local larval recruitment due to increased dispersal, they may also enhance connectivity and the influx of species from other areas. This increased exchange can support higher local richness, as sites are more frequently replenished by a wider pool of taxa, potentially increasing biodiversity at both local and regional scales. The resulting maps offer valuable information on the presence and variation of epifaunal assemblages, supporting environmental management and decision-making. This can help guide activities such as offshore construction, sediment disposal, and marine aggregate extraction to avoid areas of high biodiversity or particular assemblage types, thereby reducing the risk of adverse impacts.

Over recent decades, there has been increasing recognition of the value of reliable maps of habitats, species distributions, and community-structure metrics, both for understanding the processes shaping marine spatial patterns and for managing human activities and conservation (Davies and Guinotte 2011; Harris and Baker 2012; Lecours 2017; Fraschetti et al. 2024). This expansion in mapping has been facilitated by improved data acquisition capacity—both observational and environmental predictors (Costa et al. 2009; Heyman and Wright 2011)—greater computational power, and the growing body of evidence on cumulative anthropogenic impacts (Halpern et al. 2008). In the marine environment, most mapping efforts have focused on habitats (areas defined by spe-

cific physical, chemical, or biological characteristics) because of their utility as management units (Lecours et al. 2015). Maps of marine biological assemblage distributions have received less attention, partly because they require large, standardised biological datasets collected across broad spatial extents. Cooper and Barry (2017) demonstrated that large-scale maps of macroinfaunal taxonomic structure and univariate community metrics could be produced by integrating multiple datasets with careful harmonisation of metadata to ensure comparability. This study extends that approach to the epifaunal component of benthic ecosystems, providing new insights into the large-scale spatial patterns of epifaunal taxonomic structure across the UK shelf and North Sea, and into the key environmental drivers of those patterns.

Epifaunal taxon richness in the North Sea shows a clear south–north gradient, with significantly higher values in the north—a pattern consistent with previous studies (Jennings et al. 1999; Neumann et al. 2008; Reiss et al. 2010). This gradient, which is particularly marked near the 50 m depth contour, is often attributed to differences in water temperature and thermal stratification in deeper northern waters. However, the broader UK-shelf coverage of this study demonstrates that the northern North Sea is not the sole region of high epifaunal richness; comparably elevated richness also occurs in the English Channel, where environmental conditions differ substantially. These two regions also support distinct assemblage types (e.g. Epi_E1b in the northern North Sea and Epi_B1b in the mid-English Channel).

Recognising structural variation in epifaunal assemblages has important implications for managing human impacts. Assemblages with high biodiversity or limited spatial extent represent areas of greater ecological risk from activities that damage benthic invertebrates. Under current UK licensing frameworks for offshore activities (e.g. renewables, aggregate extraction, dredged material disposal), assessment of benthic impacts is primarily, though not exclusively, based on infaunal assemblages sampled using grabs or corers. While appropriate, this approach does not necessarily consider the ecological significance of the epifaunal component of sediment habitats, creating the risk that activities may be permitted in areas of high epifaunal ecological value (Chen et al. 2021). This study identifies regions where epifaunal assemblages exhibit relatively high taxon richness and densities, or where assemblage composition is limited to small, isolated areas. For example, Epi_A1—the most abundant and among the most taxon-rich assemblages in the study region—is predicted to occur in only a small area off the inshore east coast of England (0.1% of the study area). Similarly, Epi_B1a, occupying 0.5% of the area, is highly abundant and diverse but restricted to parts of the mid-English Channel. Such assemblages merit explicit consideration in environmental assessments to avoid irreversible loss. Consistent with the biodiversity–ecosystem function paradigm (Ali 2023), prioritising protection of biodiverse assemblages is also likely to safeguard those that are functionally important.

It should be noted that the sediment environmental rasters used in our modelling assume the presence of unconsolidated sediments throughout the study area. As a result, model predictions may be less reliable in regions where rocky substrates occur, since these assumptions do not hold. This represents a limitation in the spatial applicability of our results and should be considered when interpreting the model outputs. In addition to these data-related limitations, the modelling framework itself warrants consideration. While this study focused on random forest (RF) models, we acknowledge that other approaches—such as GLMs, MaxEnt, INLA, joint species distribution models, and neural networks—are available. RF was chosen for its flexibility, ability to model complex nonlinear relationships, and strong performance with large, heterogeneous ecological datasets. Correlated predictors generally do not reduce RF's predictive accuracy, but can bias variable importance metrics (Strobl et al. 2007), so we addressed collinearity during variable selection.

Our predictions for richness, square-root abundance, and assemblage clusters include spatial confidence maps. For richness and square-root abundance, CV values were generally low, indicating consistent predictions across model runs. However, R^2 values were moderate for richness (0.37) and lower for square-root abundance (0.13), highlighting limited explanatory power, particularly for abundance. This demonstrates that high model confidence (i.e. consistency) does not necessarily equate to high predictive accuracy, and both should be considered when interpreting results (McAlexander & Mentch 2020). Additionally, while RF offers strong predictive performance, it can be less interpretable than some parametric approaches. Although model comparison was not the aim of this study, future work could explore alternative approaches and further evaluation of model uncertainty to improve predictive performance and interpretability.

Future work should also address limitations in our approach to measuring taxon richness, which relied on simple counts of taxa per sample without incorporating taxon abundance information. This method does not adjust for variation in sampling effort, evenness, or spatial turnover. Applying statistical techniques such as rarefaction (which estimates the expected number of taxa in a standardized subsample, thereby accounting for differences in sampling effort; Gotelli and Colwell 2001; Chao and Jost 2012), as well as diversity metrics like Hill numbers (Hill 1973; Chao et al. 2014), and the Whittaker framework for partitioning diversity into alpha, beta, and gamma components (Whittaker 1960), would improve comparability and robustness of biodiversity assessments across space (see Cooper et al., in press).

To better support management of pressures and enable informed decisions about the ecological acceptability of activities affecting epifaunal assemblages, there is an urgent need to understand how assemblages differ in their sensitivity and response to anthropogenic pressures. Future research should focus on quantifying sensitivity using biological "response" traits and describing functional potential using "effect" traits. Trait-based approaches have been developed for infaunal benthic communities to assess sensitivity to activities such as trawling and dredged material disposal (Bolam et al. 2014, 2016, 2021); applying similar frameworks to epifauna would facilitate more effective protection of these communities (Lambert et al. 2014; Hewitt et al. 2018).

Finally, a more holistic understanding of the spatial variability in marine invertebrate assemblages will require integration of both infaunal and epifaunal components. The maps presented here can be used alongside existing maps for infauna (Cooper and Barry 2017; Cooper et al. 2019; Bolam et al. 2023, Cooper et al., in press; Bolam et al., in press) to identify regions of elevated ecological risk from human activities, reveal unique spatial patterns, and explore potential ecological interactions between infaunal and epifaunal communities.

Acknowledgements

We gratefully acknowledge all organisations and individuals who contributed data to the *OneBenthic* Portal, including those whose data were sourced via the Crown Estate's Marine Data Exchange https://www.marinedataexchange.co.uk/. Data provider information is available through the *OneBenthic* Data Extraction Tool (Trawl) https://rconnect.cefas.co.uk/onebenthic_dataextractiontrawl/. We also thank Paul McIlwaine (Cefas) for internal review and helpful comments on an earlier version of the manuscript, and the anonymous reviewers and editor for their constructive feedback.

Author contributions

K.C., M.C. and S.B. developed the initial concept for this paper, with input from AD. K.C. and S.B. drafted the manuscript, with contributions from M.C. and A.-L.D.. K.C. and A.-L.D. collaborated on the R code. All authors reviewed and contributed to the final manuscript.

Supplementary data

Supplementary data is available at *ICES Journal of Marine Science* online.

Conflict of interest: The authors declare no conflicts of interest.

Funding

This work, conducted under the POSEIDON (Planning Offshore Wind Strategic Environmental Decisions) project, contributes to the Offshore Wind Evidence and Change Programme funded by The Crown Estate. The views expressed are those of the authors, and neither The Crown Estate nor other project partners are responsible for any use of the information contained herein.

Data availability

All data and code associated with this study are documented in a publicly accessible metadata record on the Cefas Data Hub (https://doi.org/10.14466/CefasDataHub.188). This record includes links to the primary datasets, associated R scripts, and an API for accessing the modelled epifaunal biodiversity layers.

References

Ali A. Biodiversity–ecosystem functioning research: brief history, major trends and perspectives. *Biol Conserv* 2023;**285**:110210. https://doi.org/10.1016/j.biocon.2023.110210

Assis J, Tyberghein L, Bosh S *et al.* Bio-ORACLE v2.0: extending Marine Data Layers for Bioclimatic Modelling. *Global Ecol Biogeogr* 2018;**27**:277–84. https://doi.org/10.1111/geb.12693

Barry J, Maxwell D, Jennings S *et al.* Emon: an R-package to support the design of marine ecological and environmental studies, surveys and monitoring programmes. *Methods Ecol Evol* 2017;**8**:1342–6. https://doi.org/10.1111/2041-210X.12748

Böhner J, Selige T. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In J. Böhner, K. R. McCloy, J. Strobl (Eds.), *SAGA—Analysis and Modelling Applications. Göttinger Geographische Abhandlungen*(Vol. **115**, pp.13–28). Göttingen: Cuvillier Verlag. 2006.

Bolam SG, Coggan RC, Eggleton J *et al.* Sensitivity of macrobenthic secondary production to trawling in the Greater North Sea: a biological traits approach. *J Sea Res* 2014;**85**:162–77. https://doi.org/10.1016/j.seares.2013.05.003

Bolam SG, Cooper K, Downie A-L. Mapping marine benthic biological traits to facilitate future sustainable development. *Ecol Appl* 2023;**33**:e2905. https://doi.org/10.1002/eap.2905

Bolam SG, McIlwaine P, Garcia C. Marine benthic traits responses to the disposal of dredged material. *Mar Pollut Bull* 2021;**168**:112412. https://doi.org/10.1016/j.marpolbul.2021.112412

Bolam SG, McIlwaine PO, Garcia C. Application of biological traits to further our understanding of the impacts of dredged material disposal on marine benthic assemblages. *Mar Pollut Bull* 2016;**105**:180–92. https://doi.org/10.1016/j.marpolbul.2016.02.031

Bray JR, Curtis JT. An ordination of the upland forest communities of Southern Wisconsin. *Ecological Monographs* 1957;**27**:325–49. https://doi.org/10.2307/1942268

Breiman L. Random Forests. *Machine Learning* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324

Callaway R, Alsvag J, De Boois I *et al.* Diversity and community structure of epibenthic invertebrates and fish in the North Sea. *ICES J Mar Sci* 2002;**59**:1199–214. https://doi.org/10.1006/jmsc.2002.1288

Chamberlain S, Ooms J. worrms: World Register of Marine Species (WoRMS) Client. *R package version 0.4.3*. 2023

Chao A, Gotelli N, Hsieh TC *et al.* Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs*. 2014;**84**:45–67. https://doi.org/10.1890/13-0133.1

Chao A, Jost L. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* 2012;**93**:2533–47, https://doi.org/10.1890/11-1952.1

Chen Y-Y, Edgar GJ, Fox RJ. The nature and ecological significance of epifaunal communities within marine ecosystems. *Oceanography and Marine Biology: An Annual Review*, 2021;**59**:585–720. https://doi.org/10.1201/9781003138846-9

Clarke KR, Gorley RN. *PRIMER v7: User Manual/Tutorial. PRIMER-E Ltd*, Plymouth. 2015.

Conrad O, Bechtel B, Bock M *et al.* System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development Discussions* 2015;**8**:2271–312. https://doi.org/10.5194/gmdd-8-2271-2015

Cooper KM, Barry J. A big data approach to macrofaunal baseline assessment, monitoring and sustainable exploitation of the seabed. *Sci Rep*, 2017;7:12431https://doi.org/10.1038/s41598-017-11377-9

Cooper KM, Bolam SG, Downie AL *et al.* Biological-based habitat classification approaches promote cost-efficient monitoring: an example using seabed assemblages. *J Appl Ecol*. 2019;**56**:1085–98. https://doi.org/10.1111/1365-2664.13381

Cooper KM, Thompson MSA, Bolam SG *et al.* Mapping benthic biodiversity to facilitate future sustainable development. *Ecosphere*. 2025

Costa BM, Battista TA, Pittman SJ. Comparative evaluation of airborne LiDAR and ship-based multibeam SoNAR bathymetry and intensity for mapping coral reef ecosystems. *Remote Sens Environ* 2009;**113**:1082–100. https://doi.org/10.1016/j.rse.2009.01.015

Cressie NAC. Statistics for spatial Data: Revised Edition. *Wiley series in probability and mathematical statistics*. 1993, 424pp. ISBN 0-471-00255-0. https://rongxie.files.wordpress.com/2011/01/statistics-for-spatial-data-revised-version-1993.pdf (1 June 2025, date last accessed).

Cutler D, Edwards T, Beard K *et al.* Random Forests for Classification in Ecology. *Ecology* 2007;**88**:2783–92. https://doi.org/10.1890/07-0539.1

Davies AJ, Guinotte JM. Global habitat suitability for framework-forming cold-water corals. *PLoS One* 2011;**6**:e18483https://doi.org/10.1371/journal.pone.0018483

Desmet P.J.J., Govers G. A GIS procedure for automatically calculating the USLE LS factor on topographically complex landscape units. *Journal of Soil and Water Conservation* 1996;**51**:427–433.

Ellis J, Lancaster JE, Cadman PS *et al.* The marine fauna of the Celtic Sea. *Marine Biodiversity in Ireland and Adjacent Waters*. 2002;45–65.

. Fraschetti S, Strong J, Buhl-Mortensen L, Alexander B., Rodriguez Perez A., Kellett P., Muñiz Piniella A., Bayo Ruiz F., Bairaktari K., Heymans J. J. *et al.* Marine habitat mapping[Eds.] *Future Science Brief N°. 11 of the European Marine Board*, Ostend, Belgium. 2024 ISSN: 2593-5232. ISBN: 9789464206234. https://doi.org/10.5281/zenodo.11203128

Gotelli N, Colwell R. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* 2001;**4**:379–91. https://doi.org/10.1046/j.1461-0248.2001.00230.x

Griffiths JR, Kadin M, Nascimento FJA *et al.* The importance of benthic-pelagic coupling for marine ecosystem functioning in a changing world. *Global Change Biol* 2017;**23**:2179–96. https://doi.org/10.1111/gcb.13642

Guidi L, Fernández Guerra A, Canchaya C *et al. Big Data in Marine Science. Future Science Brief 6 of the European Marine Board* Belgium: European Marine Board, 2020. https://doi.org/10.5281/zenodo.3755793

Halpern BS, Walbridge S, Selkoe KA *et al.* A global map of human impact on marine ecosystems. *Science* 2008;**319**:948–52. https://doi.org/10.1126/science.1149345

Harris PT, Baker EK. Why map benthic habitats?In: PT Harris, EK Baker (eds) *Seafloor geomorphology as benthic habitats: GeoHab atlas of seafloor geomorphic features and benthic habitats*. Amsterdam: Elsevier, 2012, p3–22.

Hewitt JE, Lundquist CJ, Ellis J. Assessing sensitivities of marine areas to stressors based on biological traits. *Conserv Biol* 2019;**33**:142–51. https://doi.org/10.1111/cobi.13181

Heyman WD, Wright DJ. Marine geomorphology in the design of marine reserve networks. *The Professional Geographer* 2011;**63**:429–42. https://doi.org/10.1080/00330124.2011.585074

Hill MO. Diversity and Evenness: a unifying notation and its consequences. *Ecology* 1973;**54**:427–32. https://doi.org/10.2307/1934352

Huang BFF, Boutros PC. The parameter sensitivity of random forests. *BMC Bioinf* 2016;**17**, 331. https://doi.org/10.1186/s12859-016-1228-x

Jennings S, Lancaster J, Woolmer A *et al.* Distribution, diversity and abundance of epibenthic fauna in the North Sea. *Journal of the Marine Biological Association of the United Kingdom* 1999;**79**:385–99. https://doi.org/10.1017/S0025315498000502

Kaiser MJ, Rogers SI, Ellis JR. Importance of benthic habitat complexity for demersal fish assemblages. *American Fisheries Society Symposium* 1999;**22**:212–23.

Kramer MJ, Bellwood O, Fulton CJ *et al.* Refning the invertivore: diversity and specialisation in fish predation on coral reef crustaceans. *Mar Biol* 2015;**162**:1779–86. https://doi.org/10.1007/s00227-015-2710-0

Lambert GI, Jennings S, Kaiser MJ *et al.* Quantifying recovery rates and resilience of seabed habitats impacted by bottom fishing. *J Appl Ecol* 2014;**51**:1326–36. https://doi.org/10.1111/1365-2664.12277

Lecours V, Devillers R, Schneider DC *et al.* Spatial scale and geographic context in benthic habitat mapping: review and future directions. *Marine Ecology Progress Series* 2015;**535**:259–84. https://doi.org/10.3354/meps11378

Lecours V. On the Use of Maps and Models in Conservation and Resource Management (Warning: results May Vary). *Frontiers in Marine Science* 2017;**4**:288. https://doi.org/10.3389/fmars.2017.00288

Liaw A., Wiener M. Classification and Regression by randomForest. *R News* 2002; **2**(3), 18–22. https://CRAN.R-project.org/doc/Rnews/.

MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. Le Cam, J. Neyman (Eds.), 281–97. Berkeley: University of California Press. 1967.

McAlexander RJ, Mentch L. Predictive inference with random forests: a new perspective on classical analyses. *&Research & Politics*, 2020;**7**:1–7. https://doi.org/10.1177/2053168020905487

McIlwaine PSO, Barry PJ, Curtis M *et al.* Tracking long-term benthic recovery at a disused marine aggregate extraction site using monitoring tools developed for the marine aggregate industry. *Estuarine Coastal Shelf Sci* 2025;**319**:109278. https://doi.org/10.1016/j.ecss.2025.109278

Mitchell P, Aldridge J, Diesing M. *Quantitative Sediment Composition Predictions for the North-West European Continental Shelf*. UK: Cefas. V1. 2019. https://doi.org/10.14466/CefasDataHub.63

Mitchell PJ, Downie A-L, Diesing M How Good Is my Map? A Tool for Semi-Automated Thematic Mapping and Spatially Explicit Confidence Assessment. *&Environmental Modelling & Software* 2018;**108**:111–22. https://doi.org/10.1016/j.envsoft.2018.07.014

Murillo FJ, Weigel B, Marmen MB *et al.* Marine epibenthic functional diversity on Flemish cap (north-west Atlantic)—identifying trait responses to the environment and mapping ecosystem functions. *Diversity and Distributions* 2020;**26**:460–78. https://doi.org/10.1111/ddi.13026

Naimi B, Hamm NAS, Groen TA *et al.* Where is positional uncertainty a problem for species distribution modelling? *Ecography* 2014;**37**:191–203. https://doi.org/10.1111/j.1600-0587.2013.00205.x

Neumann H, Ehrich S, Kröncke I. Spatial variability of epifaunal communities in the North Sea in relation to sampling effort. *Helgoland Marine Research* 2008;**62**:215–25. https://doi.org/10.1007/s10152-008-0109-8

Newcombe EM, Taylor RB. Trophic cascade in a seaweed-epifauna-fish food chain. *Marine Ecology Progress Series* 2010;**408**:161–7. https://doi.org/10.3354/meps08589

O'Brien JM, Stanley RRE, Jeffery NW *et al.* Modelling demersal fish and benthic invertebrate assemblages in support of marine conservation planning. *Ecol Appl* 2022;**32**:e2546. https://doi.org/10.1002/eap.2546

Oksanen J, Guillaume Blanchet F, Friendly M *et al.* vegan: Community Ecology Package. *R package version 2.7-0* 2024. https://CRAN.R-project.org/package=vegan (Accessed: 2025-03-26)

R Core Team *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. 2025. https://www.R-project.org (26 March 2025, date last accessed).

Rees HL, Eggleton JD, Rachor E *et al. Structure and dynamics of the North Sea benthos*. ICES Cooperative Research Report, No. 288. Copenhagen: International Council for the Exploration of the Sea (ICES). 258pp. 2007. https://doi.org/10.17895/ices.pub.5451

Reiss H, Degraer S, Duineveld GCA *et al.* Spatial patterns of infauna, epifauna, and demersal fish communities in the North Sea. *ICES J Mar Sci* 2010;**67**:278–93. https://doi.org/10.1093/icesjms/fsp253

Rodriguez-Galiano VF, Ghimire B, Rogan J *et al.* An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification. *ISPRS J Photogramm Remote Sens* 2012;**67**:93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

Saleem S, Aslam M, Shaukat MR., A review and empirical comparison of univariate outlier detection methods. *Pak. J. Statist.* 2021;**37**:447–62.

Stelzenmüller V, Vega Fernandez T, Cronin K *et al.* Assessing uncertainty associated with the monitoring and evaluation of spatially managed areas. *Mar Policy* 2015;**51**:151–62. https://doi.org/10.1016/j.marpol.2014.08.001

Strobl C, Boulesteix A-L, Zeileis A *et al.* Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinf* 2007;**8**:25. https://doi.org/10.1186/1471-2105-8-25

Taylor RB. Density, biomass and productivity of animals in four subtidal rocky reef habitats: the importance of small mobile invertebrates. *Marine Ecology Progress Series* 1998;**172**:37–51. https://doi.org/10.3354/meps172037

Tyberghein L, Verbruggen H, Pauly K *et al.* Bio-ORACLE: a Global Environmental Dataset for Marine Species Distribution Modelling. *Global Ecol Biogeogr* 2012;**21**:272–81. https://doi.org/10.1111/j.1466-8238.2011.00656.x

Whittaker RH. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* 1960;**30**:279–338. https://doi.org/10.2307/1943563

WoRMS Editorial Board World Register of Marine Species. 2025. Available fromhttps://www.marinespecies.org at VLIZhttps://doi.org/10.14284/170. (26 March 2025, date last accessed)

Zühlke R, Alsvag J, De Boois I *et al.* Epibenthic diversity in the North Sea. *Senckenbergiana Maritima* 2001;**31**:269–81.

Handling editor: Ken Andersen